

Nucleotide Sequence of Bacteriophage λ DNA

F. SANGER, A. R. COULSON, G. F. HONG, D. F. HILL†
AND G. B. PETERSEN†

Laboratory of Molecular Biology, The Medical Research Centre
Hills Road, Cambridge CB2 2QH, England

(Received 15 September 1982)

The nucleotide sequence of the DNA of bacteriophage λ has been determined using the dideoxy chain termination method in conjunction with random cloning in M13 vectors. Various methods were studied for sequencing specific regions to complete the sequence, but all were much slower than the random approach. The DNA in its circular form contains 48,502 base-pairs. Open reading frames were identified and, where possible, ascribed to genes by comparing with the previously determined genetic map. The reading frames for 46 genes were clearly identified, though in about 20 the position of the protein initiation site could not be rigorously established. Probable positions for the *kil*, *cIII* and *lom* genes are suggested but remain uncertain. There are about 20 other unidentified reading frames that may code for proteins.

The genome is fairly compact with comparatively little non-coding DNA. In many cases the translation terminators and initiators overlap, particularly in the sequence A-T-G-A where the TGA terminates one gene and the ATG initiates the next. Such structures seem to be characterized by a purine-rich sequence, rather than by a specific "Shine and Dalgarno" sequence, before the initiator. In the whole of the left arm the codon CTA, which is normally read by a minor leucine tRNA, is absent. The distribution of other rare codons in the genes of the left arm suggests that they may have a controlling function on the relative amounts of the proteins produced.

1. Introduction

New methods for sequencing DNA (Maxam & Gilbert, 1977; Sanger *et al.*, 1977) have made it possible to determine sequences of several hundred nucleotides very rapidly. These techniques were originally applied to restriction enzyme fragments and as they improved the rate of sequencing became increasingly dependent on the isolation methods for the fragments. A major step forward was achieved by replacing standard fractionation methods with a cloning procedure using single-stranded bacteriophage as the vector. Particularly successful were the derivatives of bacteriophage M13 developed by Messing and his colleagues (Gronenborn & Messing, 1978; Messing *et al.*, 1981), which could be used in conjunction with the chain termination techniques (Sanger *et al.*, 1977, 1980). During our studies on

† Present address: Department of Biochemistry, University of Otago, Box 56, Dunedin, New Zealand.

human mitochondrial DNA (Anderson *et al.*, 1981) there were difficulties in achieving the containment facilities required at that time when using M13 vectors on human DNA; consequently, we decided to try out the method on bacteriophage lambda DNA. Progress was rapid and most of the sequence was completed by the time we were able to apply the methods to mitochondrial DNA. Because of the general interest in and importance of lambda in biochemical and genetic studies, we have completed the sequence.

Bacteriophage lambda DNA in its circular form contains 48,502 base-pairs, and codes for about 60 proteins. It has been studied extensively by genetic techniques and a fairly complete gene map prepared (Echols & Murialdo, 1978; Szybalski & Szybalski, 1979; Daniels *et al.*, 1980). In many cases the proteins encoded have been identified as bands on sodium dodecyl sulphate/polyacrylamide gels, but in only a few cases has an amino acid sequence been determined.

When we started the work, a number of DNA sequences had already been established in lambda (about 6000 base-pairs), particularly in control regions. We have not rigorously re-investigated these sequences and have usually ignored clones that were found to contain them. Other sequences have been reported while our work was still in progress. In general, our results have confirmed these sequences.

2. Materials and Methods

The lambda DNA used was the standard strain *λcJindlts857Sam7*. It was prepared at the Centre for Applied Microbiology and Research, Porton. DNA polymerase I (Klenow subfragment) was obtained from Boehringer Mannheim. Restriction enzyme *Fnu*DII was a gift from Professor M. Smith. Other restriction enzymes were from New England Biolabs. The M13 vectors were a gift from Dr J. Messing.

(a) DNA sequencing

The "shotgun" sequencing was carried out initially as described previously (Sanger *et al.*, 1980) using restriction enzyme digests, the vectors mp2 (Gronenborn & Messing, 1978) or mp7 (Messing *et al.*, 1981) and priming with the 96-nucleotide R1-R1 primer, or the 30-nucleotide primer prepared by Anderson *et al.* (1980). In later work the DNA degradation was by digestion with DNase (Anderson, 1981) or by sonication (Fuhrman *et al.*, 1981) and priming was with the synthetic 17-mer prepared by Duckworth *et al.* (1981). The "dideoxy" sequencing was carried out in Eppendorf polypropylene tubes rather than in capillaries. Considerable improvement in the reading of sequences could be obtained by drying the gels after fixing in acetic acid (see Garoff & Ansorge, 1981). This could be conveniently achieved by leaving the gel uncovered at room temperature for several hours and then heating at 67°C till completely dry. Alternatively, a conventional gel dryer could be used.

The main source of error in the sequencing is "compressions" on the gels. Instead of there being a regular distance between each band, some bands run closer together or occupy the same position. This appears to be caused by secondary structure effects, which are not broken by the denaturing conditions in the gels, with the result that a small region at the 3' end of the newly synthesized DNA becomes double-stranded and the corresponding bands run abnormally fast in the electrophoresis. The best way to overcome this problem is by determining the sequence "in both directions"; i.e. on 2 clones with the sequence in opposite orientations. It is unlikely that the compression will occur in the same place in both clones. Two other techniques have been used to overcome this problem. One is to replace dGTP in the mixtures with dITP (deoxyinosine triphosphate). I·C base-pairing is weaker than G·C, so that the secondary structures are less likely to form (Mills & Kramer, 1979). Usually it is

sufficient to run an extra dITP track in which 2 mM-dITP is used with ddGTP (0.05 mM) next to the normal ddGTP reaction on a sequencing gel. The reaction is "chased" with a mixture of 0.5 mM-dGTP and 0.5 mM-dATP. By comparing the positions of the G bands on the 2 tracks in a region where a compression has occurred, it is usually possible to resolve the problem. The method does not work on clones prepared in the vector M13mp7. Presumably the secondary structure that can be formed due to complementary sequences in the vector inhibits the progress of the DNA polymerase in the presence of dITP. Another method of resolving compressions is by running the gels in 25% (v/v) formamide.

Another occasional source of error occurs mainly at G-T sequences. At the position of the G residue, bands appear in both the G and T channels with the T band sometimes stronger than the G. The effect is usually associated with unsatisfactory polymerases and can usually be eliminated by using a higher concentration of enzyme.

(b) Identification of clones by hybridization

In order to prepare clones covering regions of incomplete or doubtful sequence, ^{32}P -labelled probes were prepared from clones that we already had and that were complementary to a sequence at or near to the required position. These had usually been stored as DNA in 10 mM-Tris-HCl (pH 7.4), 0.1 mM-EDTA at -20°C. Competent cells were transfected with a small amount (0.1 to 0.5 μl) of the DNA and plated out to give new plaques. The number of plaques varied considerably, depending largely on the age of the DNA. A better method of storing clones is to set aside about 50 μl of the supernatant after centrifuging down the bacteria, and keeping it at -20°C in small capped polypropylene tubes. The low-phosphate growth medium contained 0.15% (w/v) KCl, 0.5% (w/v) NaCl, 0.1% (w/v) NH_4Cl , 0.2% (w/v) vitamin-free Casamino acids, 0.2% (w/v) Bactopeptone, 1.21% (w/v) Tris (adjusted to pH 7.4), 0.04% (w/v) CaCl_2 , 0.5% (w/v) glucose: 0.2 ml was added to each sterile glass tube and a plaque or 5 μl of a bacteriophage supernatant was added. To the remaining low-phosphate medium, 10 μl of a log culture of *Escherichia coli* strain JM101 in 2 \times TY medium and 10 μCi of [^{32}P]phosphate (carrier free) were added per 1 ml of medium, and 1 ml of this added to each of the tubes. The cells were grown for 6 to 10 h and the DNA isolated as described (Sanger *et al.*, 1980). Usually, the cultures were combined before isolation and several regions were probed for at the same time. The isolated DNA was finally dissolved in 30 to 60 ml of the hybridizing solution (Southern, 1979). This was sufficient for 5 to 10 filters.

A random library of sonicated lambda fragments (average size 500 nucleotides) was prepared (Fuhrman *et al.*, 1981) and cloned. Individual pure plaques were usually transferred to fresh plates so that there were about 50 per plate. Blots were prepared using nitrocellulose filters and hybridization and washing carried out as described by Jeffreys & Flavell (1977). Autoradiographs were developed with fluorescent screens for 1 to 3 days and DNA prepared from the plaques giving a positive response. In order to identify plaques containing the desired sequence, a preliminary screening with the ddT or ddA reaction was carried out. The results using ddA are usually rather clearer than those with ddT and were done in the same way (Sanger *et al.*, 1980), except that the reaction mixture contained H buffer, 0.1 mM-dGTP, 0.1 mM-dCTP, 0.1 mM-TTP and 0.02 mM-ddATP.

(c) Subcloning from clones with larger inserts

In order to digest a cloned DNA with a restriction enzyme, double-stranded DNA was prepared using DNA polymerase primed with the normal flanking primer (Hong, 1981). For clones that had been stored for a long time, fresh DNA was usually prepared by growing a 1 ml culture and dissolving the final product in 25 μl of Tris-EDTA (Sanger *et al.*, 1980). To 10 μl were added 2 μl of 10 \times H buffer, 2 μl (0.4 pmol) of primer and 6 μl of water. The mixture was heated in a sealed capillary tube at 100°C for 3 min and allowed to cool slowly to room temperature. To this was added 20 μl of a solution containing 5 μCi of [α - ^{32}P]dATP (spec. act. approx. 400 Ci/mmol), 0.1 mM-dGTP, 0.1 mM-dCTP, 0.1 mM-TTP and 1.5 μM -dATP

in H buffer followed by 2 μ l (2 units) of DNA polymerase. After incubation for 15 min at room temperature, the appropriate restriction enzyme(s) was added and incubated at 37°C for a further 15 min (65°C for *TaqI*). A 2- μ l sample was usually taken at this stage and run on an acrylamide gel to determine whether the digestion was satisfactory. The ligation was done using vector DNA (M13mp8 or mp9) cut with restriction enzyme *Sma*I to generate blunt ends (Messing & Vieira, 1982), and treated with phosphatase. Therefore, if enzymes that produce staggered ends had been used, it was necessary to convert these to blunt ends. After heating at 70°C for 10 min to inactivate the enzyme, 1 μ l of 0.5 mM-dATP and 1 μ l of DNA polymerase were added and the mixture incubated at room temperature for 15 min: 1 μ l of 0.2 M-EDTA, 50 μ l of water and 50 μ l of water-saturated phenol were added. After separation of the aqueous layer, the DNA was precipitated with ethanol and the product dissolved in 5 μ l and applied to a 6% acrylamide gel (containing no urea). The required band was located by autoradiography and eluted by the method of Maxam & Gilbert (1977). After precipitation with ethanol it was dissolved in 5 μ l of water; 2 μ l of this solution was used for ligation into 20 ng of the vector and the DNA transfected into competent cells, which were then plated. Several plaques were usually screened using the ddATP reaction.

3. Results and Discussion

(a) Determination of the sequence

This work was largely carried out as a study of methods, so that we have used a variety of techniques rather than limiting ourselves to the most productive ones.

One of the most important attributes of the cloning approach is that it is a uniquely efficient purification procedure that is not affected by the size of the DNA being studied. Consequently, no attempt was made to carry out an initial restriction enzyme digestion and fractionation of the lambda into smaller fragments, but the whole DNA was digested and the resultant complex mixtures cloned directly. We believe that this approach is more rapid and efficient, though it may be less easy to carry out in a research environment where each collaborator would prefer to study a unique piece of the sequence.

Initially, we used restriction enzyme digests and inserted the fragments into the vector M13mp2 using RI linkers (Sanger *et al.*, 1980). The most successful of these were digests obtained with *Alu*I, *Hae*III, *Hinf*I, *Fnu*DII, *Taq*I, *Dde*I, and *Sau*96A. About 40 to 70 clones were normally sequenced using the "flanking" primer before redundant clones became too frequent. Good results were also obtained using *Sau*3A digests and cloning in a *Bam*HI vector that was an amber derivative of strain M13mWJ43 (Winter & Fields, 1980).

Because of their specificity, the use of restriction enzymes imposes a limitation on the number of useful clones that can be obtained from them, whereas from the random approach it is desirable to have as complex and varied a mixture as possible representing the whole of the DNA in equivalent amounts. Anderson (1981) has described a method of random cleavage using DNase I in the presence of manganese, and Deininger has used sonication (see Fuhrman *et al.*, 1981). Using these methods, it is possible to get many more useful clones from a single digest and these have been mainly used in the more recent stages of the work.

Initially, the method of sequencing was that described by Sanger *et al.* (1980). The DNA from each clone was sequenced by the chain terminating technique with a flanking primer, and the results entered into the data base using the computer programs of Staden (1980). To begin with, new data accumulated rapidly.

However, as the work proceeded more of the sequences were redundant and it seemed that it would be necessary to replace the random approach with a more specific one in order to find certain missing sequences. We chose to do this when about 90% of the whole sequence was in the data base in the form of about 10 to 15 contiguous sequences ("contigs": Staden, 1980). In fact, it would probably have been better to continue the random approach farther, since it is much more rapid than any specific methods. The final stages of the sequence analysis required, firstly, joining the separate contigs and then confirming and correcting uncertain sequences. Two methods were particularly useful for the former. The ends of contigs can often be extended by reading the DNA sequence further from certain clones, so that any improvements in the electrophoresis can prove very valuable. Normally, sequences of 250 to 300 nucleotides were obtained from each clone. Using special care and techniques it is possible to read a gel out to as much as 500 nucleotides. In our experience, using lower acrylamide concentrations (5%) and drying the gels before autoradiography have proved the most successful. Although it is frequently possible to overlap contigs in this way, the sequences obtained are usually less accurate and will require subsequent confirmation by some other method.

A useful way of extending contigs is by sequencing the insert in an appropriate clone from its opposite end. (The sequence normally determined is the complement of the plus strand of the M13 recombinant.) Two methods have been described for obtaining the sequence of the opposite strand, both of which have been used in this work. In the method of Winter *et al.* (1981), the insert in the clone is cut out with a suitable restriction enzyme and re-ligated into the vector: 50% of the resultant clones should contain the insert in the opposite orientation and can be sequenced with the normal flanking primer. Another method, developed by Hong (1981), makes use of the normal flanking primer to make an unlabelled complete copy of the insert and this is sequenced using a synthetic primer corresponding to a sequence of the M13 at the opposite end of the clone.

The main method we have used for obtaining sequences from a specific region is the use of hybridization probes to select clones from random mixtures. For probes, we chose M13 clones having the complementary sequence to that required and the DNA from these was labelled by growing on medium containing [^{32}P]phosphate. Several probes could be used at the same time to screen blots from plates containing random clones. Although not all of the positively reacting clones were from the desired region, the method was relatively rapid and enabled us to complete the sequence. Towards the end of the work when the whole sequence had been studied, there were some regions that were still uncertain, either because they had only been covered by gel readings on one strand or had been read from a region far out from the priming sequence. These could frequently be studied by taking a clone containing a larger insert, digesting it with an appropriate restriction enzyme (which could be identified from the sequence we already had) and re-cloning the smaller fragment. This was a relatively slow process but was useful at this point in the work.

With a sequence the size of lambda it is difficult to ensure that there are no mistakes, but it is believed that there cannot be many. Most of the sequence has

TABLE I
Sequences not confirmed in this work

28.462–28.807	Hoess <i>et al.</i> (1980)
35.246–35.279	Franklin & Bennett (1979)
37.710–38.040	Schwarz <i>et al.</i> (1978)
38.198–38.498	Schwarz <i>et al.</i> (1978); Roberts <i>et al.</i> (1977)
38.681–38.756	Schwarz <i>et al.</i> (1978)
39.658–39.699	Schwarz <i>et al.</i> (1980)
44.170–44.500	Daniels & Blattner (1982); Petrov <i>et al.</i> (1981)

been determined on both strands. Where this was not done there were usually several gels giving clear and unequivocal readings. The only sequences that we have not determined are listed in Table 1. They come to a total of 1454 nucleotides and were completed by others before we started sequencing.

(b) *Identification of genes*

Figure 1 shows the gene map based on both previous work and the sequence. Since transcription is in different directions in different parts of the genome, the

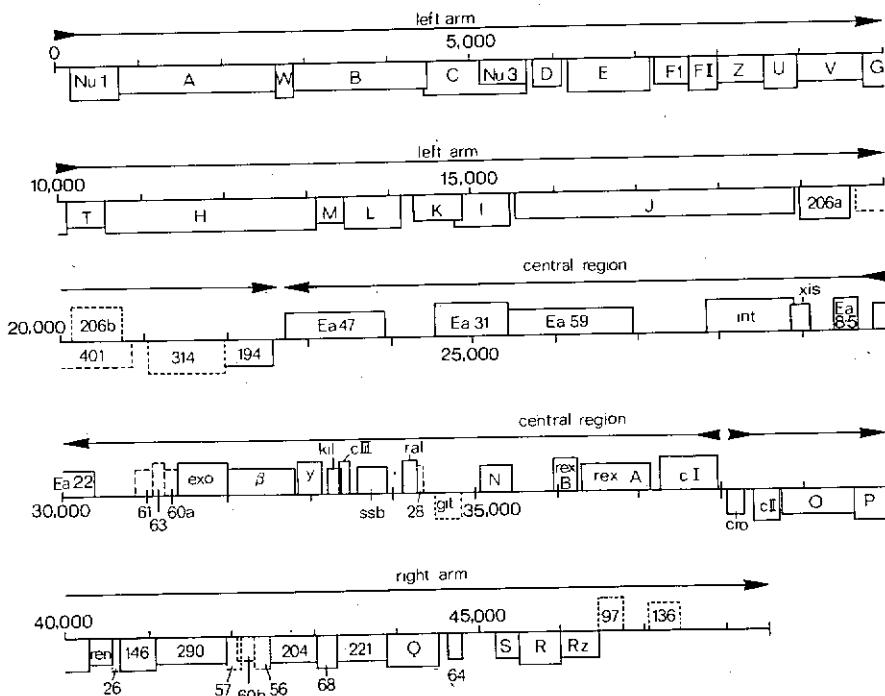


FIG. 1. Gene map of lambda showing the most probable reading frames for the proteins. Numbering is from the left-hand end. Large arrows above the genes show the main directions of transcription and translation. Genes in boxes below the line are translated from left to right; those above the line are translated from right to left. Boxes with broken lines indicate more speculative reading frames. *gil* is a gene proposed by Ineichen *et al.* (1981). Where no gene or protein product has been assigned to an open reading frame (*orf*) it has been assigned a number that corresponds to the number of amino acid residues that the putative protein would contain.

TABLE 2
Sizes of the lambda proteins

Gene	Left arm		Central region			Right arm		
	Mol. Wt.		Gene	Mol. Wt.		Gene	Mol. Wt.	
	from gels† (kdal)	from sequence‡ (daltons)		from gels† (kdal)	from sequence‡ (daltons)		from gels† (kdal)	from sequence‡ (daltons)
<i>NuI</i>	21	20,444	<i>cI</i>	27	26,245	<i>cro</i>	7-9	7365
<i>A</i>	80	73,280	<i>rexA</i>	29	31,259	<i>cII</i>	11-14	11,057
<i>W</i>	5-10	7614	<i>rexB</i>	—	15,970	<i>O</i>	34-37	33,870
<i>B</i>	60	59,474	<i>X</i>	13.5	15,376	<i>P</i>	23-27	26,522
<i>C</i>	61	45,915	<i>ral</i>	—	7605	<i>ren</i>	—	10,613
<i>Nu3</i>	19	20,789	<i>ssb</i>	10-16	13,782	<i>orf146</i>	—	16,650
<i>D</i>	11-12	11,574	<i>cIII</i>	—	6048	<i>orf290</i>	—	33,566
<i>E</i>	38	38,194		or	10,067§	<i>orf57</i>	—	6964
<i>F1</i>	17	14,310	<i>kil</i>	16	5519	<i>orf60b</i>	—	7384
<i>FII</i>	11.5	12,731		or	11,646§	<i>orf56</i>	—	6321
<i>Z</i>	20	21,564	<i>y</i>	16.5	11,646	<i>orf204</i>	—	24,119
<i>U</i>	13-16	14,651		or	16,349§	<i>orf68</i>	—	7882
<i>V</i>	25-31	25,814	<i>g</i>	28-30	29,693	<i>orf221</i>	—	25,222
<i>G</i>	33	15,610	<i>exo</i>	25-26	25,912	<i>Q</i>	23	22,476
<i>T</i>	16	16,061	<i>Ea22</i>	19-22	20,939	<i>orf64</i>	—	7084
<i>H</i>	90	92,294	<i>Ea8.5</i>	8.5	10,751	<i>S</i>	—	11,522
<i>M</i>	10	12,532	<i>xis</i>	—	8606	<i>R</i>	18	17,828
<i>L</i>	29	25,713	<i>int</i>	36-44	40,310	<i>R_Z</i>	11-13	17,232
<i>K</i>	27	23,014	<i>Ea59</i>	—	59,494			
<i>I</i>	—	23,127	<i>Ea31</i>	—	34,588			
<i>J</i>	130-140	124,440	<i>Ea47</i>	—	48,103			
<i>orf206a</i>	—	21,859	<i>orf206b</i>	—	20,186			
<i>orf401</i>	—	39,889						
<i>orf314</i>	—	32,024						
<i>orf194</i>	—	21,605						

† SDS-acrylamide gel electrophoresis. For references see Szybalski & Szybalski (1979).

‡ Calculated using programs TRANDK (Staden, 1978) and MWCALC (Staden, unpublished).

§ Ineichen *et al.* (1981).

sequence is presented in three sections, together with the proposed amino acid sequences of the protein genes (Fig. 2). Transcription and translation are shown from left to right. The numbering throughout is from the left-hand end of the genome; therefore in the central region (Fig. 2(b)) it goes backwards. In the main, genes have been identified by examining the sequence for open reading frames of the expected length (see Table 2) preceded by a suitable initiation codon and comparing with the genetic map.

The *R* gene protein was conclusively identified by comparison of the DNA sequence with the known amino acid sequence (Imada & Tsugita, 1971). In order to establish unequivocally the position of some of the structural genes, limited amino acid sequence data at the N termini were obtained (Walker *et al.*, 1982). In this way, the N-terminal sequences of the proteins of genes *D*, *E*, *V*, *B** and *H* were determined and the corresponding position in the DNA sequence identified.

(a)

GGGCGGCGHCTCGGGTTTCGCTATTATGAAAHATTTCGGGTTAAGCGTTCCTCTCGTCATAACTTAATGTTTAT
18 20 30 50 60 70 80 90 100

TTAATACCTCTGAAAGAHGGAGCAGCHCAGGTGCTGAAGCGAGGGCTTGGGCTCTGCTGTTCTCTGTTTGCGT
100 112 120 130 140 150 160 170 180

H E V N K K O L A D I F G A S I R T Q N H O E G M
GGAAATGAGATGGAGTCACAAAAGGAGCTGGTGCACCTTCCGTCAGGATCTCGAACCTGAGGAACTGGCAGAACAGGAA
190 200 210 220 230 240 250 260 270

P U L R G G G K N E U L Y D S R A U I K H Y A E R D A E I
GCCGCTTCGGAGGGGGGGAGGTGATGGGCTTATGACTCTGGCGCGTCATAAATGGTATGCCGAAAGGGAGCTGCA
280 290 300 310 320 330 340 350 360

E H E K L R R E U E E L R O A S E A D L Q P G T I E Y E R H
TGAGAACGAGAGAGCTGCGCCGGGGAGGTGAAAGAACGCTGGCGAGGGCAGATCTCCAGCCAGGAACTTGGATCGAACGCC
370 380 390 400 410 420 430 440 450

R L T R A Q A D A Q E L K N H A R D S A E U U E T A F C T F U
TCGACTTACCGCGCAGGGCAGGGAGGTGAAAGAACGCTGGCGAGGGCAGATCTCCGCTGAGTGGTGAACCCGATCTGACTTGG
460 470 480 490 500 510 520 530 540

L S R I A G E I A S I L D G L P L S U O R R F P E L E F N R H
GCTGCGGGATCGAGGTTGGGAGGTGAAATTCGCGGGCTCCCCCTGCGAGGGGGTTTCCGGAACTGGGAAACCGAC
550 560 570 580 590 600 610 620 630

U D F L K R D I T K A N N K A A A L D E L I P G L L S E Y I
TGTGATTCTGAAAGGGATATCATCAAGGCCATGAAAGCAGCCGCTGGATGAACTGATACCGGGGTTGCTGAGTGATRATAT
640 650 660 670 680 690 700 710 720

H u i P E O S G U * H N S V U N R L R H F U R A G L R S L F R P E P O T A U E H A
CGAACAGCTGGGTTAACAGGGTGGCATTTGGCCGGCGGGCTCGGTCACTGTCAGGCCGGAGCACAGGCCGCTGGATGG
730 740 750 760 770 780 790 800 810

A D A H Y L L P K E S A Y E G R W E T L P F O R A I M H A M
CGGGCTATTAATCTACTATCTCCGAAAGAACGTCATCCAGGGAGGGCTGGGGAGACTGCCCCTTCAGCGGGCCATCTGA
820 830 840 850 860 870 880 890 900

A G S D Y I R E U N U V K S A R U G Y S K M L L G U Y A Y F I
TGGCAGGGACTACATCGTGGAGGTGATGGTGAAGTCTGGCGTGGTATTCCAAATGCTGGGGTTTATGCTACT
910 920 930 940 950 960 970 980 990

A F H K Q R N T L I H L P T D G D A E N F M K T H U E P T I R
TAGAGCATAAAGCAGGCAACACCCCTATCTGGTGGGGAGGGTGTGGCAGAGACTTATGAAACCCACGCTGGGGACTATTC
1000 1010 1020 1030 1040 1050 1060 1070 1080

A D I P S L L A L A P W Y G K K H R D N T L T M K R F T N G R
GTGATATTCTGGCTGCTGGCCCTGGGATTTGGCAAAAGAACCGGGATACACGCTTACCATGAAGCGTTTCACTATGGGC
1090 1100 1110 1120 1130 1140 1150 1160 1170

A G F C P C L G G K A A K N Y R E K S U D U A G Y D E L A R F D
GTGGCTCTGGCTGGGGCTAAAGGGGCHAAGAACATCCGGTGGATGAGCTGGGGTTATGATGAGACTTGGCTGCTTGG
1180 1190 1200 1210 1220 1230 1240 1250 1260

A D D I E Q E G S P T F L G D K R I E G S U W P K S I R G S T
ATGATGATATTGACAGGGAGCTCCGGAGCTTGGGTGACAGGGTATTGAGGCTGGTCTGGCCAAGTCATCGTGGCTCCA
1270 1280 1290 1300 1310 1320 1330 1340 1350

A P K V U R G T C Q I E R H A S E S P H F M R F H U A C P H C G
CGCCAAAGTGGAGGCATCTGCGAGTGGCGTGCAGGCCAGTGGATACCCCGCATTTATGCTGGCTTCCGGCGATGG
1360 1370 1380 1390 1400 1410 1420 1430 1440

A E Q Q Y L K F G D K E T P F P G L K W T P D D P E S U F Y L C
GGGGAGGAGGAGTATCTAAATGGCAGAACAGAGCCGGCTGGCTCATGGCCGATGAGCTGGGGTTTTTATCT
1450 1460 1470 1480 1490 1500 1510 1520 1530

A E H N A C U I R O Q E L D F T D A R Y I C E K T G I W T R D
GGCGCATATGCTGGCTCATCCGCCACAGGGAGCTGGGACTTACTGATGGCCCGTTATCTCGCGAAAGACGGGATCTGGACCG
1540 1550 1560 1570 1580 1590 1600 1610 1620

A G I L W F S S S G E E I E P P D S U T F H I H T A Y S P F T
ATGGCATTCCTGGTTCTGGCATCCGGTGGAGAGATGGCCACCTGGAGCTGGAGCTGTGGCTTCACATGGACAGCG
1630 1640 1650 1660 1670 1680 1690 1700 1710

A T H W U Q I K D H M M K T K G D T G K R K T F U N T L G E T
CCACCTGGCTGGAGATTTGCAAGAGCTGGATGAAAGAGGAGATGGGAAACCTGGCTTACATGGAGCGCTGG
1720 1730 1740 1750 1760 1770 1780 1790 1800

A H E A K I G E R P D A E U M A E R K E E H Y S A P U P D R U A
CGTGAGGGAGGAAATGGCAAGCTGGCGATGGAGCTGGAGCTGGGGAGGAGCTTACCGCCGGCTTCGACCGTGG
1810 1820 1830 1840 1850 1860 1870 1880 1890

A Y L T A G G I D S Q L D R Y E M R U W G H G P G E E S H L I D
CTTACCTSACCGCCGGTATCGACTCCAGCTGGCCGATCGAACATGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
1900 1910 1920 1930 1940 1950 1960 1970 1980

A R Q I I M G R H D D E Q T T L L R U D E A I N K T Y T R R N G
ACCGCGAGATTATGGCCGGACAGGAGCTGAAGAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
1990 2000 2010 2020 2030 2040 2050 2060 2070

A A E M S I S R I C H D T G G I D P T Y E R S K K H G L F
GTGAGAACATTCGGATATCCGGATCTGGGGATCTGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
2080 2090 2100 2110 2120 2130 2140 2150 2160

A R U I P I K G A S U Y G K P U A S M P R K R N K N G U Y L T
TCCGGGGATCCCGATTAAGGGCCATCGCTGAGGCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
2170 2180 2190 2200 2210 2220 2230 2240 2250

A F I G T D T A K E Q I Y N R F T L T P E G D E P L P G A U H
CCGAAATGGTACGGATACCGCGGAAAGGAGCAGTGGATACCGCTGGAGGGGGGGGGGGGGGGGGGGGGGGGGGG
2260 2270 2280 2290 2300 2310 2320 2330 2340

A F P N N P D I F D L T A Q Q L T A E E Q U E K H U D G R . K
ACTTCCCGAATACCCGGATATTGGTACTGACCGGAGCGCAGCTGGAGCTGGAGAGGAGGGGGGGGGGGGGGGGGGG
2330 2340 2350 2360 2370 2380 2390 2400 2410 2420 2430

A K I W L D S K K R R N E A L D C F U Y A L R A L R I S I S R
AAAAATACTGGGGAGACAAAGGAGCAGCAGCTGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
2440 2450 2460 2470 2480 2490 2500 2510 2520 2530

A H O L D L S A L L A S L Q E E D G G A A T T N K K T L A D A R
GCTGGCAGCTGGAGCTGGCTGGCTGGCGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGG
2530 2540 2550 2560 2570 2580 2590 2600 2610

FIG. 2. Nucleotide sequence of lambda DNA and the most probable amino acid sequences of proteins it codes for. The sequence given is for $\lambda c\text{lindlts857Sam7}$; mutations from the wild type are at 37,589 ($c\text{lindl}$), 37,742 ($c\text{lts857}$) and 45,351 to 45,353 (Sam7). The sequence is in 3 sections according to the main directions of transcription (see Fig. 1): (a) the left arm; (b) the right arm; (c) the central region. Transcription and translation are shown from left to right throughout. The numbering is from left to right of the whole DNA and consequently goes backwards in the central region. The individual gene

W A L S G E D E M T R Q E E L A A R A A A L H D L M T G K R V
 GTGCCCTTATCGGAGGGATGATGACCGCAGAAGGAGAHTTGCCGCTGCCGTGCGCAGCTGCACTGATGACAGGAAACGGGT
 2620 2630 2640 2650 2660 2670 2680 2690 2690 2700
 W P I U Q K D D R P U E F T A T T S U S D L K K Y I A E L E U Q
 GGCAACAGTACGAAAGAGCGGCGAGGGTGGAGGTTACGGCCACTTCGGTCTGACCTGAAGAAATAATATGAGCTGGAGTGC
 2710 2720 2730 2740 2750 2760 2770 2780 2780 2790
 B M K T P T I P T L L G P I G M
 GACCGGCGATGACGACGGCGAGGGAGCTGAGATTATGATGAAAHCGCCACCCATCCCACCCITCTGGGGCCGGACGGCATG
 2800 2810 2820 2830 2840 2850 2860 2870 2870 2880
 B T S L R E Y A G Y H G G G S G F G Q L P S W N P P S E S U
 ACATCGCGCGCAATATGCGCGGTTACGGGAGGATTTGGAGGCGATGAGGGTGTGGGAGGACCCACGGAGTGAAGTGTG
 2890 2900 2910 2920 2930 2940 2950 2960 2960 2970
 B D A A L P N F T R G N A R A D D L U R N N G Y A A A N A J A
 GATGCAGCCCTGTCGCCAACCTTACCGGAGATGGCGCGAGCGATGCTGGACGCAATACCGCTATGCGCCCAACGGCATCG
 2980 2990 3000 3010 3020 3030 3040 3050 3060 3060
 B L H O D H I V G S F F P R L S H R P S H R Y R L G I G E E E A R
 CTGCGATCGGATCATCTGCGGCTTCTGGCGCTAGTCGCCCCAGCTGGCGCTATCGGCCATCGGGGAGGAAAGGCCGCT
 3070 3080 3090 3100 3110 3120 3130 3140 3150 3150
 B A F S R E V E A A H K E F A E D D C C C C J D V U E R K R T F T
 GCCTTTCCGGCGAGGTTGGAGCCATGGAGAAGGGTGGAGGAGACTGCTGCTGCGATTGAGCTTGAGCAGGAAACCGCACGT
 3160 3170 3180 3190 3200 3210 3220 3230 3240 3240
 B M M I R E G U V A M H A F N G E L F U O D T A W D T S S E R L F
 ATGATGATTCGGGAAGGGTGTGGCCATGAGCGCTTATGAGCTGCTGCTGAGGCACTGGGATACCAAGTTGCTGGCGCTTTC
 3250 3260 3270 3280 3290 3300 3310 3320 3330 3330
 B R T Q F R M U S P K R A S N P N T G D S R N C R A G U Q I
 CGGACACAGTCCGGCGATGGTCHGGCGAGGCGCATCGCAACCCGAGACATCCGGCGAGCGGAGACTGCGCTGGCGTGTGAG
 3340 3350 3360 3370 3380 3390 3400 3410 3420 3420
 B N D S G A A L G Y Y U S E D G Y P G H M P O K K W T H I P R E
 ATAGCAGCGGGTGGCGCGCTGGGGATHATAGCTGAGCGAGGCGATCTGCGCTGAGGGCAGAATGGAGCATGGAAACCCGGTAG
 3430 3440 3450 3460 3470 3480 3490 3500 3510 3510
 B L P G G R A S A F I H U F E P U D G O T P G G A N U F Y S U M
 TTACCCGGGGGGCGCCGCTGCTGATCATCGAGGCGCTGGAGGAGCGAGACTGCGCTGGCGTGCAGGTTGCTAGCGGTGATG
 3520 3530 3540 3550 3560 3570 3580 3590 3600 3600
 B E O M K M A L D T L Q N T O L O S A I U K A M Y P A T F E S E
 GAGCAGATGAGATGAGCTGGAGAGCGCTGGAGAGCGCTGGAGAGCGGATGTTGAGAGCGATGATGCCGACCATGGAGCTGG
 3610 3620 3630 3640 3650 3660 3670 3680 3690 3690
 B L D T O S A C M D F I L G A N S Q E U R E R E L T G W H I G E I A
 CTGGATAACCGAGTCAGGCGATGGATTATTCTGGCGGAGCTGAGGGAGCGGGGGAGCTGACCGCTGGATTGGTAATGGCC
 3700 3710 3720 3730 3740 3750 3760 3770 3780 3780
 B A Y Y H A A P V U R L G D A K U P H L M P G D S L N L O T A O
 GCGTATTACGGCGAGCGCCGGTCCGGTGGAGGAGGAGHAGTACCGACCTGAGTGGCGGTGACTCACTGAACCTGAGACGCGCTCG
 3790 3800 3810 3820 3830 3840 3850 3860 3870 3870
 B D T D N G Y S U F E Q S L L R Y I A A G L G U S Y E Q L S R
 GATAACGGGAGCGCTACTCGTGTGGAGCAGTCGCTGCTGCTGAGGCTGGGGTGTCTGCTATGAGCGCTTCCCG
 3880 3890 3900 3910 3920 3930 3940 3950 3960 3960
 B N Y A O M S Y S T T A R A S A N E S H A W A Y F M G R R K F U A S
 ATTACGCCAGATGAGCTACTCCAGGGAGCGGAGCTGAGCTGAGGCTGGGGGAGTATGGGGCGGCAAAATTCTGCTGCATCC
 3970 3980 3990 4000 4010 4020 4030 4040 4050 4050
 B R Q A S Q M F L C H L E E I V R R U V T L P S K A R F S F
 CGTCAGCGCGAGCGGAGCTGTTCTGCTGCTGGAGGAGGAGCATCTGCTGCGCTGGGGTGTACCTTCAGGAGCGCTTCAGTT
 4060 4070 4080 4090 4100 4110 4120 4130 4140 4140
 B Q E A R S A W G H N C D H W I G S G R M A I D G L K E U Q E A U
 CAGGAAGCGCGAGCTGGGGGGAGCTGGAGCTGGGGCTGGGGCTGGGGCTGGGGCTGGGGCTGGGGCTGGGGCTGGGGCTGGGG
 4150 4160 4170 4180 4190 4200 4210 4220 4230 4230
 B M L I A E A G L S T Y E K F C A K R G D D D Y O E I F A Q Q U R
 ATGCTGAGAGGAGCGCGAGCTGAGGAGACTACGAGAGAGCTGGCAAAACCGGGAGCAGCTACGAGGAAATTGGCCAGGAGTCCG
 4240 4250 4260 4270 4280 4290 4300 4310 4320 4320
 B E T M E R R A A G L K P P A H A A F E S G L R Q S T E E
 GAAACGATGAGCGCCGGTGCAGCCGGCTTAAKCCGCCGCTGGGGCGCTGGAGCTTAAATGCCGCGCTGCAGCATCACAGAGGAG
 4330 4340 4350 4360 4370 4380 4390 4400 4410 4410
 B E K S D S R A E A * L P H I P A S M A H F N E P L M L E P A R
 GAGAAGAGTGAAGAGCGCTGGCTTAACTGGCGATTTACCGAGCTTAAATGAGGGCTGGCTGGCTGGCTGGCTGGCTGGCTGG
 4420 4430 4440 4450 4460 4470 4480 4490 4500 4500
 C U F F C A L A G O L G I S S L T D P A U S G D S L T A O E A L
 GGTTTCTTGTGGCTTGGCTTGGAGGAGCTGGAGGAGCTGGAGGAGCTGGAGGAGCTGGAGGAGCTGGAGGAGCTGGAGGAGCT
 4510 4520 4530 4540 4550 4560 4570 4580 4590 4590
 C A T L A L S G D D D G P R Q A R S Y P V U M N G I A V L F U S
 CGCGACGCTGGCATTCAGGTGATGAGACGACGGAGCAGGGCCGAGTATCAGGGCTGAGCTGAGGAGCTGGAGGAGCTGG
 4600 4610 4620 4630 4640 4650 4660 4670 4680 4680
 C G T L U S R T R A L Q P Y S G M T G Y N G I I A R L Q O Q A R
 CGCGACGCTGGCATTCAGGTGATGAGACGACGGAGCAGGGCCGAGTATCAGGGCTGAGCTGAGGAGCTGGAGGAGCTGG
 4690 4700 4710 4720 4730 4740 4750 4760 4770 4770
 C S D P M U D G I L L D M D T P G G M U A G A F D C A D I I A
 CRGGATCGGAGCTGGAGCGAGCTGGAGCGAGCTGGAGCGAGCTGGAGCGAGCTGGAGCGAGCTGGAGCGAGCTGGAGCG
 4780 4790 4800 4810 4820 4830 4840 4850 4860 4860
 C R U R D K P U N A N D M M N C S A G Q L L A S A A S R R
 CCGTGTGGCTGACATAAACCGGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAG
 4870 4880 4890 4900 4910 4920 4930 4940 4950 4950
 C L U T D T A R T G S I G U M M A H S N Y G A A L E K P G U F
 TCTGGTCAAGGAGCGCCGGAGCGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAG
 4960 4970 4980 4990 5000 5010 5020 5030 5040 5040
 C A T C A G G T G A T T A C C G G A C C T A G G T G A C T G G C T A C G G C T G G G G A C T G C A G T C C C G
 5050 5060 5070 5080 5090 5100 5110 5120 5130 5130

products or reading frames (*orf*) are identified in the left-hand margin. Some of the more important control regions are also marked above the sequences. The start points of the main transcription products are indicated as follows: *s_R*, start site for *p_R* RNA; *s_E*, for *p_E* RNA; *s_K*, for *p_K* RNA; *s_I*, for *int* message; *s_L*, for *p_L* message. In some genes the position of initiation is uncertain and the most probable has been chosen (see Table 3 and the text).

M D A T R O M F H G N T U S L S U Q U U L D T E A A V
GATGGCGGAGCCGCGGCTTGTGGCGAGHAGGTTTCGCGCATATAACGGCGCTGCTGGCGAGGTGCTGCTGAGTACCGAGGGCTGCAGT
5140 5150 5160 5170 5180 5190 5200 5210 5220

C Y S G Q E I D A G L A D F L V N S T T U S A T U M R D A L D
GTACHGCGGTAGGGGGCAGCTTGTGGCGGCTGCGCTGAGCTTGTTAAGCGACCGATGCGATCACCGTCATGGGTGATGCACTGGAA
5230 5240 5250 5260 5270 5280 5290 5300 5310

C A R K S R L S G G R M T K E T O S T T U S A T A S Q A D U T
TGCGCGTAAHTTCGCTCTCTCGAGGGGGCGAGTGGCGAGAGACTAACAGACCTGTTTCHGCGACTGCTTGGCGAGCGTACGCTTAC
5320 5330 5340 5350 5360 5370 5380 5390 5400

C D Y U P A T E G E M H S A A Q P D U N A Q O I C T A A Y A A E N
TGACGTGGTGCAGGCGAGGGGGAGGGAGGGCGAGCGCGAGCTGAGCGAGATACCGCGAGCGGTGAGCGAGAAA
5410 5420 5430 5440 5450 5460 5470 5480 5490

C S R I H G I L N C E E H H G R E E Q A R M A E T P G M T U
CAAGCGCTTATGGGGATTCAGTCTGAGGGGGCTGAGCGAGCGAGCGAGCTGGCGAGAACCCTCCCGTATGAGCGT
5500 5510 5520 5530 5540 5550 5560 5570 5580

C K T A P R R I L A A P S Q A P S D T P D R L M Q G A P
GAAAGCGGCCCGCCGCATTCGCGCCAGCGACCHAGGTGCGCAGTGAGACTGCGCTGGATCTGATGAGCGAGGGGAC
5590 5600 5610 5620 5630 5640 5650 5660 5670

C D P P L M A G N P V S D A H N D L P U * M T S K E
GGCAGCGCTGGCTGGHGGTAGCTGGCGATCTGATGCGCTTACGATTTGCTGAGCACACCGAGTGTGAGGGATGTTTATGACGAGCGA
5680 5690 5700 5710 5720 5730 5740 5750 5760

D T F T H Y D P Q G N S D P A H T A T A P G G L S A K A P A M
AACCTTACCCATTACCGACCCAGGGGAGAGCTGAGCTGGCGCTCATACCGAACCGCCGGCGGATTGAGTGGAGCTGCGCTGC
5770 5780 5790 5800 5810 5820 5830 5840 5850

D T P L M L D T S S R K L U A H D P G T D B G P A U V I L A U
GCCCCCTGATGCTGGAGACCTCCAGCGAGCTGAGCTGGGGGAGGGCTGCTGGCGTTGGAGGATGTC
5860 5870 5880 5890 5900 5910 5920 5930 5940

D A D Q T S T L T F Y K S G T F R Y E D U L W P E A A S D E
TGCTGACGAGGACACCGACCTGAGCTGAGCTGAGCTGGGGAGGGCTGCGAGCGAGCG
5950 5960 5970 5980 5990 6000 6010 6020 6030

D T K K R T A F P A G T A I S I U *
GACGAAAGACGGACCGCGCTTGGCGGAGCGCATCGTTACCTTACCTTACATCACTAAAGGGCGCCCTGTGGGCTT
6040 6050 6060 6070 6080 6090 6100 6110 6120

E M S M Y T T A H O L A A H E D O K F K F D P L F L R L
TACGGGATTTTTTATGTCGATGTCACAGCGGAGCTGAGCTGGCGGAGATTTAGTGGCTGCTGGGCTG
6110 6120 6130 6140 6150 6160 6170 6180 6190 6200 6210

E F F R E S Y P F T T E K U V Y L S O I P G L U N I M A L L Y U S P
TCTTTTCCGTGAGGAGCTTCCACCCGGAGGAAGCTACATCTCACAATTCGGGGACTGGAAACATGGCGCTGTACGTT
6220 6230 6240 6250 6260 6270 6280 6290 6300

E I U S G E V I R S R G S T S E F T P G L U K P K H E U N P
CGATGTTCCGGTGGGGTATCCGTTCCGGCGCTACCTCTACGGGGATATGCAAGCGAAGCATGAGGTGAGT
6310 6320 6330 6340 6350 6360 6370 6380 6390

E O M T L R R L P D E D P O N L A P D P A Y R R R R I I M O N M
CGCAGATGACCGTGGCCTUCGGGATGAGAGATCGGGGAGCCGGCTTACCGCGCCGGCTCGCATCATGAGCGAGCA
6400 6410 6420 6430 6440 6450 6460 6470 6480

E R D E E L A I A P D V E E M Q A U S A U L K G K Y T M T G E A
TGGTGCACGAGAGCTGGGGCATIGCTGAGCTGGAGAGATGAGCTGGAGLAGTCTGGCGTCTTAAGGGCAAACTACACCGAGCG
6490 6500 6510 6520 6530 6540 6550 6560 6570

E F D P U E U D M G R S E E N N I T Q S G G L T E H S K R D K S
CCTTCGATCCGGTTGGGTGAGTGGCTGGGGCGAGGAGATACACGAGCTGGGGGAGCTGGAGGAGTGGAGGAGCGAG
6580 6590 6600 6610 6620 6630 6640 6650 6660

E T Y D P D P D I E A Y A L N A S G U U N I I U F D P K G H A
CCACGTGACCGAGCCGGAGATCGAGGAGCTGGCTGGGGCTGGGTGGTGAATATCATCGTGTGCTGAGTGGAGGAGAGCTGG
6670 6680 6690 6700 6710 6720 6730 6740 6750

E L F R S F K A U K E K L D T R R G S N S E L E T A U K D L G
CGCTGTCGGCTCTTCAGACGCGTCHAGGGAGAGCTGGATACCCGGCTGGCTCTAATTCGGAGCTGGAGACAGGGTGAAGAGCTGG
6760 6770 6780 6790 6800 6810 6820 6830 6840

E K A U S K G M Y D U A I U V Y S G Q Y U E N G U K K N F
GCAAGGGGTGCTTACAGGGGATGAGTGGCTGGGGCATCTGCTGCTGATTCGGGAGCTACGGTGGAGGAGACCGCTGAGAG
6850 6860 6870 6880 6890 6900 6910 6920 6930

E L P D N T M U L G N T A R G L R T Y G C I Q D A D A Q R E
TCCGGGGGACACGATGGCTGGGGAGACTCAGGGCGCGCTGGCGCATCTGGGGCATCTGGGGAGGGAGCG
6940 6950 6960 6970 6980 6990 7000 7010 7020

E G I H A S A R Y P K N H U V T G D P P A R E F T M I Q S A P L
AAGGCATTAACGCGCTTGGCCGTTAACCGGAAACTGGGTGAGCCACCGGGGATCTGGGGCATCTGGGGAGTGGAGGAT
7030 7040 7050 7060 7070 7080 7090 7100 7110

E M L L A D P D E F U S V Q L A *
TGATGCTGGCTGGCTGAGCTGAGCTGAGCTGAGCTGGGGCATCTGGGGCATCTGGGGAGGGAGTC
7120 7130 7140 7150 7160 7170 7180 7190 7200

FII M T K D E L I A R L R S L G E Q L N R D U S L T G T K E E L
CATGACGAAAGATGAGCTGGCTGGCGCTGGCTGGGGCTGAGCTGGAGGAGCTGGATGAGCTGGGGAGGGAGAG
7210 7220 7230 7240 7250 7260 7270 7280 7290

FII A L R U A E E L K E E L D D T D E T A G Q D T P L S R E N V L
GGCGCTCGTGTGGCGAGAGCTGAGAGGGAGCTGGATGAGCTGGGGAGACCCCGCTCGAGCGGGAGAATGTGCT
7300 7310 7320 7330 7340 7350 7360 7370 7380

FII T G H E N E V G S A Q P D T U I L D T S E L V U T V U A L U K
GACCGGGACATGAGAGATGGGGAGGGGGAGGGGGAGGGGGAGGGGGAGGGGGAGGGGGAGGGGGAGGGGGAGGGGGAGGGGG
7370 7380 7390 7400 7410 7420 7430 7440 7450 7460 7470

FII L H T D A L H A T R D E F U L P G T A F R U S A G V A
GCTGCGATCTGGCTGGGGAGGGGGAGGGGGAGGGGGAGGGGGAGGGGGAGGGGGAGGGGGAGGGGGAGGGGGAGGGGGAGGGGG
7480 7490 7500 7510 7520 7530 7540 7550 7560 7570

FII A E M T E R G L H R M Q *
AGCCGAATGAGAGGGCGGGCTGGCCAGAGCTGGCAATACGGGGGGCGCTGGGGCTGGGGCTGGGGAGGGGGAGGGGGAGGGGG
7570 7580 7590 7600 7610 7620 7630 7640 7650

FII R A D E T I R G Y M G T S A T I T S G E Q S G A U I R G U F
CGGCCGCGATGAGAGGGGGAGGGGGAGGGGGAGGGGGAGGGGGAGGGGGAGGGGGAGGGGGAGGGGGAGGGGGAGGGGGAGGGGG
7660 7670 7680 7690 7700 7710 7720 7730 7740

FII D D P E H I S Y A G Q G U R U E G S S P S L F U R T D E V R
 7750 GATGACCTGAAATAATCAGCATATGCCBAGAAGGGCGCTGGAGCTGCAGCCGCTGTGTTGTCGGGACTGATGAGGGTCGCG
 7760 7770 7780 7790 7800 7810 7820 7830
 FII Q L R R G D T T I G E E N F H U D R U V S P D D G G S C H L
 7840 CAGCTGCGGCCTGGAGACACGCTGAGCTGGGATGATGGGTTTCCGGGATGATGGCGGAAAGTTGTCATCTC
 7850 7860 7870 7880 7890 7900 7910 7920
 FII W L G R G V U P P A V U H P R R * M A I K G L F Q A U E N
 Z 7930 TGGCTGGGGGGGGGTGACCGCTGCGCTTHACCGTCGCCGCTGAGHAGGGGGATGTATGGGATATAAAAGGCTCTGAGCHGGCGITGAA
 7940 7950 7960 7970 7980 7990 8000 8010
 Z L S P I S K T A U V P A A M A I S Q S A I S Q S A S
 ACCCTAGGCTACGCAAGGAGCTGGGCTGGTGGCGCGCAGATGGCCATTAACTGGCTTGCTTCATCGCGATATCAGTCGGCGT
 8020 8030 8040 8050 8060 8070 8080 8090
 Z Q U A R E T K U R R K L I U K E R A P L K R A T U K N P Q A R
 8110 CACAGGTTGCCGCTGGAGAACAGGTHACGCCGAGCTGGAGHAGGGCACCGTCAAAAATCCGAGCGCGA
 8120 8130 8140 8150 8160 8170 8180 8190
 Z I K V N R G D L P U I K L G N A R H U V L S R R R R R K K G Q
 8210 GAATCAAAGTAAACCGGGGATTTGGCCGTHATCAACGGTGGGATGGTAATGGCGGGTTGCTTTCGGCCCGCAGGGCTCGTAAGGAGGGC
 8220 8230 8240 8250 8260 8270 8280 8290
 Z R S S L K G G G S S U L V U G N R P I P I Q Q L K N G R
 8310 AGCGTTCTCCCTGAHAAGGCTGGCGAGCGCTGGCTGTTGAGCTGCTGATTCAGCGCTTATCAGCAACTGAAAGGGCTGGCG
 8320 8330 8340 8350 8360 8370 8380 8390
 Z W H M Q R U H G K N R Y P I D U V U X I P M A V P L T T A F
 8410 GGTGGCATGTCAGCAGCTGGGGAAHGGCTGGCTTACCGGATGAGTGGGGTGGCCGCTGACACCGCGT
 8420 8430 8440 8450 8460 8470 8480 8490
 Z K Q N I E R I R E R L P K E E L G Y I L Q H O L P R M V I K R
 8490 TTAACAAATTAACGGGAAHGGCGAGCTGGAGACGCTGAGCTGAGCATCAACTGAGGATGGTAATTAAGG
 8500 8510 8520 8530 8540 8550 8560 8570
 U M K H T E L R H A M L D A H L E K H D T G A T F F D G R P A V
 8580 GATGAAACATAGTAACTCAGCTGGCAGCGCTACTGGATCAGTGGAGAGGATGAGACGGCTGGCCCGCTGT
 8590 8600 8610 8620 8630 8640 8650 8660
 U F D E A D F P A U A U V Y L T G A E Y T G E E L D S D T H W P A
 8670 TTTTGATGGGGGAAITTTCCCGGAGCTGGCGCTTACCGGCGCTGAGATACAGGGGAGAGCTGGGCAAGGATACCTGGCGAG
 8680 8690 8700 8710 8720 8730 8740 8750
 U . E L H I E U F L P A Q U P D S E L D P H H M E S R I Y P U M S
 8760 GGAGGCTGATATCAGGAGTTTCTCGCTGAGGCTGGGATCAGGCTGGATGAGTGGGATGGAGTCCCGGATTATCCGGTGTAG
 8770 8780 8790 8800 8810 8820 8830 8840
 U D I P A L S D L I T S N U V A S G Y D Y R R D D A G L H S S
 8850 CGATATCCGGGAGCTGAGATTGATCACGGCTGGGATGGCTGAGCTGGGAGGAGATGATGGGGCTTGTGGGAGTTC
 8860 8870 8880 8890 8900 8910 8920 8930
 U A D L T Y V I T Y E H * M P U P N P T I M P U K G A G T T
 8940 AGCCGATCTGACTTATGTCATACCTATGAAATGTCGGGRCGCTATGCCCTGACCATCAATGCCCGGTGAAAGGTCGGGACCA
 8950 8960 8970 8980 8990 9000 9010 9020
 U L H U Y K G D P P Y A N P L S D U D P H T A T G C K U K D L T
 9030 CCCGTGGGTTTATAAGGGGAGCGGTGAGCTTCCGAGCTGGCTGAGCTGGCTGCTGGCAAAAGGTTAACAGCTGA
 9040 9050 9060 9070 9080 9090 9100 9110
 U P G E L T C A E S Y P D S Y L D D E D P H T A T G C K U K D L T
 9110 CGCCCGGGAGACTGAGCTGGCTGAGCTGAGCTGGGAGGAGCAGGGAGGAGCTGGGAGCTGGGAGGAGCTGGGAG
 9120 9130 9140 9150 9160 9170 9180 9190
 U A Y K I R F P N G T U D U V F R G H U S S I G K A U T A K E V
 9290 GTGCCATTAACCGCTTCCCGAACGGCACGGCTGAGTGTGTCGGCTGGGTCAGCAGATCGGTAAGGGCGGTGACCGAAGGGAG
 9300 9310 9320 9330 9340 9350 9360 9370
 U I T R T U K U T H U G R P S M A E D R S T U T A A T G M T U
 9380 TGATCCCCCAGCTGGAGACGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAG
 9390 9400 9410 9420 9430 9440 9450 9460
 U T P A E S T S U U K G O S T T L T U A F O P E G U T D K S F R
 9460 TGACGCCCTGGAGCAGCTGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAG
 9470 9480 9490 9500 9510 9520 9530 9540
 U A U S A D K T K A T U S U S G M T I T U N G U A A G K U N T
 9550 GTGCCGTGTGGAGTACACCGCTGGCTGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAG
 9560 9570 9580 9590 9600 9610 9620 9630
 U P U S S G N G E F A A U E I T U V I S * M F L K
 9640 TTCCGGTGTGAGCTGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAG
 9650 9660 9670 9680 9690 9700 9710 9720
 G T E S F E H N G U T U T L S E L S A L O R I E H L A L M K R
 9730 AAACCGAACATTTGGAGCATACGGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAG
 9740 9750 9760 9770 9780 9790 9800 9810
 G Q H E D O A E S D S N R K F T U E D A I R T G A F L U A M S L
 9820 GGCAGGCGAGACAGGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAG
 9830 9840 9850 9860 9870 9880 9890 9900
 G H N H P O K T Q M P S M N E A U V K O I E O E U L T T H P T
 9910 TGTTGGCATACCCATCGCGAGAGACGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAG
 9920 9930 9940 9950 9960 9970 9980 9990
 G E A I S H A E N U U Y R L S G M Y E F U U U N N A P E Q T E D
 10000 CGGAGGCAATTCTCATGTCAGTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAG
 10010 10020 10030 10040 10050 10060 10070 10080
 T G A G P A E P U S A G K C S T U S * L S F H L K L A R E M G R P
 10090 ACCTCCGGGCCGAGGGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAG
 10100 10110 10120 10130 10140 10150 10160 10170
 T D W R A M L A G H S S T E Y A D W H R F Y S T H Y F H D U L
 10180 CGACTGGCGCTGGCATGCTGGAGGATGTCAGTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAGCTGGAG
 10190 10200 10210 10220 10230 10240 10250 10260

H A G I U H R G E F V F T K E P A T S R I G U G N L Y R L M R G
CAGCGGGGATTTGTCACCGGGTGAAGTTGCTTCACGGAGGAGCAGCAACCGGAGTGGCGTGGGGAAATCTTACCGGCATGATGCCG
12790 12800 12810 12820 12830 12840 12850 12860
H Y A T S G X V U G T P G S M A D S R S S Q A S G T F E Q N N H V
GCTATGCAACCGGGCGGTTHGTCGGTACCCGGGAGCAGCAACGGCGTGGCGAGCGCTGGGACGTTTGAGCAGATAACCATG
12880 12890 12900 12910 12920 12930 12940 12950
H U I N H D G T N G Q I G P A P L K X V Y D M A R K G A R D E
TGGTGATTACAGCHGCGCACGAGGAGTGGCGGCTGCTGAGGGGTGATGACATGGCCCGAAGGGTCCCCGTGATG
12970 12980 12990 13000 13010 13020 13030 13040
H I G T Q M R D G G L F S G G G R M K T F R W K V U K P G M D U
HATTCAGACAGAGTGCCTGHTGGTGGCTGCTGGAGGGTGAAGATGAGAGCTTCCGCTGGAAAGTGAACCCGGTATGGATG
13060 13070 13080 13090 13100 13110 13120 13130
M A S U P S U R K U R F G D G Y S Q R H P A G L N A N L K T Y
GGCTTGGTCCCTCTGTHGAGGGTGGCGCTTGGGTTGGCTHTCTCGAGCGGCCGCTCCGGCTGAATGCCAACCTGAAACAGTA
13150 13160 13170 13180 13190 13200 13210 13220
M S U T L S U P R E E H T A T U L E S F L E E H G M K S F L H T
CAGGGTACGACCTTCTGTCCTCGAGGGGGCACGGTACTGGAGCTTCTGGAAAGAGCACGGGCTGGAAATCTTCTGTCGG
13240 13250 13260 13270 13280 13290 13300 13310
M P P Y E H R O I K V U T C A K W S S R U S M L R U E F S A E F
GCCGCCATATGAGTGGGGCACGGTAAHAGGTGACCTGCGCAAAATGTCGCGGGTCAGHTGCTGGCTGTGAGTTCCGGCAGATT
13330 13340 13350 13360 13370 13380 13390 13400
L M O D I R Q E T L N E C T R A E Q S A S V U U L W
TGACGAGCTGGTGHACTGHTGAGGNTTCGGCGNAGGAGACACTGATGAGATGACCCGGCGAGCTGGCCAGCGTGGTGTCTGG
13420 13430 13440 13450 13460 13470 13480 13490
L E I D L T E U G G E R Y F C N E U N E K G E P U T H Q G R
GAATTCGACCTGAGGGTGGCTGGAGAGCGTTATTTTCTGTAATGAGCAGAACGAGAAAGGTGAGCCGGTCACCTGGCAGGG
13510 13520 13530 13540 13550 13560 13570 13580
L O Y Q P Y P I Q G S G F E L H N G K G T S T R P T L T U S H L
CACTGATGAGCGTACCCGATTCGGGGAGGGCTTGGAGATGAAATGCAAGGACAGCTGGCCACCGCTGACCGTTCTAACCTG
13600 13610 13620 13630 13640 13650 13660 13670
L Y G M U T T G M A E D M Q S L U V G T U V R R K U Y A R F L D
TACGGTATGCTTCGGGGAGAGATATGAGGGTCTGGTCGGCGAACGGCTGGCGCTGAAGGTTTACGCCCGTTTCTGGAT
13690 13700 13710 13720 13730 13740 13750 13760
L A U N F U N G N S Y A D P E Q E U I S R H R I E Q C S E L S
GCGGTGAGCTTCGTCACGGAAACGATACGGCGATCGAGGAGGTGATCAGCCGCTGGCGCAATTGAGCAGTGAGCAGACTGAGC
13780 13790 13800 13810 13820 13830 13840 13850
L A U S A S F U V L S T P T E T D G A U F P G R E H M L A M T C T
GCGGTGAGTGGCTCTTGTACTGTCACCGGAGAACGGATGGCTGTTTCCGGACGCTATEATGCTGGCCAACACCTGGCACC
13870 13880 13890 13900 13910 13920 13930 13940
L W T Y R G D E C G Y S G P A U A D E Y D Q P T S D I T K D K
TGACCTATCGCGGTGCGGGTGGCTTATAGCTGGCGCTGCTGGCGATGATGCGCCACCTGGCGATATCACGAGGATAAA
13960 13970 13980 13990 14000 14010 14020 14030
L C S K C C L S G C K F R N N U G N F G F L S I N K L S Q *
TCGAGCAATGGCTTCGGCAATTCGGCGATTCGGCGCTTGGCGCTTGGCGCTTGGCGCTTGGCGCTTGGCGCTTGGCGCTTGGCG
14050 14060 14070 14080 14090 14100 14110 14120
CCATGACACAGACAGATTCGGCGATTCGGCGATTCGGCGATTCGGCGATTCGGCGATTCGGCGATTCGGCGATTCGGCG
14140 14150 14160 14170 14180 14190 14200 14210
K M S P E D N L O A E M Q
GGGHAAGATATTTCUCCCTGCGTGAATATCTCCGGTGAACGGGGAGGTATTCCTGGTATGTCGCCGGAGACTGGCTGAGGAAATGCA
14230 14240 14250 14260 14270 14280 14290 14300
K G E I U A L U H S H P G G L P W L S E A D R R L Q U O S D L
GGGTGAGATTTGCGCTGGTCCACGGCCACCCGGGGTGGCTGGCTGGCTGGAGTGGCGCGCTGGCGCTGGCGCTGGCGCTGGCG
14310 14320 14330 14340 14350 14360 14370 14380
K P W H L U C R G T I H K F R C U P H L T G R R F E H G U T D
GCCGTGGTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGG
14410 14420 14430 14440 14450 14460 14470 14480
K C Y T L F R D A Y H L A G I E M P D F H R E D D H H R N G Q
CTGTTACACACTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGG
14500 14510 14520 14530 14540 14550 14560 14570
K N L Y L D N L E A T G L Y Q U P L S P A Q P G D U V L L C C F
GAATCTCATCTGAGATATGGAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
14590 14600 14610 14620 14630 14640 14650 14660
K G S S U P N H A A I Y C G D G E L L H H I P E Q L S K R E R
TGTTCTCATCTGAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
14680 14690 14700 14710 14720 14730 14740 14750
K M A A T H T L P L L A S P P G M A R I C L Y G D L Q R F
GTACGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
14770 14780 14790 14800 14810 14820 14830 14840
K G R R A I D L R U K T G A E A I R A I R A L A T Q L P A F R Q K L S
GGTCGCCACATGAGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGG
14850 14860 14870 14880 14890 14900 14910 14920
I D G H Y Q U R I A G G R D U S T S G L T A Q L H E T L P D G A
GACGGCTGGTACGGTACGGTACGGTACGGTACGGTACGGTACGGTACGGTACGGTACGGTACGGTACGGTACGGTACGGTACGG
14950 14960 14970 14980 14990 15000 15010 15020 15030
I U I H I U P R U A G H K S G G U P Q I U L G G A A A I A G S F
GTAATTCTATTTGTTCCGGAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
15040 15050 15060 15070 15080 15090 15100 15110 15120
I F T A G A T T L A A H G A A I G A G G G M T G I L F S L G A S M
TTTACGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
15130 15140 15150 15160 15170 15180 15190 15200 15210
I U L G G G V A O N L A P K H R T P R I Q T T D N G K Q N T Y F
GTGCTGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
15220 15230 15240 15250 15260 15270 15280 15290 15300

S S L D N M V A Q S N U L P U L Y G E M R G G S R U U S Q E
 TCCCTCAGGTTTACGGGAGATGTTCTGCCTGTACGGTGGGGTCACCGTGTTCTCAGGAG
 15310 15320 15330 15340 15350 15360 15370 15380 15390
 I S T A D E G D G G P U U U I G P
 ATCGGAGCCGAGCGAAGGGAGCGGTGGCTGGGTTGGTGGTGGCTGATGCAAAATGTTTATGTGAACCCCTCGCGGGGG
 15400 15410 15420 15430 15440 15450 15460 15470 15480
 J H G X G S K G H T P P R E A K D N L K S T Q
 TTGGTCAATTATGGAGCGTGAGGAGTGGGTAAGGGAGCACTAACCCCGCGGAAGCGAAGGACACCTGAAGTCACCGAG
 15490 15500 15510 15520 15530 15540 15550 15560 15570
 L L S U I D A S E C G P I E G P U D G L K S U L L N S T P U
 TTGGCTTGGTGTGATGCTGGTCTGGAGGGGGCTGGTGGGGCTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGG
 15580 15590 15600 15610 15620 15630 15640 15650 15660
 J L D T E G N T H I S G U T U V R H G E Q E Q T P P E G F E
 CTGGGAGCTGGGGAGATGGGGATGGGGATGGGGATGGGGATGGGGATGGGGATGGGGATGGGGATGGGGATGGGGATGG
 15670 15680 15690 15700 15710 15720 15730 15740 15750
 J S S G S E T U L G T E V K Y D T P I T R T I T S A N I D R L
 TCCTCCGGTCCGGAGGGGGCTGGGAGCTGGGAGCTGGGAGCTGGGAGCTGGGAGCTGGGAGCTGGGAGCTGGGAGCTGG
 15760 15770 15780 15790 15800 15810 15820 15830 15840
 J R F T F G V U O A L U E T S K G D R N P S E U R L L U Q I Q
 CGCGTTTACCTTCGGTGTACAGGGCTACTGGTGGAGGACCACTTCAGGGTGTACAGGGATCTGGGAGCTGGGAGCTGG
 15850 15860 15870 15880 15890 15900 15910 15920 15930
 J R N G G W U T E K D I T K G T T S Q Y L A S U U M G N L
 CGTAAACGGGCTGGGCTGGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGG
 15940 15950 15960 15970 15980 15990 16000 16010 16020
 J P P R P F N I R M R B M T P D S T T D P O L Q N K T L W S S Y
 CGCGCCGGCCCGCTTAAATTCGGATGGCGAGGATGGCGAGGAGCGAGGAGCGAGGAGCGAGGAGCGAGGAGCGAGGAG
 16030 16040 16050 16060 16070 16080 16090 16100 16110
 J T E I I D U K Q C Y P N T A L U G G U Q U D S E O F G S Q O U
 ACTGGAATTCGTTGATGCTGGCTACCGGAAACCGGACTGGTGGCTGAGGGACTCTGGGAGCTGGGAGCTGGGAGCTGG
 16120 16130 16140 16150 16160 16170 16180 16190 16200
 J S R N Y H L R G R I L Q U P S N Y N P Q T R Q Y S G I I W D G
 AGCGGTAATTTACHTCGCCGGCGCTHTCTCGAGGTGGCTGGCATACTACACGGGATCTGGGAGCGAGGGAGGGAGGG
 16210 16220 16230 16240 16250 16260 16270 16280 16290
 J T F K P A Y S H N M A H C L W D M L T H P R Y G M G K R L G
 ACCTTTAACCGGCACTACAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGG
 16300 16310 16320 16330 16340 16350 16360 16370 16380
 J A A D V D K H A L Y U I G Q Y C D Q S U P D G F G G T E P R
 GCCTGGCGGATGTGGATTAATGGGGCTGTATGTCATGGGGAGCTACTGGCAGGAGCTAGTGGGGGGGGGGGGGGGGGG
 16390 16400 16410 16420 16430 16440 16450 16460 16470
 J I T C N A R Y L T T Q R K A H D U L S D F F C S A M R C C M P U H
 ATCACCTGTTGCTGCTTGGGGCTGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
 16480 16490 16500 16510 16520 16530 16540 16550 16560
 J N G O T L F T U Q D P R S D K T W T Y H R S N U U M P D D G
 AACGGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGG
 16570 16580 16590 16600 16610 16620 16630 16640 16650
 J A P P F R Y S F S A L K D R H N A U E U N H I D P P N N G H E T
 GCCTGGCTTGGCTACAGGCTTACGGGGCTGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
 16660 16670 16680 16690 16700 16710 16720 16730 16740
 J A T E L U E D T Q A R Y G R N U V K T M D A F G C T S R G
 GCCTGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGG
 16750 16760 16770 16780 16790 16800 16810 16820 16830
 J Q A H R A G L H L I K T E L L E T Q T U D F S U V G A E G L R
 CAGGAGCACCCGCCGGCGCTGGCTGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCT
 16840 16850 16860 16870 16880 16890 16900 16910 16920
 J H U P G D V I E I C D P D Y A G I S T G G R U L A U N S O T
 CATGCTACCGGGCGCATGTTGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCT
 16930 16940 16950 16960 16970 16980 16990 17000 17010
 J R T L T L D R E I T L P S S G T A L I S ' L U D G S G N P U S
 CGGGAGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGG
 17020 17030 17040 17050 17060 17070 17080 17090 17100
 J U E V O S U T D G U K U V S P V U P D G U A E Y S U H E L K
 GTGGAGGGTTCAGTCCGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGG
 17110 17120 17130 17140 17150 17160 17170 17180 17190
 J L P T L R Q R L F R C U C S I R E N E D D G T Y T A I T A U Q H U
 CTGGAGGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGG
 17280 17290 17300 17310 17320 17330 17340 17350 17360
 J P E K E A I U D N G A H F D G E O S G T U H N G U T P P A U G
 CGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGG
 17390 17400 17410 17420 17430 17440 17450 17460 17470
 J H L T A E U T A D S G E Y Q U L A R W H D T P K U U K G U S F
 CACCTGGCCGGAGAGAGCTGGGGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAG
 17480 17490 17500 17510 17520 17530 17540 17550 17560
 J L L R L T U T A D D D G S E R L U S T A R T T E T T Y R F T Q
 CTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGGCTGG
 17570 17580 17590 17600 17610 17620 17630 17640 17650
 J A P A A P S R I E L T P G Y F Q I T A T P H L A U Y D P T U
 GCACCCGGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAG
 17660 17670 17680 17690 17700 17710 17720 17730 17740
 J Q F E W F S E K Q I A D I R Q V E T S T R Y L G T A L Y H
 CGTTGGTGTCTGGGTCTGGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAGGGAG
 17750 17760 17770 17780 17790 17800 17810 17820 17830
 J I A A S I N I K P G H D Y Y F Y I R S U N T U G K S A F V E
 ATAGCCGGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAG
 17840 17850 17860 17870 17880 17890 17900 17910 17920

J A U G R A S D P A E G Y L F F K G X I T E S H L G K E L L
GCCCTCGGTGGGGAGCAGTGTGCGGAAAGGTTACCTAAGTTTCAAGGCAAGATCCATCGGAAAGGACTGCTG
17920 17930 17940 17950 17960 17970 17980 17990 18000

J E K U F E L T E D N A S R L E E F S K E H K D A S D K W N A M
GAAGAACTCGAGCTGGGGGGATACCGCAAGAGTCGAGGTTTCGAAAGAGTGGAGGGATGCCAGTGTAAAGGAAATGCCATG
18010 18020 18030 18040 18050 18060 18070 18080 18090

J W A U K I E T K D G K H Y U A G I G L S M E D T E E G K L
TGCCGCTGGCTTAATGAGCAGGAGCCAAAGGGCAAGAACATTATGTCGCGGTTGGCTCAGCATGGGGACACGGAGGAAGGCAAACCTG
18100 18110 18120 18130 18140 18150 18160 18170 18180

J S Q F L U A R A F I D P A N G N E T P M F V A Q G N G
AGCCAGTTCTGGTGGCGCAGTCGATTATGACCCGGCAACGGGAGTAAACGCGGATGTTTGTGGCCGAGGGCAACCCG
18190 18200 18210 18220 18230 18240 18250 18260 18270

J L F M N D U F L K P L T A P T I T S G G N P P A F S L T P D
ATATTCATGAGCGAGCTGTTCTGAAGCGCTGAGCCGGGCCCCCATACCGCGCCGAACTCTCCGGGCTTTCGGCTACACCCGGAC
18280 18290 18300 18310 18320 18330 18340 18350 18360

J G K L T A K N A D I S G S U N A N S G T L S N U T I A E N C
GGGAAAGCTGGCGGTAAAGGATATCGAGCTGAGTGAATCGAGCTCGGAGCTGAGTAAAGTGAACGATAAGCTGAAAGACTGT
18370 18380 18390 18400 18410 18420 18430 18440 18450

J T I N G T L R E K J U V G D I U K A P S A A F P R Q R E S S
ACGATAAACGGTACGGCTGAGGGGGAAAGAAAGTCGCGGGACATCTGAAAGGGCGGAGCGGCGCTTCCCGCCAGCGTGAAGCAGT
18460 18470 18480 18490 18500 18510 18520 18530 18540

J U D W P S G T R T U T U T D H P F D R O I U V U L P L T F R
GTGGACTGGCGTAGGGTACCCGACTGTCACCGTGGAGGCGACATCTGAGTGCCTTGTGCGCTGAGTGTGGCTCCCGCTGAGCTTTCG
18550 18560 18570 18580 18590 18600 18610 18620 18630

J S S K R T U S G R T T Y S M C Y L A V U L M N G A V I Y D G A
GGAGTAAAGCTGAGCTGTCAGGGCAGGACACCTTCTGATGTTGAGTACTGATGACCGGGCGGTOATTTATGATGGCG
18640 18650 18660 18670 18680 18690 18700 18710 18720

J A N E A U U Q F S P R I V D M P A G R G R H N U I L T F T L T S T
GGCHACBGGCGGTACAGGGTCTCCCGTAACTGTCAGCGGGCTGGGGAAACGGTACGCTCACGCTTACGCTTACGCTTACGCTTACGCT
18730 18740 18750 18760 18770 18780 18790 18800 18810

J R H S A D I P P T F A S D U Q V U M V I K K G A L G I S U U
CGGCATTCGGCGAGATACTCCGGCTATACGTTGGCGAGATGTCAGGATTTGGGTATAGGAAACAGCGCTGGGCGATCAGCGTGGCG
18820 18830 18840 18850 18860 18870 18880 18890 18900

orf206a * M R N V C I A A U C
TGAGTGTGTTACHAGGGTTCTGGCGGAAACGGGGCTTTTATTAAACAGTGTAGGGGTAAACGCTAATGTTGCTATGGCGTTG
18910 18920 18930 18940 18950 18960 18970 18980 18990

orf206a U F P A A L A U V T U T P A R A E G G H G T F T U G Y F O U K P
TGTCCTTGGCGCACTGGGGAGCAGTCAGTCCTGGCGGCTGGCGGAGGCTGAGCATGTTAGCTTGTGGGGCTATTTCAAGTGAACCC
19000 19010 19020 19030 19040 19050 19060 19070 19080

orf206a G T L P S L 5 6 D T G V U S H L K G I N U K Y R Y E L T D S
GGGTCAATTGGCGCTGTTGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
19090 19100 19110 19120 19130 19140 19150 19160 19170

orf206a U G V U M A S L G G F A H S K K S E T U M T G E D T F H Y E S L
TGTTGGGGTGTGCTTGGCTGGCGCGTGGAGGAGAGCAGCAGTGTGACGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
19180 19190 19200 19210 19220 19230 19240 19250 19260

orf206a R G R Y U S U M A G P V L O I S K Q U S A Y A M A G G V A H S
GGCTGGGAGCTATGAGCTGAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
19270 19280 19290 19300 19310 19320 19330 19340 19350

orf206a R H S G S T M D Y R K T E I T P G Y M K E T T I A R P D E S A
TCGGGTGGCCGGCACTAACATGGTATCCCTAACGCGGAAACTACTCCCGGGTATATGAAAGAGGAGGAGCCACTGCCAGGGGAGCAGG
19360 19370 19380 19390 19400 19410 19420 19430 19440

orf206a M R H T S U A H S A G I O I N P A B S U U D I A Y E G S G
AATGGCGACTCTCTAGTGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
19450 19460 19470 19480 19490 19500 19510 19520 19530

orf206a S G D H R T D G F I V G U G U G Y K F
CAGTGGCGACTGGCGTACTGGCGGATTACATGCTGTTGGGGTGGGGTTAAATCTGATTAAGGCGGAGTACACAGCTGTTATGACAG
19540 19550 19560 19570 19580 19590 19600 19610 19620

orf401 M A U K I S G V L K D G T G K P V D N C T
GAACCGGGGGGGCTTTGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
19630 19640 19650 19660 19670 19680 19690 19700 19710

orf401 I Q L K A R R R N S T T U U U N T U T G S E N P D E A G R Y S M
CCATTCTAGCTGAAAGCCAGAGCTACAGCAGCAGCAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
19720 19730 19740 19750 19760 19770 19780 19790 19800

orf401 D U V E Y G O Y S U I L O U P D G F P P S H A G E T I T U Y E D S
TGCGTGTGGGGTGTAGCTGAGCTGAGCTGAGCTGAGCTGAGCTGAGCTGAGCTGAGCTGAGCTGAGCTGAGCTGAGCTGAGCTGAG
19810 19820 19830 19840 19850 19860 19870 19880 19890

orf401 Q P G T L N D F L C H M T E D A R P E P U L R P L E L M Y G E
CACACCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
19900 19910 19920 19930 19940 19950 19960 19970 19980

orf401 E U A R R N A S U V P A Q S T A D A K K S A G D A S S A Q V
AGAGGGTGGCGCGTAAACCCGCTCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
20000 20010 20020 20030 20040 20050 20060 20070 20080

orf401 S A A L V T D A T D S A R A A S T S A G Q A A S S A Q F A S S
TCGGCGGGCTTGTGAGCTGAGCTGAGCTGAGCTGAGCTGAGCTGAGCTGAGCTGAGCTGAGCTGAGCTGAGCTGAGCTGAGCTGAG
20090 20100 20110 20120 20130 20140 20150 20160 20170

orf401 G A E A A S A K A T E A F K S A A A A E S S K H A A A T S A
CCGGCGCGAGAGCGGCATCACAGAGGGCACTGGAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
20180 20190 20200 20210 20220 20230 20240 20250 20260

orf401 G A A K T S E T N A A A S O O S A A T S A S T A P T K A S E
CCGGTGGCGCGAGAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
20270 20280 20290 20300 20310 20320 20330 20340 20350

orf401 A A T S A R D A V A S K E A A K S S E T N A S S S A G R A A
AGGGCGCCACTTACGAGAGTGGGGCTCAAAAGAGGGAGGAGCAAAATCATGAGAAAGGAGCAGCAGCAGTCAAGTGGCGGCTGAG
20360 20370 20380 20390 20400 20410 20420 20430 20440

orf401 S S M T A H E N S A R A A K T I S E T H A R S S E T A A E
 CTTCTCGGAAACGGGCGGAAAGGAAATTCTGCCGGGCGCAAAHACGTCGCGAGAAGATGCGCAGGTCTGAACAGCAGCGGAACCGA
 20448 20450 20452 20470 20480 20490 20500 20510 20520
 orf401 M S A A N D M H K T H A A G G S A S T A S T K A T E A G G S T A U
 GCGCCTTGCGCCGGGAGCGCAGAACGAAACGGCGCGGGGGAGTGCCTGAGCGCTACAGGAGGCTGCGGAGAGTCGGCG
 20538 20540 20550 20560 20570 20580 20590 20600 20610
 orf401 S A S O S K S A A E A A A I P K C N S S A K P A E D I A S A U
 TATCAGCATCAGGGACAGAGTCGGCGGCAGACGGCGCTAGTCGAAAGAAATTGCGGAAACACGTGCGAGAAGATACTCTGACTG
 20620 20630 20640 20650 20660 20670 20680 20690 20700
 orf401 A L E D A D T T R K G I U Q L S S A T N S T S E T L T A A T P
 TCGCGCTTGGGGATSCGGACACAGGGAGGGAGTACTGAGCTCAGCGCTGCAACACGACGCTGAAACCGCTTGCGCACCGC
 20710 20720 20730 20740 20750 20760 20770 20780 20790
 orf401 K A A V K U U V M D E T N R K A H H T U R H *
 CAAAGCGGTTAAAGGGTAATGGGATGAGTCAGAGACGAAAGGGCAACGGCTTGTAGTCGCGACTGCGACCGAACAGCACCC
 20800 20810 20820 20830 20840 20850 20860 20870 20880
 GCTCGAGGGAAACAACTHCCCAUATTGGGAGACCGCTTTGTAATCGCGCGAGTTGAGATGTTATCGACCGCTGACCTGACCG
 20880 20890 20900 20910 20920 20930 20940 20950 20960
 orf314 M T N A L A G K Q P K
 GAATACGCTGAATGAATGGCCGCAGCGCTGGGGATGATCCAGATTTGCTACCCWCCATGACTACCGCGCTTGGGTTAAACACCGAA
 20980 20990 21000 21010 21020 21030 21040 21050 21060
 orf314 N A T L T A L A G L S T A K N K L P Y F A E N D A R S L T E
 GAATCGGACACTGAGCGGCTGGCAGGGCTTCCAGCGGAAATAATAACCTGATTTGGGAAATATGTCGCGCACCGCTGACTG
 21070 21080 21090 21100 21110 21120 21130 21140 21150
 orf314 L T Q U G G R B D I L A L K N S U A D U L E Y L G A G E N S A F P
 ACTGAGCTAGGGTGGCAGGGATHTCTGGCAAAHAAATTGCGCTTGTGAGATGTTGAATACCTTGGGGCGGGTGGGAACTTCC
 21160 21170 21180 21190 21200 21210 21220 21230 21240
 orf314 A G A P I P W P S D I U P S G Y U L M Q G O A F D K S A Y P
 6GCAAGGTCGGCGCATCGCTGCGATGATCTGGCTCTGGCTACGCTCTGATGCGAGGGGGAGGGGGTITGACAAATCAGCTTCCC
 21250 21260 21270 21280 21290 21300 21310 21320 21330
 orf314 K L A U A Y P S G V L P D M R G W H T I K G K P A S G R A U L
 AAACACTGCTGCGCTACCTCGGGGTGCTCTGATGCGAGGCTGAGCTACAGGGGAAACCGGCACGGCGCTGCTGCTGCTGCTG
 21340 21350 21360 21370 21380 21390 21400 21410 21420
 orf314 S Q F E Q D G I K S H T H S A S A S G T D L G T K T I T S S F D
 GTCTCGGACAGGGATGGAATTAAAGTCCGACACCCAGCTGGAGCTGCTGGTAGGGTTGGGGAGGAAACACATCTGCTTIGA
 21430 21440 21450 21460 21470 21480 21490 21500 21510
 orf314 Y G T K T T G S F D B Y G T K S T N H T G A H A H S L S G S T
 TTACGGGGCAAAACACAGCGAGTGTGGATTACCGCACCAATCGAGCTGGCTAGGGTAGGGTTGGGGAGGAAACACATCTGCTTIGA
 21520 21530 21540 21550 21560 21570 21580 21590 21600
 orf314 G A A G A H A H A T T S G L R M N S S G H S Q Y G T A T I T G S
 AGGGGCCGGGTGCTCATGGCCACACAGGGTTAAGGATGAGACAGTCTGGCTGAGTTGGAGACAGCAACCCATTACAGGAAG
 21610 21620 21630 21640 21650 21660 21670 21680 21690
 orf314 L S T U K G T S T Q G I A A Y L S K T D S Q G S H S H S L S G
 TTATCAGCTGTTAAAGGGACAGCACAGGGTATGCTTATTTGCGGAAACGGCATCGAGCGACAGCTATTGTCGCTCATGGCC
 21780 21790 21800 21810 21820 21830 21840 21850 21860
 orf314 T A U S A G A H A H A T U G I G A H O H P U U V I G A H A H S F
 TACAGCGCTGAGTCGGCAGTCATCGCCATACATGGCTGAGCTGGCCACCGCATCCGGTGTGTTATCGGTCGCTCATGCCCA
 21798 21808 21818 21828 21838 21848 21858 21868 21878
 orf314 S I G S H G H T I T U V N A A G N A E N T U K N I A F H N Y I U
 CGATATTGGGTACACCGGACACCATCGCTGATGCGGAAACACCGCATCCGGTAACTGCGGAAACACATGGCTTAACTGAT
 21880 21890 21900 21910 21920 21930 21940 21950 21960
 orf194 M A F R H S E Q P R T I K I Y H N L L A G T H E F I G
 orf314 R L A * M A F R H S E Q P R T I K I Y H N L L A G T H E F I G
 GAGGCTGCTCATATGGCATTCAGAGTCAGTCAGCTGGCCACCGGACCATAAAATTTATGATCTGCTGGCCGAGGAACTAATG
 21980 21990 22000 22010 22020 22030 22040 22050
 orf194 E G D A Y I P F H T G L P A N S T D I A P P D I P T A G G V A
 GAGGTGACGACATATACTGGCTCTGAGCTACCGGATATTCAGCTGGCTGAGCTGGCTGAGCTGGCTTGGGCT
 22060 22070 22080 22090 22100 22110 22120 22130 22140
 orf194 U F N S D E A S H L U E D H R G K T U Y D U A S G D A F
 GTTTACAGCTGATGGGGCATCTGTGGCAGCTCGTGGAGGACCATGCGGGTAAACCGCTCTAGCTGGCTTCCGGCGACGGT
 22150 22160 22170 22180 22190 22200 22210 22220 22230
 orf194 I S E L G D P L P A E N F T H L S P G G E Y Q K H N G D T A H W K
 ATTTCTGAGCTCGCTCGGTTACCGGAAATACTGGGTTAGTCGGGGAGGGAAATATCAGAGGTGGAGACGCGCACAGCTGGGAG
 22240 22250 22260 22270 22280 22290 22300 22310 22320
 orf194 D T E A E K L F R I R E A E E T K K S L M Q U A S E H I A P
 GATCGGGAGGAGCAAAACTTCCGGAGGGAGGAGAACAAAACAAAGCTGGTGGAGGAGCAGCTGGTGGAGGAGCAGCTGGCC
 22330 22340 22350 22360 22370 22380 22390 22400 22410
 orf194 L D D A A P L E T I A T K E E T S L L E A W K K Y R U L L N R
 CCTCAGGATCTGAGCTGGAGATCTGGAAAGGAGAACCTGGTGGAGGCTGGAGGAGAAGTGGTGGAGTGTGGTGGAGCTGACCG
 22420 22430 22440 22450 22460 22470 22480 22490 22500
 orf194 U D T S T A P D I E W P A U P V U M E *
 GTTGATCATCTAGCTGAGCTGGATATGGTGGCTGCTCTCTGTTAGGGATGATCTGTTTTGCGATTATTCCTGAGGAGATA
 22510 22520 22530 22540 22550 22560 22570 22580 22590
 AAAATACGTTCTGGGGTTAGTTAGTATATGTAAGGCTGGTGTATTGGTTTATTTGGCGATTATTCCTGAGGAGATA
 22600 22610 22620 22630 22640 22650 22660 22670 22680
 ATGACTCATTTGTCATAGTTGTTACATCGGCGCATCTGCTTTAAGGATGAGCGCATGARATAATGGTTTTCTGATG
 22690 22700 22710 22720 22730 22740 22750 22760 22770
 GCTGTGTTGATTTCTAAGGCTGGTTTTCTCTGTTCTAATCTTCTGTTGATGTTATGGGGAGTGGAGTCTGCTCTGAT
 22780 22790 22800 22810 22820 22830 22840 22850 22860
 CCAATTACCTGAGCTGGCTTCTATATAATGGCATTTGAGCTGGTGGAGGAGAAGTGGAGTGGAGTCTGCTGAGCCATAG
 22870 22880 22890 22900 22910 22920 22930 22940 22950
 ATCCAAATGAGGCTCATGGGCTTGGTGGAGCTGGTGGAGCTGGCAT
 22960 22970 22980 22990 23000 23010 23020 23030 23040

(b)

CII	<u>H E D P I T L K D Y A M R F G Q I</u> CTCTGGCGGTGHT4HHTGTTTGTCAAGGGAGGTTGTTGTTAAGGATTATGCAGCTGGCTGGCAAAAC 38018 38020 38020 38040 38050 38060 38070 38080 38090
CII	<u>K T A Y D L G U S H A I N K H I H A G R K I F L T I N A D</u> CAGAGCAGCTTAAAGHTCTCGGGGTATTCAGGGCGGTTAACAGGAGCTTCTGGAGGCGAAAGATTTTTAACTATAAACGCTGA 38100 38110 38120 38130 38138 38140 38150 38160 38170 38180
CII	<u>G S V Y H E D P F P S N K K T T A *</u> TOGAGCTGTTATGGGGHGHG01HAAGCTTCCGGGTTAACAGGAAACAAACGAA1HA1HACCCCGCTTTAACACATTCAGGCC 38190 38200 38210 38220 38230 38240 38250 38260 38270 S _c
CII	<u>GAAAGGGGCACTAAATTAAACCCACRCCATTGGTATGCTATTGCTACATTCACATCAATGTTATCTAGGAAATACTACATA</u> 38280 38290 38300 38310 38320 38330 38340 38350 38360
CII	<u>U R A A N K R P A L P I E S A S A L N K I I A M L G T E K A T A E</u> TGGTCGTCGACAAACAGCGACRCGAGGCTCTAGGAGACTGCGTTGCTAACAAARTGCACATGCTTGGAAGAACAGACCGG 38370 38380 38390 38400 38410 38420 38430 38440 38450
CII	<u>M U G H P D K S O I S R H K R D H I P K F S M L L A U L E W G</u> AACGCTGGCGCTGTTAAGTTCGAGATCACAGGGTGAAGGGACTGATTCCAAGGTTCTAACATGCTTGCTGTTGCTGTAATGGG 38460 38470 38480 38490 38500 38510 38520 38530 38540
CII	<u>U U D D D M A R L A R Q V U A A I L T N K K R P A A P T R S E</u> GGGTCGTCGACGAGCAGCTGGCTGCA1TGCGKAGAACAGTGTCTGGATCTTACCAATAAAAACGCCGGCGAACCGAGCGTCTG 38550 38560 38570 38580 38590 38600 38610 38620 38630
<u>- - - d_{ap} RNA - - -</u>	
CII	<u>S₀ M T N T A K I L L N F G R</u> AACAAATCCAGTGGAGTTCTGGAGCTTACTGGATCTAACAGGGAGCTTATGACAAATACTACAGGAAATACTACATTCGGCG 38650 38660 38670 38680 38690 38700 38710 38720
CII	<u>B N H G O E R N U A D L D D G Y A R L S N M L L E A Y S G</u> AGGTAACITTCGCGGGAGCGTWAATGCGAATTCGATGATGGTCTACGCCAGACIACTAACATATGCTGGGTTAGGGCTTATCGGG 38730 38740 38750 38760 38770 38780 38790 38800 38810
CII	<u>A D L T K R O F L L A I L R K T Y G H N K P M D R I T D</u> CGAGAGTCTGACCAAAGCGCAGTTTAAAGTCTGCTTGGCATTCGGTAAACCTATGGTTGGAATAACCAATGGACAGAACTACCGA 38820 38830 38840 38850 38860 38870 38880 38890 38900
CII	<u>S G L S E I T K L P V K R C N E A K L E L U V R M N I I K O O</u> TTCTCAACTTACGCGGATTAAACGTTACCTGTCAAGCGTGTCAATGAAAGCTTACGACTGCTGAGAATGATAATTACGAGCAGA 38920 38930 38940 38950 38960 38970 38980 38990
CII	<u>G G M F G P P N K N I S E H C I P Q N E G K S P K T P D K T S</u> AGGCCGATGTTGGACCAATAAAACATCTCAAGHATGTCATGCTTAAACAGGGAAAAATCCCTAAACGAGGGATAAAACGAGGGATAAAACATC 39000 39010 39020 39030 39040 39050 39060 39070 39080
<u>- - - d_{ri} mutations - - -</u>	
CII	<u>L K L G D C Y P S K O G D T K P D T K E I T K E K R K D Y S S E N</u> CCTCHATTGGGGATTGCTATCCCTCAAAACAGGGGACAAACAGACTTACAAAGGAAAGGAAAGATTATTCGAGAGAA 39100 39110 39120 39130 39140 39150 39160 39170
CII	<u>S G E S S D Q P E N N D L S U U K K P D A A I O S G S 5 K H G T A</u> TTCTGGCGAACTCTCTGACCACGACCTTGTGGTAAACCGGGATGCTGCAATTCAGAGCGCAGCGAGTGGGGAGACG 39180 39190 39200 39210 39220 39230 39240 39250 39260
CII	<u>E D L T A A E H N F D M U K T I P A R K P H F A G W N A N</u> AGGAGCCTGACCGCCGAGAGTGGATGTCAGTGGAGACATACGCAACATGGCCAGAAACCGAATTTGGTTGGGCTAA 39270 39280 39290 39300 39310 39320 39330 39340 39350
CII	<u>D I R L N R E P D G R N H R D M C U L F R H A C Q D D N F H S</u> CGATATCCGCCCTGATGGCTGACGACGCTAACCCGGCAGATGTTGGCTGCTGGCATGGCGCATGGCGACANTTCTGGTC 39360 39370 39380 39390 39400 39410 39420 39430 39440
CII	<u>G N U L S P A K L R D K U T Q L E I I N R N K O O A G U T A S</u> CGGTAACTGCTGGGCCGCGCAACACTCCGGCATAGTGGACGGCCACATGGCAACCCGAGAACAGGAGGGTGACAGCG 39450 39460 39470 39480 39490 39500 39510 39520 39530
CII	<u>P K P K L D D L T N T D H I Y G U D L * M K N I A A Q M V N F D R</u> CAACCAAAACTCGACCTGCAACACAGAGCTGGATTAACAGGGGTGGCTGATGTTAAACATCGCCGACAGAGATGGTTAACTTGGCGT 39540 39550 39560 39570 39580 39590 39600 39610 39620
CII	<u>E Q M R R I A N N N P E D Y D E K P Q U O D U A Q I I N G U</u> GAGCAGATGGCTGGAGTCACGACACAGACATGGCGAGAGGCGCAGGGTAGCAGCAGAGTGGCGAGATCATACGGTGTG 39630 39640 39650 39660 39670 39680 39690 39700 39710
CII	<u>F S Q L L A T F P A S L A N D N G E U N E I R R Q H U L A</u> TTCAGCGAGTTAGGGCAACTTCCCGCGAGGGCTGGCAACAGAGCTGGAGCTGGCGAGGCTGGGTGATGTTACGAGGT 39720 39730 39740 39750 39760 39770 39780 39790 39800
CII	<u>F R E N G I T T H E Q U N A G M R U A R R Q N R P F L P S P</u> TTTCGGGAAACCGGGATCACACAGATGGCAACGGGAGCTGGCGTAGGCCGTCGGCAGAACATGGCGACTTCTGGCTGTTACGCCAC 39810 39820 39830 39840 39850 39860 39870 39880
CII	<u>P G Q F U A H C R E E A S U T A G L P N U S E L U D M U Y E Y</u> GGGCAGTGGCTGGAGCATGGCGGGAGGCGATCCGGCTTACCGCCGACTGCGAACAGGCTGGGTGATGTTACGAGGT 39900 39910 39920 39930 39940 39950 39960 39970 39980
CII	<u>C R K R G L Y P D A E S Y P H K S N A H Y H L U T N L Y Q N</u> TGCCGGAKGGCGAGCGCTGTTGGCGAGTCTGGGAAATCAACAGGCGACTTCTGGCTGGTTACCAACCTGTATCAGAC 39990 40000 40010 40020 40030 40040 40050 40060 40070
CII	<u>M R A N A L T D A E E L R R K A A D E L U H M T A R I H R G E</u> ATGGGGGGCAATGGCGCTTACTGGTGGCAATTACGGCGTAAAGGGCGAGGATGGCTTGGCTGAGAATTACCGGGTGGTGG 40080 40090 40100 40110 40120 40130 40140 40150 40160
CII	<u>A I P E P U K G L P U N G G R P L N R A Q A L A K I A E I K</u> GCGATCCCTGACCAAGTAAACAACTCCGGTCAATGGCGGTAGACCGCTAACATGGCACAGGGCTGGCGAGAGTCGAGAACTC 40170 40180 40190 40200 40210 40220 40230 40240 40250
ren	<u>A K F G L K G A S U *</u> GCTAGTTCGGACTGAGGGAGCHGHTGATGAGGGCAAGAGGGCHATTATCTACCTGGGGAGGCTAACTGCTCTGTGCGCCCG 40260 40270 40280 40290 40300 40310 40320 40330 40340
ren	<u>V A A L T G A H T U T S I N Q A A K M A R A G L L U I E G K</u> ACGTTGGCGCGCTAACGGCGCAAGCTAACACGGCTAACAGGCGCTAACATGGGCCGCGCTAACATGGCACGGGAGGTTCTGGTTATCGAGGGTA 40350 40360 40370 40380 40390 40400 40410 40420 40430

FIG. 2 (b).

ren V W R T U Y Y R F A T R E E P E G K * S T N L V F K E C R Q
 AGGTCTGACCGTGTATACCGCGTTACCGGGAAAGGAACGGCAAGATGACCTGGTTTAAGGAGTGTCGCQ
 40448 40450 40452 40454 40456 40458 40460 40462 40464 40466 40468 40470 40472 40474 40476 40478 40480 40482 40484 40486 40488 40490 40492 40494 40496 40498 40500 40502 40504 40506 40508 40510 40512 40514 40516
 ren S A M K R U L Y G U K R * I R S P A H Q Q N A I H
 AGAGTCCCGCATGAAACGGGTATTGGCGGTATATGGGTTAACAGGAGCCTACATTACTGAGCTAATAACAGGCGTCTGCTGGTAAT
 40530 40540 40550 40560 40570 40580 40590 40600 40610 40620 40630 40640 40650 40660 40670 40680 40690 40700 40710 40720 40730 40740 40750 40760 40770 40780 40790 40792 40794 40796 40798 40799 40800 40802 40804 40806 40808 40810 40812 40814 40816 40818 40820 40822 40824 40826 40828 40830 40832 40834 40836 40838 40840 40842 40844 40846 40848 40850 40852 40854 40856 40858 40860 40862 40864 40866 40868 40870 40872 40874 40876 40878 40880 40882 40884 40886 40888 40890 40892 40894 40896 40898 40900 40902 40904 40906 40908 40910 40912 40914 40916 40918 40920 40922 40924 40926 40928 40930 40932 40934 40936 40938 40940 40942 40944 40946 40948 40950 40952 40954 40956 40958 40960 40962 40964 40966 40968 40970 40972 40974 40976 40978 40980 40982 40984 40986 40988 40990 40992 40994 40996 40998 40999 41000 41002 41004 41006 41008 41010 41012 41014 41016 41018 41020 41022 41024 41026 41028 41030 41032 41034 41036 41038 41040 41042 41044 41046 41048 41050 41052 41054 41056 41058 41060 41062 41064 41066 41068 41070 41072 41074 41076 41078 41080 41082 41084 41086 41088 41090 41092 41094 41096 41098 41100 41102 41104 41106 41108 41110 41112 41114 41116 41118 41120 41122 41124 41126 41128 41130 41132 41134 41136 41138 41140 41142 41144 41146 41148 41150 41152 41154 41156 41158 41160 41162 41164 41166 41168 41170 41172 41174 41176 41178 41180 41182 41184 41186 41188 41190 41192 41194 41196 41198 41200 41202 41204 41206 41208 41210 41212 41214 41216 41218 41220 41222 41224 41226 41228 41230 41232 41234 41236 41238 41240 41242 41244 41246 41248 41250 41252 41254 41256 41258 41260 41262 41264 41266 41268 41270 41272 41274 41276 41278 41280 41282 41284 41286 41288 41290 41292 41294 41296 41298 41300 41302 41304 41306 41308 41310 41312 41314 41316 41318 41320 41322 41324 41326 41328 41330 41332 41334 41336 41338 41340 41342 41344 41346 41348 41350 41352 41354 41356 41358 41360 41362 41364 41366 41368 41370 41372 41374 41376 41378 41380 41382 41384 41386 41388 41390 41392 41394 41396 41398 41400 41402 41404 41406 41408 41410 41412 41414 41416 41418 41420 41422 41424 41426 41428 41430 41432 41434 41436 41438 41440 41442 41444 41446 41448 41450 41452 41454 41456 41458 41460 41462 41464 41466 41468 41470 41472 41474 41476 41478 41480 41482 41484 41486 41488 41490 41492 41494 41496 41498 41500 41502 41504 41506 41508 41510 41512 41514 41516 41518 41520 41522 41524 41526 41528 41530 41532 41534 41536 41538 41540 41542 41544 41546 41548 41550 41552 41554 41556 41558 41560 41562 41564 41566 41568 41570 41572 41574 41576 41578 41580 41582 41584 41586 41588 41590 41592 41594 41596 41598 41600 41602 41604 41606 41608 41610 41612 41614 41616 41618 41620 41622 41624 41626 41628 41630 41632 41634 41636 41638 41640 41642 41644 41646 41648 41650 41652 41654 41656 41658 41660 41662 41664 41666 41668 41670 41672 41674 41676 41678 41680 41682 41684 41686 41688 41690 41692 41694 41696 41698 41700 41702 41704 41706 41708 41710 41712 41714 41716 41718 41720 41722 41724 41726 41728 41730 41732 41734 41736 41738 41740 41742 41744 41746 41748 41750 41752 41754 41756 41758 41760 41762 41764 41766 41768 41770 41772 41774 41776 41778 41780 41782 41784 41786 41788 41790 41792 41794 41796 41798 41800 41802 41804 41806 41808 41810 41812 41814 41816 41818 41820 41822 41824 41826 41828 41830 41832 41834 41836 41838 41840 41842 41844 41846 41848 41850 41852 41854 41856 41858 41860 41862 41864 41866 41868 41870 41872 41874 41876 41878 41880 41882 41884 41886 41888 41890 41892 41894 41896 41898 41900 41902 41904 41906 41908 41910 41912 41914 41916 41918 41920 41922 41924 41926 41928 41930 41932 41934 41936 41938 41940 41942 41944 41946 41948 41950 41952 41954 41956 41958 41960 41962 41964 41966 41968 41970 41972 41974 41976 41978 41980 41982 41984 41986 41988 41990 41992 41994 41996 41998 41999 42000 42002 42004 42006 42008 42010 42012 42014 42016 42018 42020 42022 42024 42026 42028 42030 42032 42034 42036 42038 42040 42042 42044 42046 42048 42050 42052 42054 42056 42058 42060 42062 42064 42066 42068 42070 42072 42074 42076 42078 42080 42082 42084 42086 42088 42090 42092 42094 42096 42098 42100 42102 42104 42106 42108 42110 42112 42114 42116 42118 42120 42122 42124 42126 42128 42130 42132 42134 42136 42138 42140 42142 42144 42146 42148 42150 42152 42154 42156 42158 42160 42162 42164 42166 42168 42170 42172 42174 42176 42178 42180 42182 42184 42186 42188 42190 42192 42194 42196 42198 42200 42202 42204 42206 42208 42210 42212 42214 42216 42218 42220 42222 42224 42226 42228 42230 42232 42234 42236 42238 42240 42242 42244 42246 42248 42250 42252 42254 42256 42258 42260 42262 42264 42266 42268 42270 42272 42274 42276 42278 42280 42282 42284 42286 42288 42290 42292 42294 42296 42298 42300 42302 42304 42306 42308 42310 42312 42314 42316 42318 42320 42322 42324 42326 42328 42330 42332 42334 42336 42338 42340 42342 42344 42346 42348 42350 42352 42354 42356 42358 42360 42362 42364 42366 42368 42370 42372 42374 42376 42378 42380 42382 42384 42386 42388 42390 42392 42394 42396 42398 42400 42402 42404 42406 42408 42410 42412 42414 42416 42418 42420 42422 42424 42426 42428 42430 42432 42434 42436 42438 42440 42442 42444 42446 42448 42450 42452 42454 42456 42458 42460 42462 42464 42466 42468 42470 42472 42474 42476 42478 42480 42482 42484 42486 42488 42490 42492 42494 42496 42498 42500 42502 42504 42506 42508 42510 42512 42514 42516 42518 42520 42522 42524 42526 42528 42530 42532 42534 42536 42538 42540 42542 42544 42546 42548 42550 42552 42554 42556 42558 42560 42562 42564 42566 42568 42570 42572 42574 42576 42578 42580 42582 42584 42586 42588 42590 42592 42594 42596 42598 42600 42602 42604 42606 42608 42610 42612 42614 42616 42618 42620 42622 42624 42626 42628 42630 42632 42634 42636 42638 42640 42642 42644 42646 42648 42650 42652 42654 42656 42658 42660 42662 42664 42666 42668 42670 42672 42674 42676 42678 42680 42682 42684 42686 42688 42690 42692 42694 42696 42698 42700 42702 42704 42706 42708 42710 42712 42714 42716 42718 42720 42722 42724 42726 42728 42730 42732 42734 42736 42738 42740 42742 42744 42746 42748 42750 42752 42754 42756 42758 42760 42762 42764 42766 42768 42770 42772 42774 42776 42778 42780 42782 42784 42786 42788 42790 42792 42794 42796 42798 42800 42802 42804 42806 42808 42810 42812 42814 42816 42818 42820 42822 42824 42826 42828 42830 42832 42834 42836 42838 42840 42842 42844 42846 42848 42850 42852 42854 42856 42858 42860 42862 42864 42866 42868 42870 42872 42874 42876 42878 42880 42882 42884 42886 42888 42890 42892 42894 42896 42898 42900 42902 42904 42906 42908 42910 42912 42914 42916 42918 42920 42922 42924 42926 42928 42930 42932 42934 42936 42938 42940 42942 42944 42946 42948 42950 42952 42954 42956 42958 42960 42962 42964 42966 42968 42970 42972 42974 42976 42978 42980 42982 42984 42986 42988 42990 42992 42994 42996 42998 42999 43000

(c)

S_R

TTGTTCCATACACCTCCATTAGTAAGTGCHACCATATTACCGCCAGGGTAAATTAGTCACACGCACGGTGTGTTAGATAATTATCCCTT 38840 38830 38820 38810 38800 37999 37988 37979 37960	
CGCGTATAGATTAACTGATGGCACACAAAAAGAACCAATTAGACAGACGCGCTTGGAGGAGCAGCTGGCCTTAAAGCAATTATGGA 37950 37940 37930 37920 37910 37900 37899 37880 37870	
K K K N E L G L S Q E S U M D K M G M G Q S G U G A L F N G 37860 37850 37840 37830 37820 37810 37800 37790 37780	
I H A L N A Y N H A L T K I L K S U E F S P S I A R E CATCAATGCTTAACTGCTTAAACGCCATTGCTTACACAAATTCTAACAGTAAAGCTGGGAGGAGCTGGGAGCTGGGCTTAAAGCAATTATGG 37770 37760 37750 37740 37730 37720 37710 37700 37690	
T Y E M Y E A V S M O P S L R S E Y E Y P U F S H V Q A G M AACTCTAGAGATGTAGAAGCGGTAGTAGTGGCAGCTTACAGTGGAGTACCCCTGTTTCATGTTCAAGCAGGAGAT 37680 37670 37660 37650 37640 37630 37620 37610 37600	
F S P K L R T F T K G D A E R H U S T T K K A S D S A F N L GTTCCTACCTAACAGTAGAACCTTTAACAGAGTGGGAGAGATGGGAGACACCAACAAACAGTGTGTTCTGGCT 37590 37580 37570 37560 37550 37540 37530 37520 37510	
E U E G N S M T A P T G S P S F P D G M L I L U D P E Q A TGAGGTGAGGTAATTCTCATGCGACACGGCTCAAGCCAACTTCTGGCAGACGATGTTACTTCATGTTGACCTGGCAGCAGC 37500 37490 37480 37470 37460 37450 37440 37430 37420	
U E P G D F C I A R L G G D E F T F K K L I R D S G Q U F L TGTTGAGCCAGGTGATTCTGATGCCAGACTGGGGTGTGATTACCTTACAGAACGACTGACGGAGATGGCTCAGGGTGT 37410 37400 37390 37380 37370 37360 37350 37340 37330	
A P L H P Q X P M I P C N E S C S V V G K V I A S Q W P E E ACACACACTAACAGCACAGTACCCATGCACTTCCGGTGTGTTGGGGAGAATTACGCTAGTCAAGTGGCTCAGGGAGA 37320 37310 37300 37290 37280 37270 37260 37250 37240	
T F G * GACGTTGGCTGATGCCAAGGTGTTCTGGCGCAGTAGCTGATAACAAATGNGCAAGAACTTCATGCAATTAGGGGAAATTTC 37230 37220 37210 37200 37190 37180 37170 37160 37150	
rexA	M K N G F Y A T T Y R S K N K G K D K R CCCCCTAGACATACAGTAGATAATGGATTGHHATTAGAGATGTTTATGCGACTTACCGCAGCAAAATAAAGGAAAGATAGC 37140 37130 37120 37110 37100 37090 37080 37070 37060
rexA	S I N L S U F L N S L A D N H H L Q U G S N Y L Y I H K I GCTCANAACTCTGTGTTTCTCTTAATTCTGCTGGCTGATACTCATCACCTCAGGGTGGCTCCAATTATTGTATATCATAAAA 37050 37040 37030 37020 37010 37000 36990 36980 36970
rexA	D G K T F L F T K A T D K S L U Q K I N R S K A S U E D I K TCGATGGAAAACCTCTTCTTACCAAACAAACAGACAGACTGGTTCTGGCTCAGAGATAATCTGCTTAAAGCTTCAAGTGGAGAT 36960 36950 36940 36930 36920 36910 36900 36890 36880
rexA	N S L H D D E S L L G F P S F L F U E G D T I G F F A R T U F G AGARCAAGCTCGCAGATGACGAACTTGGGGATTCCCCATTTTGTGTTGGAGGGCACACCATGGGTTTCCAGAACACTGTT 36870 36860 36850 36840 36830 36820 36810 36800 36790
rexA	P T T S D P L T D F L I G K G M S L S S G E R U Q I E P L M R GGCCGACCACTCGCTGACAGATTTTAACTGGGGAGGAATGATTACAGCTGGAGAGGCGCGCTCAGATAAGGCCACTGATGA 36780 36770 36760 36750 36740 36730 36720 36710 36700
rexA	G T T K D D U M M H M H F I G R T T Y V K U E A K L P V F G D I GGGGAAACCAACAAAGCAGTGTATGACATATGCTTCTGCGCCGAAACACGGTGAGGGAGGCAAGCTACCTGTTGGCGATA 36690 36680 36670 36660 36650 36640 36630 36620 36610
rexA	L K V U L G A T I E G E L F D S L D I U I K P K F K R D I K TATTAAGGCTTAACTGGGGCACAGATAATGAGGGGAGGTTTGTCTGATGTTAGCTTAAAGCCAAATTAAAGGGGATA 36600 36590 36580 36570 36560 36550 36540 36530 36520
rexA	K U A K D I I F N P S P Q F S D I S L R K A D E A G D I L T AAAAGGTGCAAGGTTATTTTAACTGGGGCACAGATAATGAGGGGAGGTTTGTCTGATGTTAGCTTAAAGGGGATA 36510 36500 36490 36480 36470 36460 36450 36440 36430
rexA	E H Y L S E K G H L S A P L N K U T N A E I A E E M A Y C Y CAGACACATTCTATGAGAAAGGCACTCTGGCCCTTGACCAAGGTGACAGATAAGCTGAGATAAGCTGAGAGTGGCAT 36420 36410 36400 36390 36380 36370 36360 36350 36340
rexA	A R M K S D I L E C F K R Q U G K V K D * M R N R ACGCAAGATGAAAGTGTATACTGGATGTTTAAAGGCGAGGTTGGCCAAAGGTTAAAGGAGTTAATTACAGGAGTAATTATCGCGACA 36330 36320 36310 36300 36290 36280 36270 36260 36250
rexB	I M P G U Y I U I P Y U I V S I C Y L F R H Y I P G U S GAATCATGCTGGTTTACATAGATAATCTCTTACGCTTATGCTGAGCATTTCTGCTATCTCTCCGGCACTACATCTGGTGT 36240 36230 36220 36210 36200 36190 36180 36170 36160
rexB	F S A H R D G L G A T T L S S Y A G T M I A I L I A A L T F L L CTTCTGCTGAGATGGCTGGGGGACATGGCTGCTGAGGAGCATGTTGCAATCTGATGCTGGCTTGGAGCT 36150 36140 36130 36120 36110 36100 36090 36080 36070
rexB	I G S R T P R R L A K I P E Y G Y M T S U U I U Y A L S F U E TAATCGGGAAACAGAACCGCCGHCCTGGCAAAAGATGAGGGTATGAGCTATGACATCTGGAGTATTGCTTATGCT 36060 36050 36040 36030 36020 36010 36000 35990 35980
rexB	L G H L F F C G L L L L S S I S G Y M I P T I A I G I A S A PGCTTGGGGCTTCTGGGGCTTCTGGGGCATACAGATAACGGCTACATGATAACCCACTATGCCATCGGCCATCGCT 35970 35960 35950 35940 35930 35920 35910 35900 35890
rexB	S F I H I C L U F O P L Y N L T R E Q * CATGGTTCATCTATATGCTCCTTGTCTTCAACATATAATTGAGAGAACAGATAACGGCTTACCGCCTCAGCGCGGGTTTCTT 35880 35870 35860 35850 35840 35830 35820 35810 35800
	TGCCCTCACGATCGCCCCCAAAACACATAACCAATTGATTTGAAAGATAATAGATAACACTACATAACATAGCAATTCTGAGCT 35790 35780 35770 35760 35750 35740 35730 35720 35710
	TCACCTCACCAACATGCCCCCTGCAAAAGATAATCATATAAAACACATAACGATAACCATCTGGGGTGTAAATTATCTGGGG 35780 35690 35680 35670 35660 35650 35640 35630 35620
	TGTTGACATATAACCACTGGGGTGTACTGAGCAGCATCAGCAGGACGCACTGGACACCATGAGAGGTGACGCTTAAAGGG 35610 35600 35590 35580 35570 35560 35550 35540 35530
	TGAAGAAGGGCGACATTCAAGCAGAGGGCTTGGGGTGTGATGAGCAACGAGATGGCCGTAAGTGCAGATTCCG6GATTAGCT 35520 35510 35500 35490 35480 35470 35460 35450 35440
N	M C Q S R G U F U O D Y N C H T P P K L T D R R I Q M D A Q AATGTCGCAATCGGGGGTTTCGCTGGAGCTACACTGACACACACACAAAGCTAATGACAGGAGATCAGATGGATGCGACA 35430 35420 35410 35400 35390 35380 35370 35360 35350

FIG. 2 (c).

N T R R R R E R R R H E K D H O H K A A A N P L L U G V S A K P U N
HACGGCGCCGCGAACGTCGCOCAGAGAACGGCTCATGGAAAGCAGCAATCCCTGTTGGTGGGTAAGGCCAAACCGTAA
35340 35330 35320 35310 35300 35290 35280 35270 35260
N R P I E S L M P K P K S H L N P I D L T U L E F Y H
CCGCCCTATTCCTCGCTGAATCGAACCGAATCACAGTAGAACCGCACTTAAATCCGATAGACCTTACAGTGCTGGTGAATCCA
35250 35240 35230 35220 35210 35200 35190 35180 35170
N K Q I E S N L O R I E R K K N O R T W Y S K P G E R G J I T C S
CAACAGATGAAAGCAACCTGCACGCTTACGGCAGAACATCGCHACATGGTACAGCAAGCCTGGCAAGCGCCATACATCGAG
35160 35150 35140 35130 35120 35110 35100 35090 35080
H G R O X K K G K S I P L I
TGGGCAGGAAATTAAGGGAAATTAAGGGCTTACGTTCTTATCTAGTCTAGATTGGCTTGGCTTATCTCAATATTATATGGATCAT
35070 35060 35050 35040 35030 35020 35010 35000 34990
AGCTGCAACTAACTCAGTCAGTCAATAATCTCTCTAGGAAATAATATGCTTCCATCCATGGGAAAGGTTTGTTCACACAC
34980 34970 34960 34950 34940 34930 34920 34910 34900
CAAGCTCAATCAACTCAACTATGGATTGTTGATGTAACACATCTTCTGCTTCAATTAGGGCTGCGCACARACCATAG
34890 34880 34870 34860 34850 34840 34830 34820 34810
ATTGCTCTTGTAAAGGTTTGTAAAGTACTGATCGACTTTCATGCTTACATTTAGCTTAAATCGCTTATATCTGG
34800 34790 34780 34770 34760 34750 34740 34730 34720
CGCTGGCAATAGCTGATAATCGATGCAATTAATTCCTAGCGAAAHATGCAAGAGAGAACATGCCACACATGAGGAATACCGA
34710 34700 34690 34680 34670 34660 34650 34640 34630
TTCTCTCATTAACATATCAGGGCATTCTGGGCTTAAGCAGAGTCCAAACGATAACGATCATATAACATGGTTCTCCAGAG
34620 34610 34600 34590 34580 34570 34560 34550 34540
GTTCAATTACTGAACTCGTCCAGGATAACGAGTGGATGCTTACTCATCAACTGTAAGGGTTGTAATAGTTATCCGATTC
34530 34520 34510 34500 34490 34480 34470 34460 34450
TCGCTGTAGGGGTACACGAGAACACCAGGCGCTGTGTTAAAGAGACAGGCAACATCTTACTACCCCACTTAAAGGTGAT
34440 34430 34420 34410 34400 34390 34380 34370 34360

ral M E F F E E F E E E H P O D M E Q Y Q D Y P Y D Y *
orf28 ATATGGAGAGRAGATTGGAGAGTCGAAGCATCCTAGGATGTGATGGACAAATACAGGACTATCCGTATGACTACGACTATGAT
34350 34340 34330 34320 34310 34300 34290 34280 34270
ral K H Q W C G O F K R C N G C K L Q S E C M U K P E E M F P U
AAAATCATGTTGAGACAAATTCAAGCCATGCAATGGATGCAAGCTGCAATGGTGTGAAAGGAAATGTTCCCTGTA
34260 34250 34240 34230 34220 34210 34200 34190 34180
ral M E D G K Y U D K H W A I R T T A M I A R E L G K Q N N K A A
ATGGAAGATGGAAATATGCTGATAATGGCAATACGAGCAGCAGCAGTGGTAAACAGAGACTGGTAAACAGAACACAGACTGCC
34170 34160 34150 34140 34130 34120 34110 34100 34090
TGATAGTGGCCCTTATTTGGCATTAATAACAGAAATAACACTGCACTGTGTTATCCATTCACAGGTGAAATACAGGAGCAATGTCG
34080 34070 34060 34050 34040 34030 34020 34010 34000
TCGTAACTAACAGGAGCCGACTTGTCTGATTATTGGATCTTGCCTTCCAGTGTGAGGGCGATTTTTATCTGTGAGGGATATG
33990 33980 33970 33960 33950 33940 33930 33920 33910
ssb M S N I K K Y I I D Y D H K A S I E I E I D H D U M T E F
AACAGATGTCACATACATGATGTTACGACTGAGAACATGAGATGAAATGACCATGACGTAATGAGAG
33980 33980 33980 33980 33980 33980 33980 33980 33980
ssb K L H Q I N Q N F W H S D S E Y R L N K H G S U L N A U L I M L
AAAATCTTACCCAGATAATAATTTCTGGTCAGACTCTGATACCGACTCAATAACAGCCTGCTGATTAATGCTGTTAAATCATG
33910 33900 33890 33880 33870 33860 33850 33840 33830
ssb P O H A L L I A I S S D L N A Y G U V U C E F D W H M D G H G Q
TGGCGCACTGCTGCTTACGACTGCAATGGATGTTGAGGAAATAAGGAAATACGCACTGCAATGGTGTGAGGTTGACTGGAATGGTC
33720 33710 33700 33690 33680 33670 33660 33650 33640
ssd E G N P P M D G S E G I R I T D I D T S G I F D S D D M T I
AGGRAGGAGCTGCCCTCAATGGATGTTGAGGAAATAAGGAAATACGCACTGCAATGGATATGTTGAGGAAATGGTC
33630 33620 33610 33600 33590 33580 33570 33560 33550
ssd K A R *
TCAGGGCCGCTGAGTGGCTGTTTACCGCATACCAATHACGCTTACCTGAGGGCTTTTCGTTATGTTATAATAGGAGCACACATGC
33540 33530 33520 33510 33500 33490 33480 33470 33460
cIII Y A I A G H P U A G C P S E S L L E R I T R K L R D G W K R
AATATGCCATTGCGGGCTGCTGTTGCTGCTGCCCTCGCATTTACTGAAAGAAATCCCGTAAATTACGTGAGGGATGGAAAC
33450 33440 33430 33420 33410 33400 33390 33380 33370
kii L I D I L N Q P G U F P K N G Y P D *
GCCCTATGCACTACATTAATCAGGCCAGGAGCTCCAAAGGATGAACTTACGACTTACGACTTACGACTTACGACTTACGACTT
33360 33350 33340 33330 33320 33310 33300 33290 33280
kii I G P E K M F R E H U D A Y K K W I L I L I L K L R S S K S I H
ATTGGCGATGAAAGATGTTCTGAGGCGCTCGAACGCTTATAAAATGATTAATGAAACTGAGATCAAGCAAAAGCATTCA
33270 33260 33250 33240 33230 33220 33210 33200 33190
Y * M N A Y Y
TAACCCCCCTTCTGTCTTCTAATCAGGCCGGCATTTCCGGGGCATTTTCAGGCTTACGAGGTTCAAGCCATGACGCTTATI
33180 33170 33160 33150 33140 33130 33120 33110 33100
Y I O D R L F A P S H A R H Y Q Q L A R E E K E A E L A D D M
ACATTCAAGGATCGTCTGAGGGCTCAGAGACTGGGGCGCTCACTACCAAGACTCTGCCCTGAGAGAGAGGGAGACTGGCAGACGACA
33090 33080 33070 33060 33050 33040 33030 33020 33010
Y E K G L P Q H L F E S L C I D H L Q R H G A S S K K S I T R A
TGGAAAGGAGCTGCCAGGACCTGTTGAGTGCATGCTGATGCTGATTCAGGCTTACGGTCAAGGCCACGGGCCAACAAATCCATTACCGTG
33080 32990 32980 32970 32960 32950 32940 32930 32920
Y F D D D U E F D E R M A E H I R Y M U E T I A H H O U D I
CGTTTGTGAGCATGTTGAGGAGCCATGCAAGAACATCGGTACATGGTGAACACCATGCTTACGGTCAACGGTGTGACCGACAGGTTGATATG
32910 32900 32890 32880 32870 32860 32850 32840 32830
β M S T A L A T L A G K L A E R V G M D S U D P Q
γ S E V *
ATTAGAGGATATAAACGAAATGAGTACTGCACTCGCAACGCTGGCTGGGAAGCTGGCTGAGCTGTCGGCATGGATTCTGTCGACCCACA
32820 32810 32800 32790 32780 32770 32760 32750 32740

NUCLEOTIDE SEQUENCE OF BACTERIOPHAGE λ DNA

751

β E L I T R O T A F K G P A S D H Q F I A L L I V A N Q Y
 G G A C T G A T C A C C A T C T T G C C A G H C G G A T T H A G G T 3 A T G C C H C G A T G C A T T A C T G A T C G T G C C A C C A G T A
 32730 32720 32710 32700 32690 32680 32670 32660 32650
 β G L N P W T K E I Y H F P D K Q N G I U P U U G G U D G H S R
 C O G C C T T A T C C G T G G A C G A H G H A A T T A C G C C T T C C T G A T H A G C A G A T G G C A T G C T I C C G G T G G G C G T I G A T G G C G T G G C C G
 32640 32630 32620 32610 32600 32590 32580 32570 32560
 β I N E N Q Q F D G M D F E Q D N E S C T C R I Y R K D
 C U T C A T C A T G A H A R C A C G G A T T G A T G C A T G G A C T T G A G C A G G A C A T G A T C T G C A T C A T G C C G G A T T A C C G C A G G S C C G T A A
 32550 32540 32530 32520 32510 32500 32490 32480 32470
 β H P I C U T E W M D E C R R E P F K T R E G R E I T G H Q
 T C A T C C G A T C T G G T A C C G A T G G A T G A T G C C G C G C G A T C A A T C T G C G G A G G A A T T A C G G C G G C G T G G C A
 32460 32450 32440 32430 32420 32410 32400 32390 32380
 β S H P K R N M R H K M I O C A R L F G F A G G I Y D K D E
 G T G C A T C C H C A H A C G G A T T H C G T C A H A C C H T G A T G C A T G G C C T T G C G C T T G G G A T T T G C T G G T A T C T A T G C A C A G G A T G A
 32370 32360 32350 32340 32330 32320 32310 32300 32290
 β A E R I U E N T A Y T A E R Q P E R D I T P U N D E T M Q E
 H G C C G A G G C A T I S T G A A A T A C T G C A T G C A H G T C A G G C C G G A A C G C G A C A T C A C T C C G G T T A A C G T G A A A C C A T G C A G G A
 32280 32270 32260 32250 32240 32230 32220 32210 32200
 β I N T L L I A L D K T H D D P D L L P L C S Q I F R R D I R A
 G A T T H A C A T C T G C T G A T C G G C C T G G A T A A A C T G G G A T G A C G A C T T A T T C C G C T C T G G T T C C A G A T T T C G C C G C G A C A T C T G C
 32190 32180 32170 32160 32150 32140 32130 32120 32110
 exo M T P
 β S S E L T Q A E R U K H L G F L K Q K A A E Q K V A A *
 H T G C T G A G A T C G A C A G G C G G A A S C A T H A A G C T T G G A T T C C T G G A A C A G A H A G C G C G A G A C G A R G G T G C A G A C C G
 32090 32080 32070 32060 32050 32040 32030 32020 32010
 exo D I I L Q R T G I D U R A H U E Q G D D A W H K L R L G U I T
 G A C A T T C C T G C A G G G T H C C G G A T C G A T G T G H A C A G G G G A T O A T C G G T G C A C A R A T T A C G G C T G G C G T C A T C C C
 32010 32000 31990 31980 31970 31960 31950 31940 31930
 exo A S E U H N V I A K P R S G K K K H P D M K M S Y F H T L L A
 G C T T C A G A G T T C H C A H C G I G A T G A A A C C C G C T C C G G A A A G A G T G G C C T G A C A T G A A A A T G T C T C A T T C C A C C C T G C T T G C T
 31920 31910 31900 31890 31880 31870 31860 31850 31840
 exo E U C T G V A P E U N A K A L G K Q Y E N D A R T L L F E
 G A G G T T T G C C G G T G G C T C C G G A A G T T A A C G A C T G G C C T G G G G A A A A C A G A T C A G G A G C G A C G C C A G A C C C T G T T T G A A
 31830 31820 31810 31800 31790 31780 31770 31760 31750
 exo F T S G U V N U T E S P I I Y R D E S M R T A C S P D G L C S
 T T C A C T T C C G C G T G H A T G I T A C T G A A T C C C G A T A C T A T C G C A C G A A A G T A T G C G T A C C G C T G C T C T C C C G A T G G G T T T G C A G T
 31740 31730 31720 31710 31700 31690 31680 31670 31660
 exo D G N G L E L K C P F T S R D F M K F R L G G F E A I K S A
 G A C C G C A A C G C G C T T G A C T G A A T G C C C G G A T T T C A T G A G T I C C G G T C G G G T C G G T G C A T A A A G G C T C A G C T
 31650 31640 31630 31620 31610 31600 31590 31580 31570
 exo Y M A U Q O Y S M H U T R K N A H Y F A N Y D P R M K R E G
 T A C A T G G C C C A G G T G C A G T C A G C A T G T G G G G A A A A T G C T G A T C T G G C C A C T A T G A C C C C G T A T G A G G C G T G A A G G C
 31560 31550 31540 31530 31520 31510 31500 31490 31480
 exo L H Y U O E R D E K U I S F D E I U P E F I E K M D E A
 C T G C A T T A T G T G A T T G G G C G G A T G A A G T A C A T G C C G A G T T T C A G G A T C G T G G C C G G A G T T C A T C G A A A A A T G G C A G G G A
 31470 31460 31450 31440 31430 31420 31410 31400 31390
 exo L A E I G F U F G E O W R *
 orf60a M T H P H D N I R U V G A I T T F U Y S
 C T G G C T G A A A T T G G T T G A T T G G G G A C A T G G C G A T G A C G C T C T C A C G A T A A T A T C C G G T A T G G C G A A T C A C T T T C G T C A T T
 31380 31370 31360 31350 31340 31330 31320 31310 31300
 orf60a U T T K R G H U F P G L S U V I R N P L K A Q R L A E I N X
 C C G T T A C A A R G C G G A C T G G G T T A T T C C G G T T A T C C G A A T C C A C T G A A A C C A C G C G G C T G G C T G G A G G A G A T A A T A R T A
 31290 31280 31270 31260 31250 31240 31230 31220 31210
 orf63 M H K A S S V E L R T S I E M A H S L A Q I G I R F
 orf60a R G A U C T K H A L L S *
 A A C G A G G G G C T G T A G C A A A G C A T T C T C G T G A G T T A A G A A C G A T T C G A G A T G G C A C A T A C C C T G C T C A A A T T G G A A T C A G G I T
 31200 31190 31180 31170 31160 31150 31140 31130 31120
 orf63 U P I P U E T D E F H T L A A S L S Q K L E M M M U A K A E
 T G T G C C A A T A C C A G T A G A C A G A C G A G A A G A T T C A T G C C C T T C C C T T C A C A A A A G C T G G A A R A T G T G G T G C C G G A A A C C G A
 31110 31100 31090 31080 31070 31060 31050 31040 31030
 orf63 A D E R N Q U *
 orf61 M H R E A H E Q D N H G M H F S G S G L H I L C A Y A C R H G
 A G C A G A T G A G A G A A C C A G G T A T G A C A C C C A C G G A A T T T C A T G G C A G C G G G C T C A T A T G C T G T G C T G C A C A T G
 31020 31010 31000 30990 30980 30970 30960 30950 30940
 orf61 T C S M T P Q Q E N A H L R S I A R Q A N S E I K K S O T A C A G C A G
 G G A C T T G T C A T G A C C C T C A C G G A A A C C G C G C T C A G G C T A T T C G G C A T T C A T T G A A A T T G G A A T T G G A A T T G G A A T T G G A
 30930 30920 30910 30900 30890 30880 30870 30860 30850
 orf61 S G *
 T T T C C G G A T A A A C G C T G A T G A C A T T G C C G T G A C G T A C T G A A G A A G C A C C G C G A A A C C G T A A C G C T G A T G G G A T T C A C A C C G A C T C A T
 30840 30830 30820 30810 30800 30790 30780 30770 30760
 T T T A A G C C T G G C A T G C G G C T G T T A A T G G A A C G A T G A A A H G C A A A T C A T C A G G G G C T A C A G G C T C C T T T T
 30750 30740 30730 30720 30710 30700 30690 30680 30670
 T A T T A T C G C A T T C A C C C T C A M H G C T A T T A A C C A M H C T T C A G G G G A T T A A T G A A A G T G G C A G A C A T C A T T G A T T C A G G C T C A G A A A T A G
 30660 30650 30640 30630 30620 30610 30600 30590 30580
 A A G A T T A C A G C G C A C A C A G C H A T A A A A T G C G C G C C T G A C C A C C A G G C T A T A T C G G C A C T C A T T G T G A T G T G G C G A T C C G A
 30570 30560 30550 30540 30530 30520 30510 30500 30490
 T A G A T G A A C G A A G A C G C C T G G C T G G G A C T T G T G C A A G T T G C C A G G G A G A T C T G G A A C T T A T C A G T A A A C A G A G A G G T T
 30480 30470 30460 30450 30440 30430 30420 30410 30400
 E 22 U S E I N S Q A L R E A H E Q D N H D D W G F D A D L F H
 C G A G T G A G C G A A A T T A C T C T C A G G C A C T G C G T G A A G C G G C A C R G C A G G C A A T G C A G C A T G G G A T T T G A C C A G A C C C T T C C A
 30390 30380 30370 30360 30350 30340 30330 30320 30310
 E 22 E L U T P S I V U L E L L D E R E R N Q Q Y I K R R D Q E N F
 T G A T T G G T A A C C A C T C A G T G T G C T G G A A C C G G A A A C C G A C A T C A A C C G C C G C A C C A G G A G A C G A
 30300 30290 30280 30270 30260 30250 30240 30230 30220

E₂₂ D I A C L T U J F L R U E L E T A K S X L N E Q R E Y Y E G U
GSATATTGCGCTAHCAGTHGGGAAHACTCGTGTGAGCTTGAAACAGAACAAATCAAGACTCAAGCAGCAGTCGAGTATTACGAAGGTG
30210 30200 30190 30180 30170 30160 30150 30140
E₂₂ I S D G S X R I A K L E S N E V E D G H M O F L U U R P G
TATCTCGGATGGGAGTTHAGCGTATGCTGAAHCAACAGACTCCGTTGAGCAGGAAACAGCTTCCTGTTGCGGCATCCTGG
30120 30110 30100 30090 30080 30070 30060 30050
E₂₂ K T P U I K H C T G D L E E F L R Q L I E Q D P L T I D I
GAAGACTCTGTATCACAGCAGTCGACTGGTGAAGAGTTCTCGCGCTGAACTGAGACTTAATCGAACAGACCCGTTAGTAACTATCGACAT
30030 30020 30010 30000 29990 29980 29970 29960
E₂₂ I T H R Y Y G U G G O H U U O D A G E F L R H M M S D A G I R I
CATTCAGCGCTGCTATTCGGGGTCTGGGTTCAAGGAGTCAGGTGAGTGTATCTGCATATGATGTCGACGCTGCGATTCGGAT
29940 29930 29920 29910 29900 29890 29880 29870
E₂₂ K G E
CAAGGGAGAGTGAATCGGTTTGTAAAGATAACGCTTGTAAGAAATQCTGAATTTGCGCTGCTTCAGCGCATGCCAGAGTCGCTGAG
29850 29840 29830 29820 29810 29800 29790 29780
TGTCAAGATGACCGTACTAACATCGGGTTGHTGTTATCTTACTGTTCTTACATACATGCTGATACCGTTAGCTGAACAG
29750 29740 29730 29720 29710 29700 29690 29680
E₂₂ K S P F C N U R E M T D T U Q N Y Y H E Y G G H D
KCATTCATGCAAGGGTTTATAATAGTAGTCAATGAGTCTGAGCTGAGAAATGGGGCTTACATGTTGAGCTGAGGTTATTCATGAAAGGTAT
29670 29660 29650 29640 29630 29620 29610 29600
E₂₂.5 M S F N E L E S E Q K D W A L S M H L E R S G
KCATTCATGCAAGGGTTTATAATAGTAGTCAATGAGTCTGAGCTGAGAAATGGGGCTTACATGTTGAGCTGAGGTTATTCATGAAAGGTAT
29580 29570 29560 29550 29540 29530 29520 29510
E₂₂.5 K S N E D K S P F C N U R E M T D T U Q N Y Y H E Y G G H D
AAGTCATTAAGAACAAATCCGCAATGTCGAGAAAATGACTGATACCGTGGAAACATTCACGAGTAGGGCTTACATGAAAGGTAT
29490 29480 29470 29460 29450 29440 29430 29420
E₂₂.5 T C P L C T K H I D D
ACTTGCCCTCTGTCACAAACATATAGATGATTAACCCCATATATACATAACATCCTCGCACTCGGGGGATTTATTTATCTGACAT
29480 29390 29380 29370 29360 29350 29340 29330
CGCTACGGGGGGTTTGTGGAGATGATAATGCACTTCCGAGTCAGGGAGATGGAATGGAGAGCCATTCAACAGAGTTATCGA
29310 29300 29290 29280 29270 29260 29250 29240
AGCGGAGAACATCHAGCACTGTCACGACCACTGGATATGGGCGCATAGCACATCGACGACATACCAATATTGAAATTGAGAGACT
29220 29210 29200 29190 29180 29170 29160 29150

S_i

x₁₅ G A A G A A C C A C H G C G C C T G A T G G C G G T T T T C T G C G T G A T T G C G G A G A C T T T G C G A T G T A C T T G A C L T C A G G A G T G G A C C
29120 29110 29100 29090 29080 29070 29060 29050
x₁₅ R S P R P R S L E T U R R H U R E C R I F P P P U K D G R E
ACGCCAGCGACGCTCAGAGGAGCTGAGCTGAGCTGAGGATTTCCGACCTCCGGTTAAGGATGGAGAG
29040 29030 29020 29010 29000 28990 28980 28970
x₁₅ Y L F H E S A V U K V D L N R P U T G G L L K R . I R N G K K A
GTATCTGTTCCACGAACTGCGGTAAGGGTTGACTTAATCGACCACTGACAGGTTGCCCCCTTGAAGGAGATCAGAAATGGGAGAACG
28950 28940 28930 28920 28910 28900 28900 28870
x₁₅ K S T
S H E R R D L P P N L Y I R N G Y C Y R D D P R T G K E F
GAATGTCATGAGCCGGGATTACCCCCTAACCTTATATAGAAACATGGATATACTGCTACAGGGACCCAAAGGAGCGGTTAAGAGT
28860 28850 28840 28830 28820 28810 28790 28780
int G L G D R R I A I T E A I O A N I E L F S G H K H K P L T
TTGGATTACGAGAGCAGGGCAATCGCAATCACTGAGGCTATACAGGCCAACATGGAGTTATTTAGAGCACACACAAAGCCTCTGA
28770 28760 28750 28740 28730 28720 28710 28700
int A R T I N S P N S U T L H S H L D R Y E K I L A S R O J K E K
CAGCGAGAACTCARCACTGTCATTTCCGTTACCTGCTGAGGAAATCTGGCCACAGAGGAAATCAAGCAG
28680 28670 28660 28650 28640 28630 28620 28610
int T L I N Y M S K I K A P I R R G L P D A P L E D I T T K E I A
AGACACTCATTTACATGAGCAAATTAAGCACTAAAGGGGCTGCTGCTGAGGACATCACACAAAGGAGTT
28590 28580 28570 28560 28550 28540 28530 28520
int A M L H G Y I D E G K A S R K L I R S T L S D A F R E A I
CGGCAATGTCATGGGATACAGCAGGGGGGGCTGAGGCAATTAATCAGATCAACACTGAGGAGCTGGCATCCGGAGGGCA
28580 28490 28480 28470 28460 28450 28440 28420
int A E G H I T T N H V A A T R A A K S E U R R S R L T A D E Y
TAGCTGAGGGCCATAACACAAACATGCTGCTGCTGAGGAAATCAGAGGATAGGAGGACTGAGCTTGGCTGACGAGT
28410 28400 28390 28380 28370 28360 28350 28330
int L K I Y Q A E S S P C H L R L A M E L A V U T G O R V D
ACCTGAAATTATTCAGCAGCAATCACTGAGGCTGAGGACTGCAATGGCTGTGTTACCGGCAACAGAGTTGGG
28320 28310 28300 28290 28280 28270 28260 28250
int L C M K W H S D I U D G Y L Y U E Q S K T G U K I A I P T A
ATTATCGGAAATTAGGATGGCTGATATCTGAGTGGATCTTGTGAGCAATGGCTGAGGAAATGGGCAACAGAGTTGGG
28230 28220 28210 28200 28190 28180 28170 28160
int L H J D A L G I S M K E T L D K C K E I L G G E T I I A S T
CATTCGATATGATGCTCGGAAATCATGAGGAAACACTTGATAATGCAAGAGAGCTTGGGGAGAACCTAAATGCACTA
28140 28130 28120 28110 28100 28090 28080 28070
int R R E P L S S G T V U S R Y F M R A R K A S G L S F E G D P P
CTCGTCGCGAACCGCTTTCACCGCACAGTAAAGGTTTATCGGGAGCATCGCTCTCGGGGGGATCCGG
28050 28040 28030 28020 28010 28000 27990 27970
int T F H E L R S L S A R L Y E K Q I S D K F A Q H L L G H K S
CTACCTTCACGAGTTGGCGATTGCTGAGACTCTAGGAAACGAGATAAGGAGTTGGCTGACATCTCTCGGGCATAAGT
27960 27950 27940 27930 27920 27910 27900 27880
int D T M A S Q Y R D D R G R E H D K I E I K *
CGGACACCATGGCATCACAGTATCGTGTGAGCAGAGGAGGAGGAGGACAAATCAAAATGATTTTRTTTGACTGATAAGT
27870 27860 27850 27840 27830 27820 27810 27800
-----att site-----
GACCTGTTGCGAAACAAATTGATAAGCAATGCTTTTTATATGCACTTGTGAGTATAAGCTGAGGAGAACGTAACATGAT
27780 27770 27760 27750 27740 27730 27720 27710 27700
AAATATCAATATTAATTTAGATTTGCAAAAAACAGACTACATAATCTGTAACACACATATGCACTATGAAATCAACTA
27690 27680 27670 27660 27650 27640 27630 27620 27610

CTTAGATGGATTAGTGACCTGTAACAGAGCATTAGCCGAAAGGTGATTTTGTCCTTGGCTAATTTTGTCAACCTGTCGCA
 27600 27598 27580 27578 27560 27558 27548 27538 27528
 CTCAGAGAACGACAAGGCCTCGCACTTCAGTGCAHAGCTTGTGCAACCCACTACGCCCTGCATAACAGTAAGAAGATAGCAGTGA
 27510 27508 27490 27480 27470 27460 27450 27440 27430
 TGTCAAACGACGCAAGCTGCTTCTTCTTCAGCHACCTCCCACACCCAGCATGCATACTTCCGCCATAACTGTAGTGAATGTCG
 27420 27410 27400 27398 27390 27370 27360 27350 27348
 TATGAGCGAGGAGCGGAAAGTAAACCTATGAAARATGGCTACGAAAGCTCGGCTATCATCGCTTATTAGTAGCTTGAACACCTCT
 27330 27320 27310 27300 27290 27280 27270 27260 27250
 CGAGAGCTAAAGAGCTACTCGTCGGCATACTATATGCAATTAGACTATATCGTGGTATACAGTCGACCATGCAACATGAA
 27240 27230 27220 27210 27200 27190 27180 27170 27160
 TAACAGTGGGTATACCAAAAGGAAGCAGAAAGCTAAATATGGAAACATACGATGCCCGTTAGTTCAATACTAAATTAGG
 27150 27140 27130 27120 27110 27100 27090 27080 27070
 ATGGAAACGATATGTAATAGAGAGACTCTGIGTGGCCAAATAATTGGCTTATTTAAATAATTAAAGGT
 27060 27050 27040 27030 27020 27010 27000 26990 26980

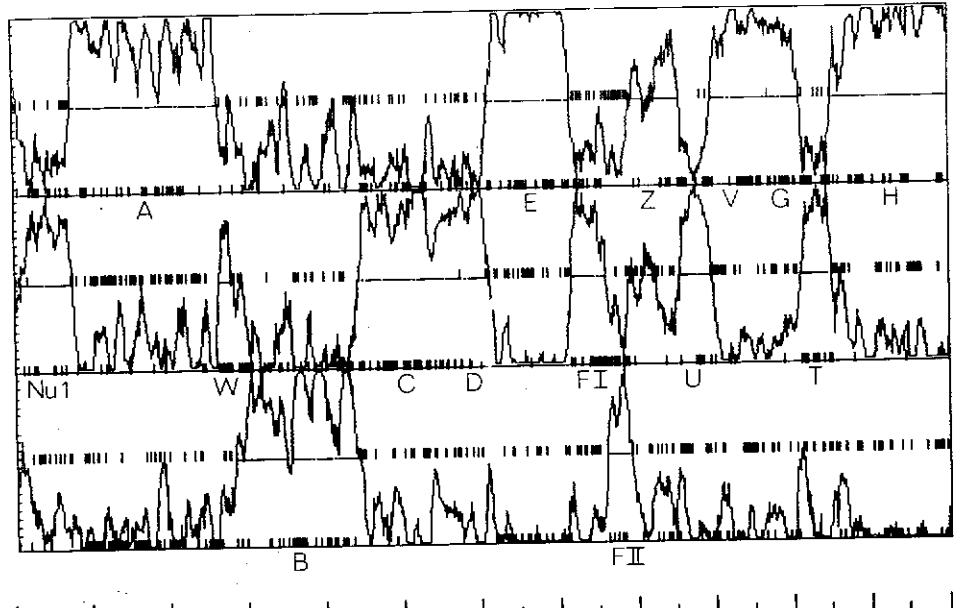
Ea59 M L E F S U I E R G G Y I P A U E K N K R F L R A D G H
 TACTATATGTTGGAGTTAGTGTATTAAGAAGGGGGGTTATTCCTCGACTAGAAAAAAATAAGGATTCCTACGAGCAGATGGTTGG
 26970 26960 26950 26940 26930 26920 26910 26900 26890
 Ea59 N D Y S F U T M F Y L T U F D E H G E K C D I G N U U K I G F
 AATGACTATCTCTTGTACATGTTATCTACTCTGTCATGAGCTGTTAGAAATCGHTATCGGAATGTTAAATTCG
 26880 26870 26860 26850 26840 26830 26820 26810 26800
 Ea59 U G O K E E L V S T I D K K F S Q L P E M F F S L G E S
 GTAGGTCAAAAGAACGAAAGCCTTATTCATTAAAGATAAAAATCGTCACCTCCTGAAATGTTTTCTTAGGTGAAAGC
 26790 26780 26770 26760 26750 26740 26730 26720 26710
 Ea59 I D Y Y U N L S K L S D G F K H N L L K A I Q D L U U W P N
 ATTGACTACTATGTAATCTCAAGAACGCTTAAACATACCTTCATTAAGCTTACGATGTTAGTAGTATGGCAAAAT
 26700 26690 26680 26670 26660 26650 26640 26630 26620
 Ea59 R L A C D I E N E T S L R G U T L S E F I H G Q F A R
 CGATAGCCGACATTAAGGCTCTAACACCTCATTAGAGGGGAACTTCAAGAAATTATGGACAGTCGCACGT
 26610 26600 26590 26580 26570 26560 26550 26540 26530
 Ea59 U L N G L P E L S D F H F S F N R K S A P G F S D L T I P
 GTGTTAATGTTGGCAGATGTCAGATTCTACTTATAGAAAGGAACTTCAAGAAATTATGGATTTAGCT
 26520 26510 26500 26490 26480 26470 26460 26450 26440
 Ea59 E U T V N S M P S T N I H A F I G R N G C G K T T I L N G M
 GAGGGTACGCTTAATCTATGCCACAGCAHATTATGCTTTATCGGGCGGATGGTGTAAACACAAATTGTAATGGAAATG
 26430 26420 26410 26400 26390 26380 26370 26360 26350
 Ea59 I G A I T N P E N N E Y F S E E N N R L I E S R I P K G Y F
 ATTGGTCAATCACCAACCCAGAAACAAATGAAATAGACTTCTGAAATATAGACITCGAGTCAGAAATCCCCAAAGGGATATTT
 26340 26330 26320 26310 26300 26290 26280 26270 26260
 Ea59 R S L V U S F S A F D P F T P P K E Q P D P A X G T Q Y F
 CGATTCGCTTTCAGTTGGTACCTTGTGATCTTACTCTCCCTAAAGAACACCTGCGGCAAGAACATACACT
 26250 26240 26230 26220 26210 26200 26190 26180 26170
 Ea59 Y I G L K H A A S N S L K S L G D L R L E F I S A F I G C M
 TATATGGACTAGAGATGTCGACCAATAGTAAATCACTACGATTCCTCGCTTAAAGGATTTCAATTTCAGCATTTATGGTGTATG
 26160 26150 26140 26130 26120 26110 26100 26090 26080
 Ea59 R U D R K R Q A L H L E A I K K L S S D E N F S N M E L I S L
 AGAGTAAATGAAAGAACGAACTCTGCTTGAAGCTATCAAACAAACTATGAGTAACTTCAAAATATGGAAACTCATCGACCTC
 26070 26060 26050 26040 26030 26020 26010 26000 25990
 Ea59 I S K Y E E L R R N E P O I U D D K F T K L F Y D N I Q
 ATTTCTAAATGAGAGTAAAGACCTAAATGAAACCAAGATTCAAGTGAGCATGATAAAATTCACAAATTGTTATGACAATATCCAG
 25980 25970 25960 25950 25940 25930 25920 25910 25900
 Ea59 K Y L L R M S S G H A I U L F T I T R L U D U V U G E K S L U
 AAATATCTTCGCAATGAGCTCTGGCATCTGCTTACCTTCTGCTTACGAAACATTAAAGGACTTACTCGATGCACGCAATGGT
 25890 25880 25870 25860 25850 25840 25830 25820 25810
 Ea59 L F D E P E U H L H P P L L S A F L R T L S D L L D A R N G
 TTATTCGATGACCAAGCTCTGGCATCTGCTTACCTTCTGCTTACGAAACATTAAAGGACTTACTCGATGCACGCAATGGT
 25780 25770 25760 25750 25740 25730 25720 25710 25700
 Ea59 U A I I A T H S P U V U L S Q E U P K S C M W K U L R S R E A I
 GTGCAAAATTCGCACTTCCCACTGAGTACTGAGGAGGTTCCCAAAATCTGCTGAGGAAAGCTTACGGTCAGAGAAAGLARIA
 25710 25700 25690 25680 25670 25660 25650 25640 25630
 Ea59 N I I R P D I E T F G E F N L G V T R E F U V F L L E U T N S G
 AAATATTCGGTCCGGATATGGACATTCGGTGAGAGCTTGGTCTTAACTCTGAGGTGTTTACTGAGTAACTGACAATATCTGG
 25620 25610 25600 25590 25580 25570 25560 25550 25540
 Ea59 Y H H L L S Q S U D S E L S Y E T I L K N Y N G Q I G L E G
 TACCAACCATATTCGGCTGGTGGATTAGCAGCTTCTTAAAGAACATTCAGGATTAAGGATTAATGTCAGATGACATTAG
 25530 25520 25510 25500 25490 25480 25470 25460 25450
 Ea59 R T V L K A M I M N R D E G K U Q * M K K L P L P A R T Y S E M
 CGAACCGTTTAAAGCGATGATAATGAAAGAGATGAAAGTAAGTACATGAAACAAACTACCTCTCCAGCGAAACTTATGGAAAG
 25440 25430 25420 25410 25400 25390 25380 25370 25360
 Ea59 L N K C S E G M M Q I N U R N F I T H F P T F L Q K E Q Q
 TGCTTAATGAGCTGGAGGTATGGAGATAATGTTAGAAATATTTCTACTACACTTCCCTTTGGAGAAACTTATGGAAAGAACAC
 25350 25340 25330 25320 25310 25300 25290 25280 25270
 Ea59 Y R I L S S T G Q L F T Y D R T H P L E P T T L U U G N L T
 AAATAGAATATGAGCTGAGCAGGTAGTATTTACCTACGAGCAGACACCCCTCTGAGCTACACCTTACGTTAGTGTAAACCTG
 25260 25250 25240 25230 25220 25210 25200 25190 25180
 Ea59 K V K L E K L Y E N N L R D K N K P A R T Y D D M L U S S
 CAAGGGTAAATGAGAAAGCTTTAGAAATATCTCCGAGAGATAAAACAAACCGCTGAAACATATTAGCATGACATGCTTGTTC
 25170 25160 25150 25140 25130 25120 25110 25100 25090
 Ea59 G E K C P F C G D I G O T K N I D H F L P I A H Y P E F S V
 CAGGTGAAATGTCGCTTGGTGTGATATGGAGCAGACAAATAATGACATTTCTCTTACGACATTATCTGAGTAACTTCTGG
 25080 25070 25060 25050 25040 25030 25020 25010 25000
 Ea59 M P I N L U P S C R D C H N M G E K G Q U F A U D E V H Q A I
 TGATGCCTTAAATTTAGTCCATGCGCCGAGCTGAGTATGGAGAGAAAGGTCAGTTTCGAGTAGATGGGGTACACCAAGCGA
 24990 24980 24970 24960 24950 24940 24930 24920 24910

As a further identification of reading frames, we have used the computer method of Staden & McLachlan (1982), which calculates the frequency with which codons occur in the different reading frames. The method is now contained in a more general analytical program called ANALYSEQ. Known genes are chosen and the frequency of occurrence of the different codons calculated and used as a standard. A region of the sequence is then scanned and the frequency of codons in each reading frame calculated and compared with that of the known genes. The results are expressed as "probability" that the reading frame is coding for a protein, assuming that coding regions will have a similar codon distribution. Figure 3 shows results using ANALYSEQ on the lambda sequences. In these diagrams the probability is plotted against the position in the sequence for the three different reading frames. Thus the presence of a coding region is indicated by a peak in a particular reading frame. In most cases the results are quite clear and confirm the expected positions of the genes. We have also used these diagrams to help identify the position of protein initiation. At the beginning of the coding region there is usually a sharp rise in the probability curve and we have measured the positions where this crosses the 50% probability mark. These figures are given in Table 3. It appears that the distribution of codons differs in the different transcription units of lambda (see Table 5). Thus, when the ANALYSEQ results were calculated using as standard a gene from the left-hand end of the sequence (e.g. gene *J*), they were clear for other genes in the left-hand end but less so for those in the right-hand end; and, conversely, good results were obtained for the right-hand end when the standard was from that end (e.g. genes *O* and *P* or *exo* and β). Differences between the two ends of lambda have been noted previously, particularly in the nucleotide composition, the left-hand end being G+C-rich and the right relatively A+T-rich.

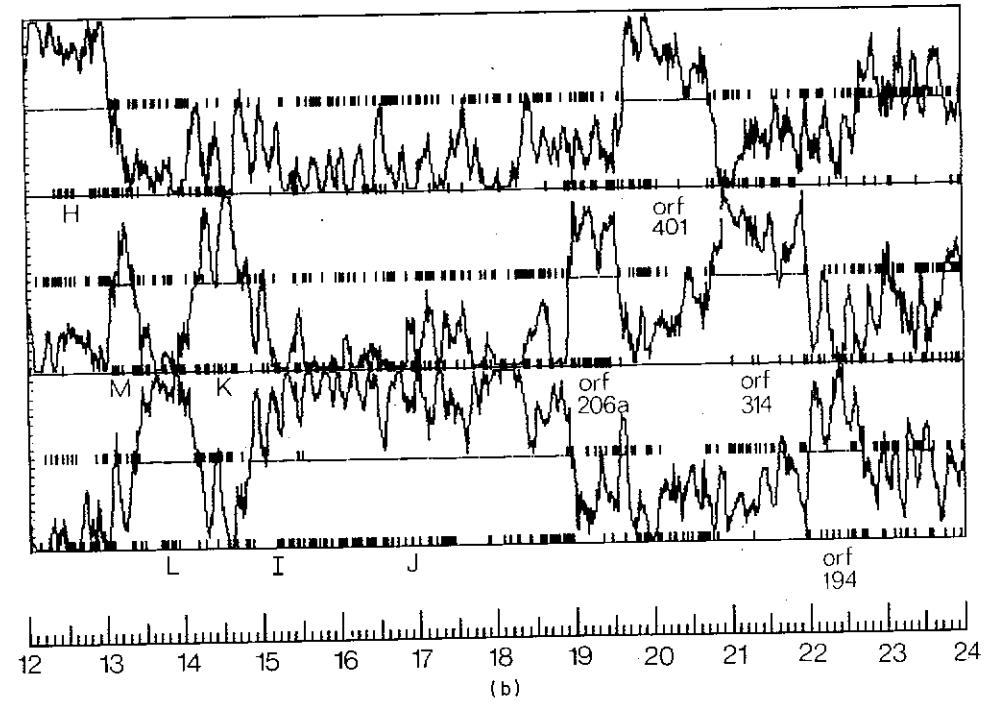
(c) Protein initiation sites

Table 3 lists the proposed start points of the gene products. In many cases they are not rigorously established and depend largely on choosing an appropriate AGT or GTG preceded by the best "Shine and Dalgarno" (SD) sequence (Shine & Dalgarno, 1974). Various attempts to define more precisely the nucleotide sequence important for initiation in the vicinity of initiation sites have met with only limited success. In an extensive study on 124 initiation sites, Stormo *et al.* (1982) have shown that other sequences both before and after the ATG may contribute to the formation of an initiation signal and have suggested additional rules that may be used to identify sites. These "Stormo" rules have also been taken into account in this work.

An interesting feature of the lambda sequence is the frequent occurrence of overlapping termination and initiation sites, particularly in the sequence A-T-G-A, where the TGA is the termination codon for one reading frame and ATG the initiation of the next. These structures are listed in Table 4. The initiators are usually preceded by an SD sequence, but more characteristic is a very purine-rich region separated from the ATG or GTG by a few nucleotides. This effect is particularly marked in the *nin* region (see below), where there is a continuous series of 13 overlapping reading frames. It is not known what the function of this

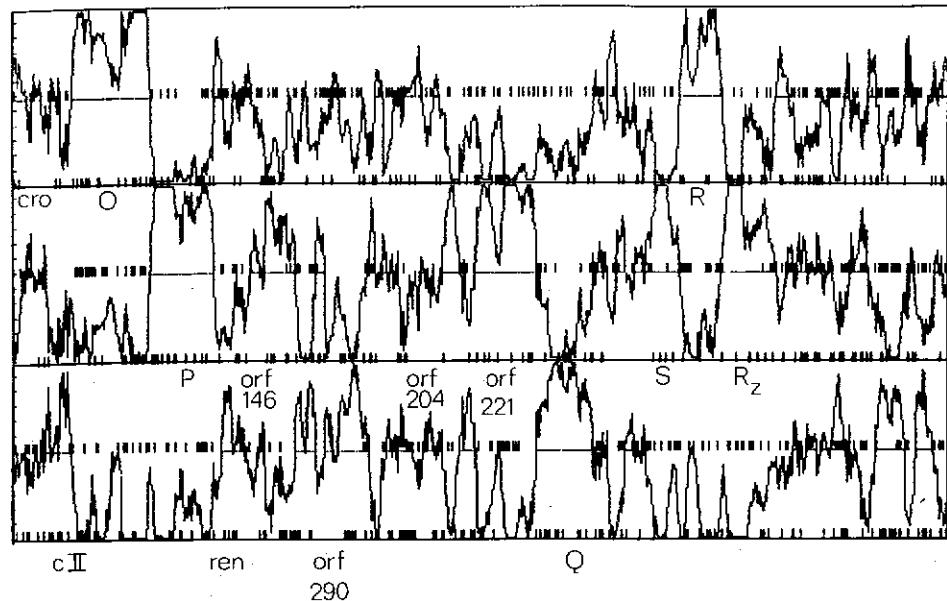


(a)

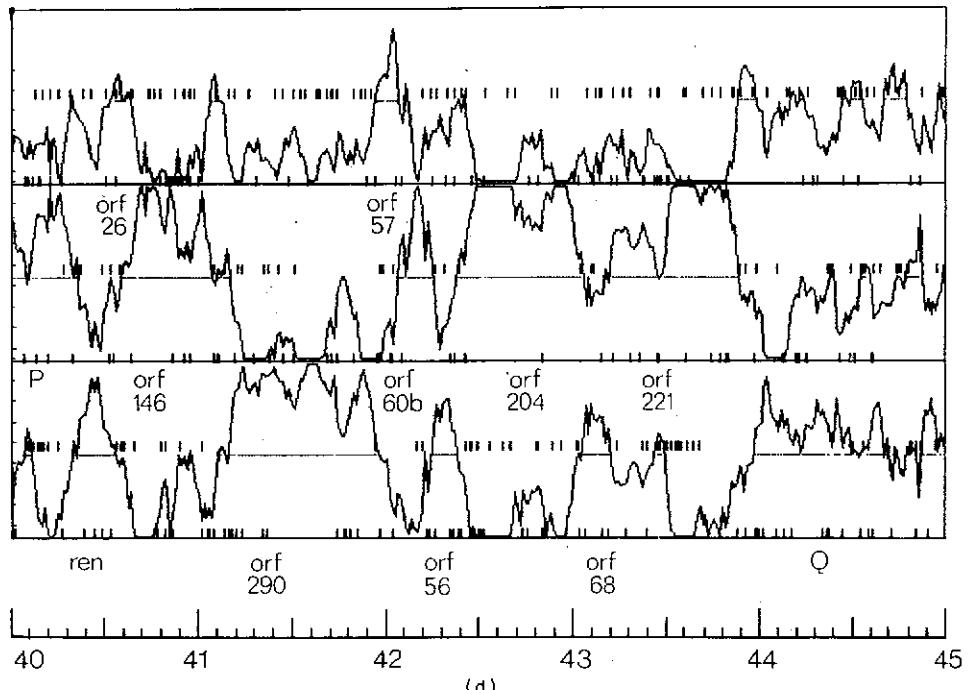


(b)

FIG. 3. Gene predictions for the DNA of lambda using the codon preference method of Staden & McLachlan (1982), which is one function of ANALYSEQ, a general sequence analysis program. The x axis represents sections of the lambda sequence (Figs 1 and 2) and the probability of coding is plotted in the y direction. The method assumes that the codon preferences for neighbouring genes are similar, and hence that the codon usage of some of the known genes can be used as standard.

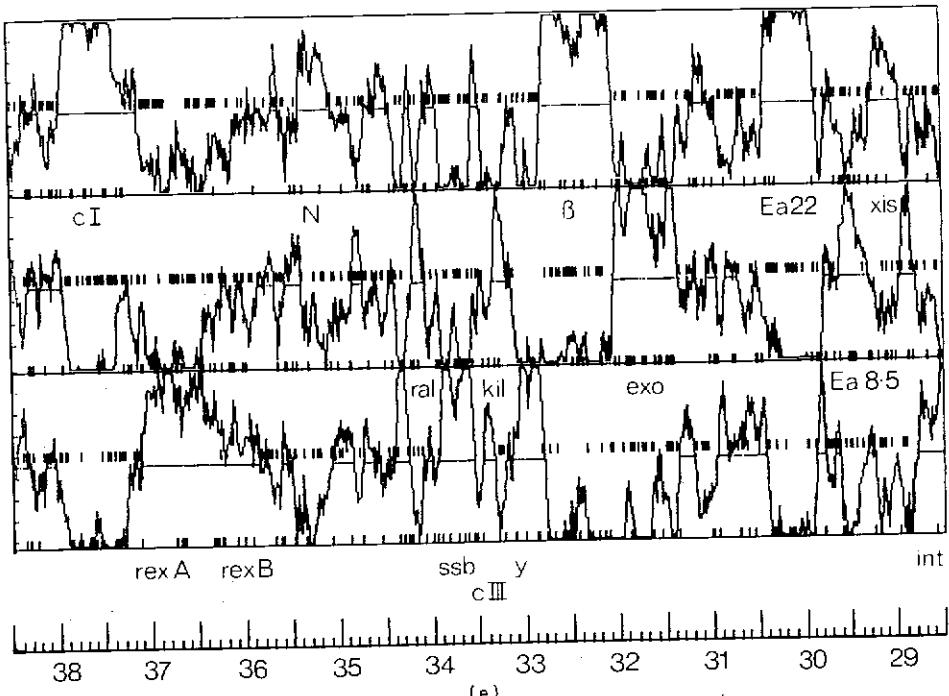


(c)

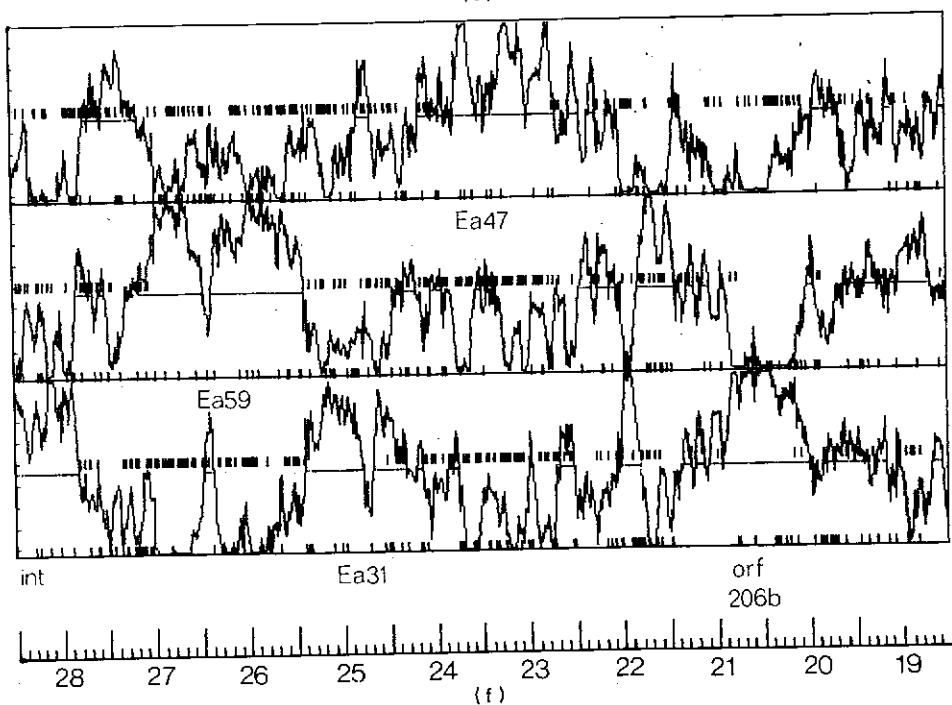


(d)

Probabilities of coding are calculated by sliding a window of 35 codons along the sequence, one codon at a time. For every position of the window, the codons found in each of the 3 reading frames are compared with those in the standard and the corresponding probabilities of coding calculated. The probabilities for each of the 3 frames have been plotted, one above the other, every 5 codons. A solid horizontal line at the mid-point of a reading frame (this is at the 50% level of probability) indicates



(e)



(f)

which of the 3 frames is most likely to be coding. The initiation codons ATG, GTG are marked as vertical bars along the base of each plot and the termination codons as vertical bars along the 50% level. (a) and (b) The left arm: genes *A*, *B*, *C*, *D*, *E*, *FI*, *V*, *H*, *L*, *J* used as standards; (c) the right arm, using genes *O*, *P*, *Q*, *R* as standards; (d) positions 40,000 to 45,000 of the right arm using reading frames *orf146*, *orf290* and *orf204* as standards; (e) and (f) the central region using genes *cI*, *rexA*, *ssb*, γ , β , *exo*, *Ea59* and *Ea31* as standards.

TABLE 3
Most probable initiation sites for the λ proteins

Gene	Proposed start		Shine & Dalgarno sequence†	Status (also see the text)
	From sequence	From ANALYSEQ		
<i>NuI</i>	191	210	G-G-A(7)ATG	One of 2 possible
<i>A</i>	711	770	G-G-G-T(8)GTG	Uncertain
<i>W</i>	2633	2660	G-G-A-G(8)ATG	Established by amber mutation
<i>B</i>	2836	2880	A-G-G-A(8)ATG	One of 2 possible
<i>C</i>	4418	4430	A-G-G-A-G-G(8)GTG	Best of 6 possibilities
<i>D</i>	5747	—	A-A-G-G(6)ATG	Amino acid sequence
<i>E</i>	6135	6150	G-G-A(7)ATG	Amino acid sequence
<i>FI</i>	7202	7170	G-G-A-G-G(7)ATG	Most probable of 2
<i>FII</i>	7612	7620	G-G-A-G-G(6)GTG	Most probable of 2
<i>Z</i>	7977	7960	A-A-G-G-G(6)ATG	Most probable of 3
<i>U</i>	8552	8550	G-G-A(13)ATG	Only possible
<i>V</i>	8955	8970	G-A-G-G(7)ATG	Amino acid sequence
<i>G</i>	9711	—	G-G-A-G(9)ATG	Most probable of 2
<i>T</i>	10,115	10,110	G-G(6)GTG	Uncertain
<i>H</i>	10,542	10,520	A-G-G-A-G-G(10)ATG	Amino acid sequence
<i>M</i>	13,100	13,090- 13,160	G-G-A-G-G-T(8)ATG	Most probable of 5
<i>L</i>	13,429	13,410	A-G-G-G-T-T(9)AGT	Only probable
<i>K</i>	14,276	14,250	G-G-A-G-G(12)ATG	Uncertain
<i>I</i>	14,773	14,860	G-A-G-G-T(13)ATG	Most probable of 3
<i>J</i>	15,505	—	G-G-A-G(9)ATG	Most probable
<i>orf206a</i>	18,965	18,960	G-A-G-G-T(8)ATG	Most probable of 3
<i>orf401</i>	19,650	19,660	G-G-G(6)ATG	Most probable
<i>orf314</i>	21,029	20,845	A-G(12)ATG	Most probable (see the text)
<i>orf194</i>	21,937	21,990	G-A-G-G(10)ATG	Only probable
<i>cro</i>	38,041	37,980	A-A-G-G-A-G-G-T(6)ATG	Schwarz <i>et al.</i> (1978); Roberts <i>et al.</i> (1977)
<i>cII</i>	38,360	38,370	A-A-G-G-A(11)ATG	Schwarz <i>et al.</i> (1978); Roberts <i>et al.</i> (1977)
<i>O</i>	38,686	38,680	A-G-G-A-G(6)ATG	Schwarz <i>et al.</i> (1978)
<i>P</i>	39,582	39,560	G-G-G-T(9)ATG	Schwarz <i>et al.</i> (1980)
<i>ren</i>	40,280	40,370	A-A-G-G-A-G(8)ATG	Only probable one
<i>nin</i> region			See the text and Table 4	
<i>Q</i>	43,886	43,910	G-G-A-G(8)ATG	Daniels & Blattner (1982); Petrov <i>et al.</i> (1981)
<i>orf64</i>	44,621	—	G-G-G-T(9)ATG	Daniels & Blattner (1982)
<i>S</i>	45,186	45,130	G-G-G-G-T(8)ATG	Daniels & Blattner (1982)
<i>R</i>	45,493	45,490	G-G-A-G(7)ATG	Amino acid sequence (Imada & Tsugita, 1971)
<i>R_z</i>	45,966	45,960	G-A(7)ATG	Most probable
<i>cl</i>	37,940	37,950	‡	Ptashne <i>et al.</i> (1976)
<i>rexA</i>	37,114	37,100	G-G-A(7)ATG	Only possible
<i>rexB</i>	36,259	—	A-G-G-A-G(6)AGT	Landsmann <i>et al.</i> (1982) One of 2 possible
<i>N</i>	35,438	35,390	G-G-A(10)ATG	Landsmann <i>et al.</i> (1982) Franklin & Bennett (1979)
<i>ral</i>	34,287	34,210	A-G-G-A(8)ATG	Ineichen <i>et al.</i> (1981)
<i>ssb</i>	33,904	33,930	A-G-G-A(9)ATG	Ineichen <i>et al.</i> (1981)
<i>cIII</i>	33,463	33,460	A-A-G-G-A-G(7)ATG	See the text
<i>kil</i>	33,330	33,340	A-G-G-A-G(10)ATG	See the text
<i>γ</i>	33,112	33,100	A-G-G-A-G(8)ATG	See the text

TABLE 3 (continued)

Gene	Proposed start		Shine & Dalgarno sequence†	Status (also see the text)
	From sequence	From ANALYSEQ		
β	32,810	32,800	G-A-G-C(12)AGT	Only probable
<i>exo</i>	32,028	32,040	A-A-G-G-C-G(6)ATG	Most probable
<i>Ea22</i>	30,395	30,430	G-A-G-G(8)GTG	Only probable
<i>Ea8·5</i>	29,655	29,600	A-A-G-G-A(8)ATG	Only probable
<i>xis</i>	29,078	29,250	G-G-A-G(9)ATG	Hoess <i>et al.</i> (1980)
<i>int</i>	28,882	28,780	A-G-G-A(6)ATG	Hoess <i>et al.</i> (1980)
<i>Ea59</i>	26,973	27,040	A-A-G-G(7)ATG	Only probable
<i>Ea31</i>	25,399	25,410	A-A-G-G(8)ATG	One of 2 possible
<i>Ea47</i>	23,918	23,750	G-G-A-G(7)ATG	Most probable
<i>orf206b</i>	20,767	20,900	A-A-G(8)GTG	See the text

† The figures in parentheses are the number of nucleotides between the first residue of the initiating ATG or GTG and the A of the G-G-A-G-G in the SD sequence, or the corresponding position in the SD sequence if there is no A there.

‡ The initiating ATG for the *cJ* protein is at the 5' terminus of the mRNA when the mRNA is transcribed from p_M' and consequently has no SD sequence.

characteristic overlapping structure is, though it is tempting to suggest that it could allow the ribosome to read through from one frame to the next. Oppenheim & Yanofsky (1980) showed that there was "translational coupling" between the *trpD* and *trpE* genes, which are connected by an overlapping termination and initiating codon in the sequence T-G-A-T-G (i.e. that the expression of *trpD* was dependent on the expression of *trpE*) and suggested that this effect may be due to a ribosome, or a component of it that terminates on the first reading frame, being able to initiate more efficiently on the second one than a free ribosome. We have no evidence as to whether there is translational coupling between the lambda proteins that are connected by overlapping termination and initiation codons, but it is certainly an attractive hypothesis.

(d) Sizes of the proteins

Table 2 shows the molecular weights of the various proteins calculated from the most probable reading frames shown in Figure 2. These are compared with values found previously by sodium dodecyl sulphate/acrylamide gel electrophoresis. The agreement seems reasonably good. The most notable exception is *G*, which is discussed below. There is a similar discrepancy in the case of the *xis* gene (Hoess *et al.*, 1980), where a molecular weight of 32,000 was suggested by acrylamide gel electrophoresis (Hendrix, 1971). In this case there seems to be ample proof for the smaller value obtained from the DNA sequence.

(e) The codon CUA

An interesting feature of the left arm is that the codon CUA is completely absent from its 7476 codons. This is not the case in the other regions, where there are 40 CUAs in 6220 codons, or in other *E. coli* systems. CUA is a leucine codon and is

TABLE 4
Overlapping initiation and termination sites of λ proteins

Terminator	Initiator	Purine content†	Gap size‡	Structure§
<i>A</i>	<i>W</i>	10/11	0	<u>ATGA</u>
<i>W</i>	<i>B</i>	11/15	8	<u>ATGA</u>
<i>B</i>	<i>C</i>	14/14	0	<u>GTG(14)TAA</u>
<i>Z</i>	<i>U</i>	13/16	2	<u>ATGA</u>
<i>G</i>	<i>T</i>	6/6	0	<u>GTG(13)TGA</u>
<i>T</i>	<i>H</i>	8/8	6	<u>ATG(2)TGA</u>
<i>H</i>	<i>M</i>	8/9	2	<u>ATGA</u>
<i>M</i>	<i>L</i>	11/14	3	<u>TGATG</u>
<i>K</i>	<i>I</i>	14/19	0	<u>ATG(97)TGA</u>
<i>orf314</i>	<i>orf194</i>	7/12	0	<u>TAA TG</u>
<i>O</i>	<i>P</i>	7/8	3	<u>ATGA</u>
<i>P</i>	<i>ren</i>	11/12	3	<u>ATGA</u>
<i>ren</i>	<i>orf26</i>	9/11	0	<u>ATGA</u>
<i>orf26</i>	<i>orf146</i>	14/14	2	<u>ATGA</u>
<i>orf146</i>	<i>orf290</i>	25/29	4	<u>ATGA</u>
<i>orf290</i>	<i>orf57</i>	10/10	4	<u>ATGATG </u>
<i>orf57</i>	<i>orf60b</i>	13/13	3	<u>GTG(28)TAA</u>
<i>orf60b</i>	<i>orf56</i>	11/12	0	<u>GTGA</u>
<i>orf56</i>	<i>orf204</i>	10/10	0	<u>ATGATG(2)TAA </u>
<i>orf204</i>	<i>orf68</i>	15/18	4	<u>ATGA</u>
<i>orf68</i>	<i>orf221</i>	14/16	0	<u>ATG(17)TGA</u>
<i>orf221</i>	<i>Q</i>	10/10	3	<u>ATGA</u>
<i>S</i>	<i>R</i>	16/19	0	<u>ATG(11)TAA</u>
<i>R</i>	<i>R₂</i>	9/11	3	<u>ATGA</u>
β	<i>exo</i>	14/17	1	<u>ATGA</u>
<i>exo</i>	<i>orf60a</i>	10/12	2	<u>ATGA</u>
<i>orf60a</i>	<i>orf63</i>	21/25	4	<u>ATG(22)TAA</u>
<i>orf63</i>	<i>orf61</i>	12/14	0	<u>ATG(14)TGA</u>
<i>xis</i>	<i>int</i>	12/14	0	<u>ATG(17)TGA</u>
<i>Ea59</i>	<i>Ea31</i>	14/16	4	<u>ATGA</u>

† Number of purine residues/total nucleotides in purine-rich region preceding the initiator.

‡ Number of nucleotides between the purine-rich sequence and the initiator.

§ The initiator is underlined and the terminator overlined. Where appropriate, the number of nucleotides between the initiator and the terminator is shown in parentheses. The hyphens have been omitted for clarity.

|| These sequences have 2 consecutive ATG triplets, either of which could be the initiator.

recognised by a unique tRNA, which is present in a relatively low concentration (Ikemura, 1981b). It seems that this unusual feature must have some biological significance. Possibly, at the late stages of infection CUA becomes a limiting codon so that the structural proteins can be synthesized at the expense of other proteins of the bacteriophage or host that contain CUA codons. Such a situation could be

achieved if the CTA tRNA were inactivated by one of the late proteins of lambda and it would be interesting to know if this is indeed the case. In this connection, Yudelevich (1971) has shown that a different leucine tRNA, which recognizes the codon CUG, is split into two fragments by an enzyme produced by bacteriophage T4 when it infects the cell.

(f) Yield of the protein products

There are probably 25 proteins coded for by the left arm of lambda (positions 1 to 22,560). These are produced in very different amounts (Muriel & Siminovitch, 1972; Hendrix, 1971) but are probably made from a single transcriptional unit initiated at the late promoter, p'_R . Ray & Pearson (1974, 1975) have produced evidence that the various regions of the RNA are present in equal amounts during the production of the proteins; therefore, the varying yield of proteins is probably due to differences in translation. There is no obvious relationship between the nature of the initiation sites and the yield of proteins. This does not of course preclude the possibility that initiation is involved, as the characteristics of the sites are still not well-understood and it may be that secondary structures are concerned in the control of initiation (Iserentant & Fiers, 1980), although this is not apparent from the primary structure. Another possible mechanism of translational control may be the frequency of certain codons that correspond to minor transfer RNAs and may limit the rate of protein synthesis (Ikemura, 1981a,b). Table 5 lists the frequency of rare codons in the proteins of the left arm and the approximate yield of some of these proteins (Muriel & Siminovitch, 1972). It can be seen that there is a general correlation; the codons are more frequent in proteins that are produced in small amounts than in the major components. The figures are probably not accurate enough to define an exact relationship, but it seems likely that this may be an important mechanism for translation control.

As mentioned above, the two arms have different codon distributions and the rarer codons also differ. Only CTA and ATA seem to be rare in both. In the right arm, the rare codons seem to be related mainly to its being an A + T-rich region. For instance, the rarest codon is CCC, and UUG is considerably rarer than UUA, whereas the reverse is true in the left arm.

(g) Possible secondary structures

Although the new sequences reported here have not been rigorously searched for possible secondary structures, it was noticed that there are a number of possible looped structures in the left arm occurring in intergenic regions before the initiation sites of certain proteins. These are shown in Figure 4. Some of these (preceding *E*, *J*, *orf401*; Fig. 1) are followed by a run of T residues, which is a characteristic of transcription termination sites. These would seem to be unlikely places for such sites, since all the proteins are believed to be translated from one long mRNA initiated at the p'_R promoter, and in any case they would probably be anti-terminated by the *Q* protein (Schechtman *et al.*, 1980). Another possibility is that these loops may be sites for RNA processing by ribonuclease III. It is interesting that four of these loops (*E*, *Z*, *J*, *orf401*) contain the sequence C-C-G-C-C.

TABLE 5
Frequency of certain rare codons in the left arm of λ

Gene	Frequency of codon per 100							Relative concentration†
	AUA	UUA	CAA	CGA	CGG	AG _G	UCA	
<i>NuI</i>	1.10	0	0	1.65	2.75	1.10	0.55	
<i>A</i>	0.62	1.16	0	0.31	1.09	0.62	0.47	1
<i>W</i>	0	0	0	4.3	1.45	2.9	0	
<i>B</i>	0.56	0.38	0.19	0.56	2.4	0.38	0.93	
<i>C</i>	0.24	0.24	0.48	0.48	1.7	0	0.73	
<i>Nu3‡</i>	0	0	0.57	0.57	0	0	1.7	230
<i>D</i>	0	0	0	0	0.9	0	0	870
<i>E</i>	0	0	0.88	0	0	0	0.58	470
<i>FJ</i>	0	0	1.5	0	0	0.76	0.76	360
<i>FII</i>	1.7	0	0	0	4.2	0	1.7	
<i>Z</i>	2.1	0	1.6	0.52	3.1	3.1	1.6	
<i>U</i>	0	0	2.3	0	1.5	0	2.3	375
<i>V</i>	0	0	0	0	0	0	0.41	265
<i>G</i>	0	0	0	0	2.1	0.71	2.1	18
<i>T</i>	0	0	0	0.69	1.4	0	1.4	
<i>H</i>	0.12	0	0.59	0.35	1.6	0.70	0.59	12
<i>M</i>	0.92	0	0	0.92	1.8	0.92	0	
<i>L</i>	0	0	0	0.43	0.86	0	0	44
<i>K</i>	0	0	0.50	1.0	3.0	0.50	1.5	14
<i>I</i>	0.45	1.3	0.45	0.45	1.3	0.90	2.2	
<i>J</i>	0.88	0.09	0.26	0.26	1.1	0.71	0.35	13
<i>orf206a</i>	0.49	0.49	0.97	0	0.97	0.49	0.49	
<i>orf401</i>	0.75	0	0.75	0.25	0.75	1.2	5.7	
<i>orf314</i>	0	1.3	0.32	0.32	0	0.96	1.6	
<i>orf194</i>	0.52	1.5	0.52	0	2.5	0.52	0.52	
	AUA	UUA	CAA	CGA	CGG	AG _G	UCA	CUA
Total for left arm	0.45	0.21	0.38	0.38	1.38	0.62	1.06	0
Total for central region	1.91	2.35	1.41	0.72	0.44	2.43	2.07	0.58
Total for right arm	0.65	0.77	1.77	1.50	0.88	2.27	1.12	0.73

† Relative number of copies of protein in cells 45 min after infection with lambda (Muriel & Siminovitch, 1972).

‡ Assuming initiation at position 5132.

(h) The reading frames

In this section we discuss certain features of some of the individual reading frames. Other aspects of the sequence will be discussed in a separate paper (Daniels *et al.*, 1982).

NuI

There are two possible starts at positions 191 and 269. Both have satisfactory SD sequences that satisfy Stormo rule 2D. We have chosen the former from the ANALYSEQ result.

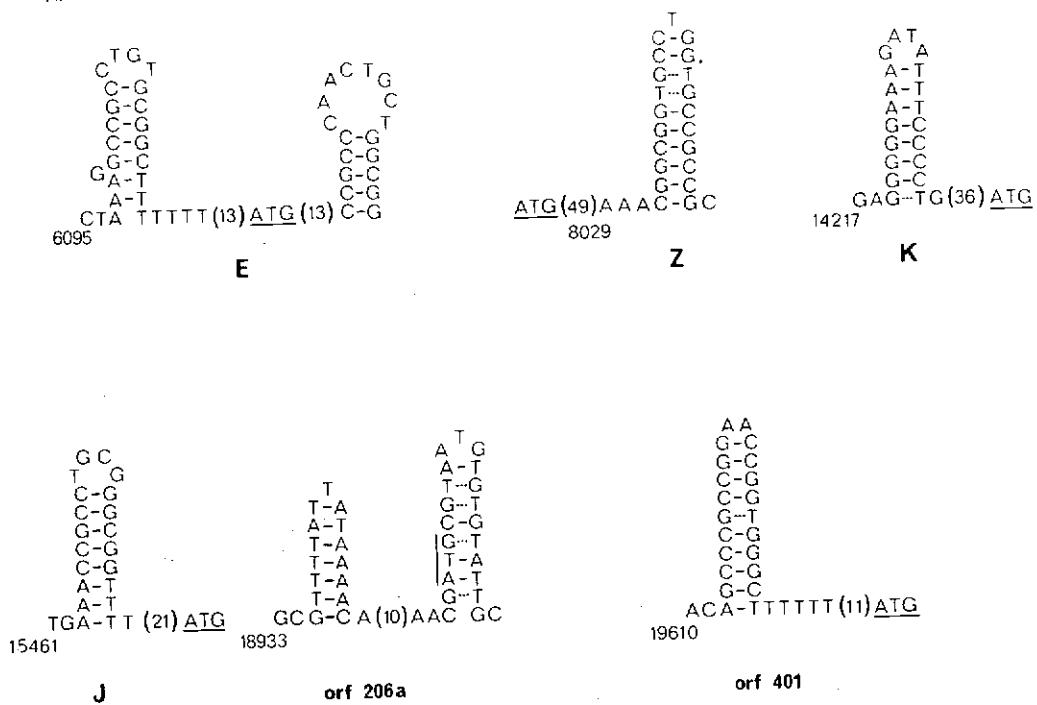


FIG. 4. Possible looped structures found near the initiating sites of some of the genes in the left arm of lambda. The initiating ATG is indicated by underlining and its relative position is shown by the numbers of nucleotides in parentheses. Base-pairing is indicated by dashes and sequence hyphens have been omitted for clarity.

A

There are five possible sites. The ATG triplets at positions 891 and 900 have very poor SD sequences, whereas the GTG triplets at 930 and 933 have good ones. However, the ANALYSEQ results suggest that this is a coding frame back to 770 and there is a GTG at position 771 preceded by G-G-G-G-T, which is probably an acceptable SD sequence involving a single G-U base-pair.

W

G. F. Hong (unpublished results) has shown that a strain of lambda with a mutation in gene W has T at position 2642 instead of the normal C. This establishes the ATG at 2633 as the only possible start.

B

The start of the main protein product *B** has been shown by amino acid sequencing to be at position 2902 (Walker *et al.*, 1982). This is a processed protein and so the initiation site must precede this. There are two possible ATG triplets at 2836 and 2878, both of which have SD sequences. That at 2836 is better and obeys Stormo rule 2B. It also overlaps with the termination codon of the previous protein (*W*) in the sequence A-T-G-A. This seems to be a feature of several of the lambda (*W*) genes.

protein starts, though in this case it is not preceded by the usual purine-rich sequence (see Table 4).

C

The GTG at position 4418 is preferred to five other possible start sites as it has a much better SD sequence. It also overlaps the end of the *B* protein and is preceded by a run of 14 purines. It is present in the sequence G-T-G-A but, unlike most other sequences of this type (see Table 4), the TGA is not the terminator of the previous protein. The size of this protein is considerably smaller than that estimated by acrylamide gel electrophoresis (see Table 2). Even if the first possible GTG (position 4283) were used, the molecular weight would be only 48,475.

Nu3

Nu3 is believed to be a protein of 19,000 M_r , initiating within the *C* gene and in the same reading frame. There are a number of ATG and GTG triplets that would give a protein of about this size, but none has a satisfactory SD sequence. Better sites are at positions 5219 and 5342; they would give proteins of 29,555 and 25,255 M_r , respectively.

E

In spite of being one of the major protein components, it has a rather poor SD sequence (GGA), but does obey Stormo rule 2C. It is preceded by an interesting structure, which has the characteristics of a transcription termination site (Fig. 4). There is also a smaller loop after the ATG containing the sequence C-C-G-C-C, which is also present in the other loop. It is tempting to suggest that in some way these structures might play a part in translation control and ensure that what appears from its primary structure to be a poor initiation site may produce relatively large amounts of protein.

Z

Although the reading frame is very clear here with only one possible start site with an SD sequence, the *Z* gene gives a rather poor peak in the ANALYSEQ program (see Fig. 3), suggesting an unusual codon composition. This may be partly due to the relatively high frequency of the rare codons (see Table 5) and to the small amount of asparagine and glutamine codons. There is a possible looped structure containing the sequence C-C-G-C-C within the coding region (see Fig. 4).

U

There is a very poor SD sequence, but the ATG overlaps the terminator of gene *Z* and is preceded by a purine-rich sequence (see Table 4).

G

Murialdo & Siminovitch (1972) identified the *G* gene product as a protein of about 33,000 M_r , or approximately 330 amino acids, whereas the reading frame in this position contains only 140 amino acids. The DNA sequence has been checked carefully. If there were an extra nucleotide between positions 10,068 and 10,133

there would be a reading frame of 279 amino acids; however, there would then be no reading frame for the *T* gene.

T

There are three possible initiation sites, at positions 10,115, 10,160 and 10,184, all of which have only G-G as an SD sequence. The one at position 10,115 is chosen as it overlaps the end of the *G* gene and is preceded by a purine-rich sequence (see above).

H

The site at 10,542 corresponds to the N terminus of the *H** protein (Walker *et al.*, 1982; Hendrix & Casjens, 1974). Thus the processing of *H* to *H** does not involve removal at the N terminus.

M

The start from the ANALYSEQ program is not clear here, but there is only one initiator with a good SD sequence and this overlaps the terminator of *H*. This small protein has an unusually high tryptophan content (5 residues out of 109). This may account for the low peak in the ANALYSEQ program.

K

There are two possible start sites. The one at 14,276 has a good SD sequence but it is 12 residues from the ATG; the other one at 14,306 has AGG seven residues from the ATG. Neither obeys Stormo rules 2 or 6. The former has been chosen rather arbitrarily. The coding sequence is preceded by a gap of 148 nucleotides that contain a possible loop structure (see Fig. 4). The protein has a somewhat unusual composition with a high tryptophan and histidine content (9 and 12 residues, respectively, out of 199). This may partly account for the poor peak in the ANALYSEQ program.

I

The ATG at position 14,773 is selected as the only initiator with an SD sequence, although it is separated by 13 residues. Other possibilities are an ATG at 14,812 or a GTG at 14,872. The latter overlaps the terminator of the *K* gene but is not preceded by a run of purines, which seems to be a characteristic of such structures. If position 14,773 is correct, there is an overlap of 103 nucleotides with the *K* gene.

J

There are several GTG triplets farther downstream of the probable start site, but these have poorer SD sequences. There is a gap of 61 residues between the *I* and *J* genes containing a looped structure followed by a run of T residues, which thus resembles a transcription terminator (see Fig. 4).

orf206a

There are two GTG triplets at positions 19,118 and 19,142 that could be initiation sites, but these have poorer SD sequences than the ATG at 18,965. This site is within a possible looped structure (see Fig. 4).

orf401

There is no obvious start site. Besides the ATG at 19,650, there are several GTG triplets farther downstream, but the ANALYSEQ result makes the former more probable. Position 19,650 is preceded by a looped structure resembling a transcription termination site (see Fig. 4). An unusual feature of this reading frame is the high content (5.7%) of the codon UCA. This is normally a rare codon, no other frame on the left arm having more than 2.3% (Table 5). There are two quite long open reading frames on the opposite strand in this region. One (20,744 to 19,876) has no good initiation site and gives a trough with ANALYSEQ; the other (20,955 to 20,149), which we call *orf206b*, gives a good peak and has a possible initiation GTG at position 20,767 (Fig. 5). Translation of this reading frame would give a protein with a very unusual amino acid composition (Table 6). It would contain 26% alanine and be very hydrophobic. For this reason, it seems possible that it may be the gene (*lom*) for the outer membrane protein identified by Reeve & Shaw (1979) (see below). It is interesting that if *orf401* makes a protein it also would be very rich in alanine. This is probably due mainly to the fact that the alanine codon, GCA, will inevitably be found on both strands in a different reading phase.

orf314

This long reading frame has no really satisfactory start site. The nearest possible one to the position indicated by ANALYSEQ is at 21,029. It contains 12 nucleotides between the SD sequence (A-G) and the ATG. It may be noted that there is an ATG in a different phase at position 20,815 which contains a good SD site (A-A-G-G-G-G-T). This could be the initiation site if there were an error (insertion) in the sequence between positions 20,818 and 20,964. However, the sequence has been determined on both strands and there is no evidence of any ambiguity. One possible explanation could be that there is no protein encoded by this reading frame in the strain of lambda used. The *b* region appears to have no essential function in phage grown in the laboratory, so that a fortuitous frame-shift mutation could have taken place and been preserved in the strain without having any effect on its growth or other properties.

```

CGTTGCAGCAGCGTTTCAGACGTGGTACTGCTGAGCTGGCACTATCCCCTTCGCTTGCTGCCATCTCAAGCGCAC
 20780 20778 20760 20750 20748 20738 20720 20710 20700
      A E A I S S A R F A E F F A R I A A A S A A L L L C D A P T
 20690 20680 20670 20660 20650 20640 20630 20620 20610
      A L P A A S V A F U D A V D A L P A A V F A S A A E E A L
 20600 20590 20580 20570 20560 20550 20540 20530 20520
      R S A A V S D D L A F U S D P A A L F S A A U A E E E A
 20510 20500 20490 20480 20470 20460 20450 20440 20430
      CCGTCCGCTGCTGCTTCAAGATGACCTGGATTCTGCTCTGGCGTITGGCGGACTCCCGCCCGCTGTTTGCGTCTGCCGCGAGAGGCC
 20420 20410 20400 20390 20380 20370 20360 20350 20340
      A R P A L D D A F U S D D F A A S F E A T S R A E U A A S
 20420 20410 20400 20390 20380 20370 20360 20350 20340
      D G G A F U A A U E A D U A A D P C C D A A F U S D U F A A P A
 20390 20380 20370 20360 20350 20340 20330 20320 20310
      L U A A A F F E D S A A A A L F F S A S V A F A D A A S A P E
 20240 20230 20220 20210 20200 20190 20180 20170 20160
      D A S *
 20150 20140 20130
      GGACGCTTCTGAGCTGACGATGCAAG

```

FIG. 5. Nucleotide and amino acid sequence of *orf206b*.

The *nin* region

This region, which is not essential for the growth of the phage, has an unusual structure. It contains eight or nine possible reading frames, all of which start with an ATG or GTG that overlap the end of the previous reading frame. Their structures are shown in Table 4. It will be seen that there is also a very short reading frame (*orf26*) that overlaps the ends of the *ren* and *orf146* gene with similar structures.

Altogether, there is a continuous series of 13 overlapping reading frames from the beginning of the *O* gene to the end of the *Q* gene. The *O*, *P* and *Q* genes make proteins, but it is not known whether the others do. Figure 3(c) shows the application of the ANALYSEQ program to the right transcription region using genes *O*, *P*, *Q* and *R* as standards. Although there are peaks corresponding to the *nin* reading frames, they are not very clear. Figure 3(d) shows the results for the *nin* region using *orf146*, *orf290* and *orf204* as standards. Here, all the other reading frames show up as definite peaks, indicating that they all have a similar codon distribution.

orf64

This is a reading frame proposed by Daniels & Blattner (1982). It does not give a peak in the ANALYSEQ program.

S

There are two ATG triplets close together. The first is chosen as having the better SD sequence. The reading frame has a TAG at position 45,351. This is the site of the *S7* mutation, which was present in the strain used.

The right-hand end (46,428 to 48,502)

Other than the *cos* site, no function has yet been ascribed to this region and there do not appear to be any possible rightward reading frames in it. However, there are two possible reading frames on the opposite strand with acceptable SD sequences (see Fig. 6). In the ANALYSEQ program, using standards from the left arm or from the central region, they do show as peaks, but the values are not very high. As far as we know, there is no other evidence for leftward transcription or for gene products in this region.

rexB

There are two ATG triplets with good SD sequences at 36,259 and 36,244.

orf28

In the ANALYSEQ program there is a high peak (position 34,357 to 34,271) just preceding the reading frame assigned to the *ral* gene (Ineichen *et al.*, 1981). It corresponds to a reading frame having a satisfactory possible initiating site. It would code for a protein of only 28 amino acids, of which eight would be glutamic acid, four aspartic acid and five tyrosine. Such a bizarre protein could have interesting properties.

M A Q U A I F K E I F D Q U R
 CGATAATGCHAACATACGGCCCTCGTATCACATGGAGGGTTTACCAATGGCTCAGGGTGCATTAAAGAAATATTGATCAAGTC
 47498 47480 47470 47468 47458 47448 47438 47420 47418
 K P L D C E L F Y S E L K R H N U S H Y I Y L A T D N I H
 GAAAGAGTTTAAAGCTGTGATTGTTTACTGAACTAAACGTCACACCGTCACATTATATTACTATCTAGCCACAGATAATATTC
 47480 47398 47380 47370 47360 47350 47340 47330 47320
 I U L E N D N T U L I K G L K K U U N U K F S R N T H L I E
 ACATCGTGTAGGAAACGATAAACACCGTGTAAATAAAAGGACTTAAATGTTAATCTCAAGAAACAGCATCTTATAG
 47310 47300 47290 47280 47270 47260 47250 47240 47230
 T S Y D R L K S R E I T F Q O Y R E H N L A K A G U F R H W T
 AAACGTCATGTAGGGTGAATCAAGAGAAATCACATTCAGCAATCAGGGAAATCTTGCTAACAGGAGTTTCCGATGGGTTA
 47220 47210 47200 47190 47180 47170 47160 47150 47140
 N I H E H K R Y Y T F D N S L F T Y S I Q N T T Q I F P
 CAAATATCCATGAACTAAAGHTATTACTATCCCTGATTAATCTTACTATTACTGAGAGACATTGAGACACTACAAATCTTC
 47130 47120 47110 47100 47090 47080 47070 47060 47050
 R *
 CACGCTTAAATCATACGGTCCGGTTCTTCGGTGTACGCCGGGGCTGGCATATAATGCAATACGGTGTACCGCGCTAAACCGTGTGTCAT
 47040 47030 47020 47010 47000 46990 46980 46970 46960
 CGTTTTATYATTCCCGGACACTCCCGCAGAGAAGTCCCCGTCAGGGCTGTGGACATAGTTAATCGGGAAATACATGACGATTCTCG
 46950 46940 46930 46920 46910 46900 46890 46880 46870
 CACCTGACATACATTAATATTAACAAATATGAAATTCAACTCATTTAGGGTTTACACATACGATTCTGGGA
 46860 46850 46840 46830 46820 46810 46800 46790 46780
 M K K M L L A T A L A L L I T G C A Q Q T
 ACTTCACAAAGCATTGGAAATACACATGCTCGTACTGGCTGGCCCTGCCTTACAGGATGTGCTCAACAGCG
 46770 46760 46750 46740 46730 46720 46710 46700 46690
 F T U O N K P S P U A P K E T I T H H F F U S G T G Q K K T
 TTACTGTCACAAACACCGGACGAGTAGCAGCAAGGAAACCTCACCCATCATTCTCGTTTGGAATTGGCAGAAGAAACT
 46680 46670 46660 46650 46640 46630 46620 46610 46600
 U D A A K I C G G A E N V U K T E T Q O O T F U N G L G F I
 GTGCGATCGGCAAAATITGTGGCGCGCAAGAAATGTGTTAAACAGAACCCAGAAACATTGCAARTGGATTCGTCGGTTTAT
 46590 46580 46570 46560 46550 46540 46530 46520 46510
 T L G I Y T P L E A R U Y C S Q *
 ACTTAAAGCTTATCTCCGCTGGAGCGCGTGTGATTTGCTCACATAATTGCTATG

FIG. 6. Nucleotide sequence of the *r* strand (i.e. the opposite strand to that shown in Fig. 2(b)) in the region 46,452 to 47,500 and the amino acid sequence of 2 open reading frames.

cIII, kil and γ

There are various interpretations possible in this region, none of which is entirely satisfactory. The DNA sequence has been determined by Ineichen *et al.* (1981) and by us independently, so it is very unlikely that there are any mistakes in it. According to Ineichen *et al.*, the *cIII* protein starts at 33,456, γ at 33,232 and *kil* at 33,112 in the same reading frame as γ . We present here an alternative interpretation that was suggested by the ANALYSEQ results (see Fig. 3 and Table 3) and which seems to fit the results better, although giving rather small proteins.

The ATG start proposed by Ineichen *et al.* for the *cIII* gene at 33,456 is preceded by a good SD sequence at 33,475 to 33,470, 14 nucleotides upstream. If this is indeed a ribosome binding site, it would seem much more likely that it would initiate at the ATG at 33,463, which is seven nucleotides away. This would give a protein of molecular weight 6040, which is probably the *cIII*. There are no data concerning its molecular weight. There is an alternative start upstream in the same phase at 33,535, a GTG with a good SD sequence that would give a protein of molecular weight 9006. However, it would not fit so well with the ANALYSEQ result. This would leave the reading frame from 33,477 to 33,189 free and we suggest that it codes for the *kil* gene, probably starting at the ATC at 33,330, and giving a protein of 47 amino acids. The initiation site at 33,232 proposed by Ineichen *et al.* for the γ gene has a very poor SD site, and the ANALYSEQ program would suggest that the site at 33,112, which has a better SD sequence, may be the actual start although it would give rise to a smaller protein than predicted (Karn *et al.*, 1974). Clearly, further work is required to identify the reading frames for these genes.

exo

There are a number of possible start sites farther downstream but none has as good an SD as that at 32,028. The A-T-G-A overlap with gene β also makes this probable.

Region 31.347 to 30.395

According to Epp (1978), there should be a gene (*Ea9*) coding for a protein of 9000 M_r in this region. There does not appear to be a reading frame of this size, though there are three possible smaller frames starting at 31.351, 31.196 and 31.024 that could code for proteins of 60, 63 and 61 amino acids, respectively. They all have overlapping termination and initiation sites, the first one overlapping with *exo*, and could form a continuous series like that found in the *nin* region. They give weak peaks in the ANALYSEQ program. We refer to these reading frames as *orf60a*, *orf63* and *orf61*. Two functions have been mapped in this region by Court *et al.* (1980a,b) and are probably encoded by these reading frames. Most of this region has also been sequenced by Luk & Szybalski (1982). The results agree with ours, except at position 31.139–31.140, where they have only one T residue.

xis

The ANALYSEQ results are not very satisfactory in the region of the *xis* gene. This may be related to the unusual amino acid composition, particularly the high content of basic residues (Hoess *et al.*, 1980).

Region 27.810 to 26.972

This region, which contains the *att* site, probably does not code for any proteins. There is a short reading frame starting at 27,604, but it has an unlikely initiation codon.

The b region

The *b* region is not essential for lytic growth and has not, therefore, been studied in detail by genetic methods. Several proteins, identified by their size, have been shown to be encoded by this region (Hendrix, 1971; Murielado & Siminovitch, 1972; Epp, 1978). In the DNA sequence there are eight possible reading frames: four on the left arm and four in the central region. *orf401* and *orf206b* overlap on opposite strands (Fig. 1).

There are three clear reading frames in the central region, which are clearly identified as *Ea59*, *Ea31* and *Ea47* (Fig. 1, Hendrix, 1971; Murielado & Siminovitch, 1972; Epp, 1978). Epp *et al.* (1981) have identified a product (*Ea24*, M_r 24,000) that maps in the same region as *Ea59*. There is, however, no other reading frame of this size in this position. It may be initiated within the same reading frame and use the same terminator as *Ea59*. There is a possible, but unlikely initiating ATG at position 26,082.

Reeve & Shaw (1979) identified a lambda encoded protein of 21,000 M_r in the outer membrane of the infected cell and mapped the gene (*lom*) encoding it near the end of the *J* gene. There are two possible reading frames, *orf206a* and *orf206b*. The amino acid composition expected from frame *orf206b* (Table 6) shows it to be very

TABLE 6

Amino acid composition (number of residues) of proteins that would be coded by reading frames orf206a and orf206b

Amino acid	orf206a	orf206b
Phe	7	19
Leu	8	16
Ile	8	3
Met	8	0
Val	23	21
Ser	21	24
Pro	6	6
Thr	18	4
Ala	22	74
Tyr	10	0
His	5	0
Gln	4	0
Asn	3	0
Lys	9	0
Glu	8	19
Asp	8	11
Cys	1	4
Trp	3	0
Arg	10	5
Gly	24	0
Polarity index†	42	30

† Capaldi & Vanderkooi (1972).

hydrophobic, as would be expected for a membrane protein. *orf206a* is less hydrophobic but contains local regions of high hydrophobicity, which might anchor it to the membrane. Reeve & Shaw believed *lom* to be a late protein, which would favour *orf206a*, though *orf206b* could possibly be translated on a different mRNA from the other early proteins. There is thus no clear evidence as to which reading frame is coding for this protein. If *orf206b* is *lom*, *orf206a* may be *La21* (Hendrix, 1971). Hendrix identified late products (*La43*, *La40* and *La38*) that he suggested were related to one another by degradation. These probably correspond to *orf401* though *La38* could perhaps be related to *orf314* (see above). Muralaldo & Siminovitch identified late products *p16* and *p17* (23,000 and 20,000 M_r , respectively), which could correspond to *orf206a* and *orf194*. They also found evidence for a late protein product, *p8*, of 79,000 M_r in the *b* region; however, there is no reading frame of this size.

Other reading frames

There are a number of fairly long open frames reading from right to left on the left arm. They are found in the same position as the coding frames on the opposite strand and are probably related to the fact that the codons UCA, CUA and UUA, which are complementary to the termination codons UGA, UAG and UAA, respectively, are absent or rare in the coding regions (Table 5). Whether or not this effect has a more general biological significance or whether it is a coincidence

applying only to this region of sequence is not clear. The reading frames on the strand opposite to genes *E*, *L* and *J* show slight peaks with the ANALYSEQ program, but do not have satisfactory initiation sites. ATG and GTG are in fact rare codons in these frames, since the complementary codons CAC and CAT, which code for histidine, are rare on the coding strand. The only possible initiation site in these reading frames is an ATG at position 17,853, which could code for a protein of 472 amino acids.

Some of the sequencing was carried out by C. Howe, V. Devalia and C. J. Edge while they were summer students in this laboratory, to whom we are very grateful. We wish to thank J. Messing for gifts of the M13 vectors, F. R. Blattner and D. L. Daniels for supplying us with up-to-date lambda data and for helpful advice, K. McKenney for advice on interpretation of the data and other members of this laboratory, especially B. G. Barrell, for help and advice. The computer analyses were very much dependent on the help of R. Staden, to whom we are very grateful. D.F.H. thanks the Medical Research Council of New Zealand for post-doctoral support and G.B.P. thanks the Royal Society and Nuffield Foundation for a Commonwealth Bursary.

REFERENCES

- Anderson, S. (1981). *Nucl. Acids Res.* **9**, 3015-3027.
 Anderson, S., Gait, M. J., Mayol, I. & Young, I. G. (1980). *Nucl. Acids Res.* **8**, 1731-1743.
 Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H. L., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J. H., Staden, R. & Young, I. G. (1981). *Nature (London)*, **290**, 457-465.
 Capaldi, R. A. & Vanderkooi, G. (1972). *Proc. Nat. Acad. Sci., U.S.A.* **69**, 930-932.
 Court, D., Gottesman, M. & Gallo, M. (1980a). *J. Mol. Biol.* **138**, 715-729.
 Court, D., de Crombrugghe, B., Adhya, S. & Gottesman, M. (1980b). *J. Mol. Biol.* **138**, 731-743.
 Daniels, D. L. & Blattner, F. R. (1982). *Virology*, **117**, 81-92.
 Daniels, D. L., de Wet, J. R. & Blattner, F. R. (1980). *J. Virol.* **33**, 390-400.
 Daniels, D. L., Sanger, F. & Coulson, A. R. (1982). *Cold Spring Harbor Laboratory*, in the press.
 Duckworth, M. L., Gait, M. J., Goelet, P., Hong, G. F., Singh, M. & Titmus, R. C. (1981). *Nucl. Acids Res.* **9**, 1691-1706.
 Echols, H. & Muriel, H. (1978). *Microbiol. Rev.* **42**, 577-591.
 Epp, C. (1978). Ph.D. thesis, University of Toronto.
 Epp, C., Pearson, M. L. & Enquist, L. (1981). *Gene*, **13**, 327-337.
 Franklin, N. C. & Bennett, G. N. (1979). *Gene*, **8**, 107-119.
 Fuhrman, S. A., Deininger, P. L., La Porte, P., Friedmann, T. & Geiduschek, E. P. (1981). *Nucl. Acids Res.* **9**, 6439-6456.
 Garoff, H. & Ansorge, W. (1981). *Anal. Biochem.* **115**, 450-457.
 Gronenborn, B. & Messing, J. (1978). *Nature (London)*, **272**, 375-377.
 Hendrix, R. W. (1971). In *The Bacteriophage Lambda* (Hershey, A. D., ed.), pp. 355-370. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.
 Hendrix, R. W. & Casjens, S. R. (1974). *Virology*, **61**, 156-159.
 Hoess, R., Foeller, C., Bidwell, K. & Landy, A. (1980). *Proc. Nat. Acad. Sci., U.S.A.* **77**, 2482-2486.
 Hong, G. F. (1981). *Biosci. Rep.* **1**, 243-252.
 Ikemura, T. (1981a). *J. Mol. Biol.* **146**, 1-21.
 Ikemura, T. (1981b). *J. Mol. Biol.* **151**, 389-409.
 Imada, M. & Tsugita, A. (1971). *Nature New Biol.* **233**, 230-231.
 Ineichen, K., Shepherd, J. C. W. & Bickle, T. A. (1981). *Nucl. Acids Res.* **9**, 4639-4653.

- Iserentant, D. & Fiers, W. (1980). *Gene*, **9**, 1-12.
- Jeffreys, A. J. & Flavell, R. A. (1977). *Cell*, **12**, 429-439.
- Karn, A., Sakari, Y., Echols, H. & Linn, S. (1974). In *Mechanisms of Recombination* (Grell, R. F., ed.), pp. 95-106, Plenum Press, New York.
- Landsmann, J., Kroger, M. & Hobom, G. (1982). *Gene*, **20**, 11-24.
- Luk, K.-C. & Szybalski, W. (1982). *Gene*, **17**, 247-258.
- Maxam, A. M. & Gilbert, W. (1977). *Proc. Nat. Acad. Sci., U.S.A.*, **74**, 560-564.
- Messing, J. & Vieira, J. (1982). *Gene*, in the press.
- Messing, J., Crea, R. & Seeburg, P. H. (1981). *Nucl. Acids Res.*, **9**, 309-321.
- Mills, D. R. & Kramer, F. R. (1979). *Proc. Nat. Acad. Sci., U.S.A.*, **76**, 2232-2235.
- Murialdo, H. & Siminovitch, L. (1972). *Virology*, **48**, 785-823.
- Oppenheim, D. S. & Yanofsky, C. (1980). *Genetics*, **95**, 785-795.
- Petrov, N. A., Karginov, V. A., Mikriukov, N. N., Serpinskii, O. I. & Kravchenko, V. V. (1981). *FEBS Letters*, **133**, 316-320.
- Ptashne, M., Backman, K., Humayun, M. Z., Jeffrey, A., Maurer, R., Meyer, B. & Sauer, R. T. (1976). *Science*, **194**, 156-161.
- Ray, P. N. & Pearson, M. L. (1974). *J. Mol. Biol.*, **85**, 163-175.
- Ray, P. N. & Pearson, M. L. (1975). *Nature (London)*, **253**, 647-650.
- Reeve, J. N. & Shaw, J. E. (1979). *Mol. Gen. Genet.*, **172**, 243-248.
- Roberts, T. M., Shimatake, H., Brady, C. & Rosenberg, M. (1977). *Nature (London)*, **270**, 274-275.
- Sanger, F., Nicklen, S. & Coulson, A. R. (1977). *Proc. Nat. Acad. Sci., U.S.A.*, **74**, 5463-5467.
- Sanger, F., Coulson, A. R., Barrell, B. G., Smith, A. J. H. & Roe, B. A. (1980). *J. Mol. Biol.*, **143**, 161-178.
- Schechtman, M. G., Alegre, J. N. & Roberts, J. W. (1980). *J. Mol. Biol.*, **142**, 269-288.
- Schwarz, E., Scherer, G., Hobom, G. & Kössel, H. (1978). *Nature (London)*, **272**, 410-414.
- Schwarz, E., Scherer, G., Hobom, G. & Kössel, H. (1980). *Biochem. Int.*, **1**, 386-394.
- Shine, J. & Dalgarno, L. (1974). *Proc. Nat. Acad. Sci., U.S.A.*, **71**, 1342-1346.
- Southern, E. (1979). *Methods Enzymol.*, **68**, 152-176.
- Staden, R. (1978). *Nucl. Acids Res.*, **5**, 1013-1015.
- Staden, R. (1980). *Nucl. Acids Res.*, **8**, 3673-3694.
- Staden, R. & McLachlan, A. D. (1982). *Nucl. Acids Res.*, **10**, 141-156.
- Stormo, G. D., Schneider, T. D. & Gold, L. M. (1982). *Nucl. Acids Res.*, **10**, 2971-2996.
- Szybalski, E. H. & Szybalski, W. (1979). *Gene*, **7**, 217-270.
- Walker, J. E., Auffret, A. D., Carne, A., Gurnett, A., Hamish, P., Hill, D. & Saraste, M. (1982). *Eur. J. Biochem.*, **123**, 253-260.
- Winter, G. & Fields, S. (1980). *Nucl. Acids Res.*, **8**, 1965-1974.
- Winter, G., Fields, S. & Ratti, G. (1981). *Nucl. Acids Res.*, **9**, 6907-6915.
- Yudelevich, A. (1971). *J. Mol. Biol.*, **60**, 20-29.

Edited by S. Brenner

Note added in proof: Kröger, M. & Hobom, G. (*Gene*, **20**, 25-38 (1982)) have recently published the DNA sequence from position 40,218 to 43,972. Our results agree with their sequence except at positions 41,978 and 43,082. At position 41,978 we originally had the sequence C-G, but have now re-examined the sequencing gels and find that they were misread and that the sequence is G-C, which agrees with Kröger & Hobom. This has now been corrected in Figure 2(b). At position 43,082 they find G whereas we find A. We believe that this may be due to a difference in the strain of lambda used.