J. Mol. Biol. (1982) **162**, 729–773

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.

0022–2836/82/360729–45 $03.00/0

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 *λcI*indlts857*S*am7. 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.*, 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 2mM-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 × 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 × 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 μl (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 *Taq*I). A 2-μl 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 μl of 0·5 mM-dATP and 1 μl of DNA polymerase were added and the mixture incubated at room temperature for 15 min: 1 μl of 0·2 M-EDTA, 50 μl of water and 50 μl of water-saturated phenol were added. After separation of the aqueous layer, the DNA was precipitated with ethanol and the product dissolved in 5 μl 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 μl of water; 2 μl 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, *Hin*fI, *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 1

Sequences not confirmed in this work

28,462–28,807	Hoess *et al.* (1980)
35,246–35,279	Franklin & Bennett (1979)
37,710–38,040	Schwarz *et al.* (1978)
38,198–38,498	Schwarz *et al.* (1978); Roberts *et al.* (1977)
38,681–38,756	Schwarz *et al.* (1978)
39,658–39,699	Schwarz *et al.* (1980)
44,170–44,500	Daniels & Blattner (1982); Petrov *et al.* (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

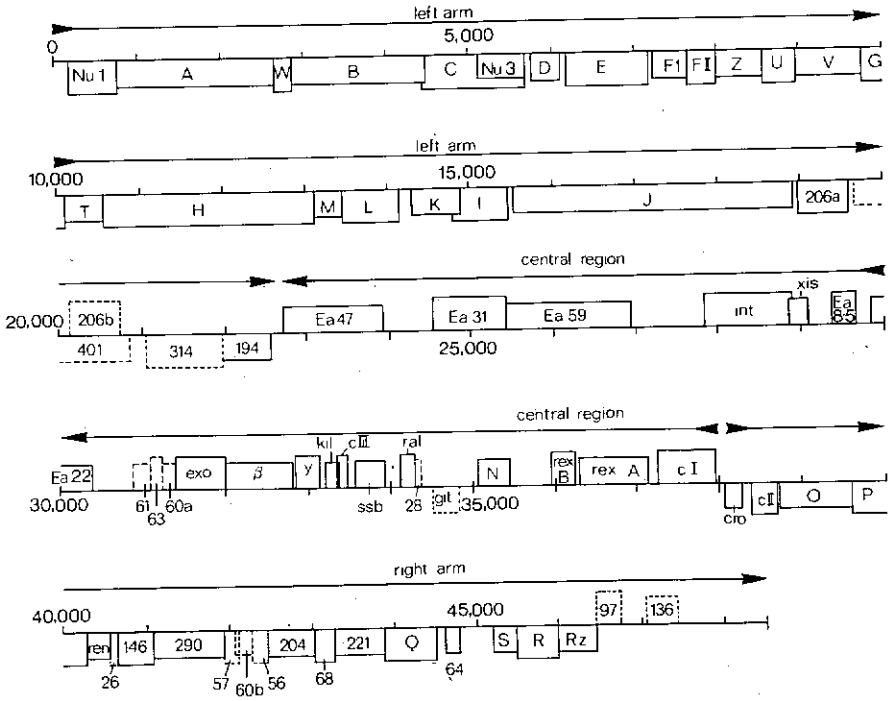


FIG. 1. Gene map of lambda showing the most probable reading frames for the proteins. Numbering is from the left-hand end. Large arrows above the genes show the main directions of transcription and translation. Genes in boxes below the line are translated from left to right; those above from right to left. Boxes with broken lines indicate the more speculative reading frames. *git* 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

Left arm			Central region			Right arm		
Gene	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)
NuI	21	20,444	*cI*	27	26,245	*cro*	7–9	7365
A	80	73,280	*rexA*	29	31,259	*cII*	11–14	11,057
W	5–10	7614	*rexB*	—	15,970	*O*	34–37	33,870
B	60	59,474	*N*	13·5	15,376	*P*	23–27	26,522
C	61	45,915	*ral*	—	7605	*ren*	—	10,613
Nu3	19	20,789	*ssb*	10–16	13,782	*orf146*	—	16,650
D	11–12	11,574	*cIII*	—	6048	*orf290*	—	33,566
E	38	38,194			or 10,067§	*orf57*	—	6964
FI	17	14,310	*kil*	16	5519	*orf60b*	—	7384
FII	11·5	12,731			or 11,646§	*orf56*	—	6321
Z	20	21,564	*γ*	16·5	11,646	*orf204*	—	24,119
U	13–16	14,651			or 16,349§	*orf68*	—	7882
V	25–31	25,814	*β*	28–30	29,693	*orf221*	—	25,222
G	33	15,610	*exo*	25–26	25,912	*Q*	23	22,476
T	16	16,061	*Ea22*	19–22	20,939	*orf64*	—	7084
H	90	92,294	*Ea8·5*	8·5	10,751	*S*	—	11,522
M	10	12,532	*xis*	—	8606	*R*	18	17,828
L	29	25,713	*int*	36–44	40,310	R_Z	11–13	17,232
K	27	23,014	*Ea59*	—	59,494			
I	—	23,127	*Ea31*	—	34,588			
J	130–140	124,440	*Ea47*	—	48,103			
orf206a	—	21,859	*orf206b*	—	20,186			
orf401	—	39,889						
orf314	—	32,024						
orf194	—	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.

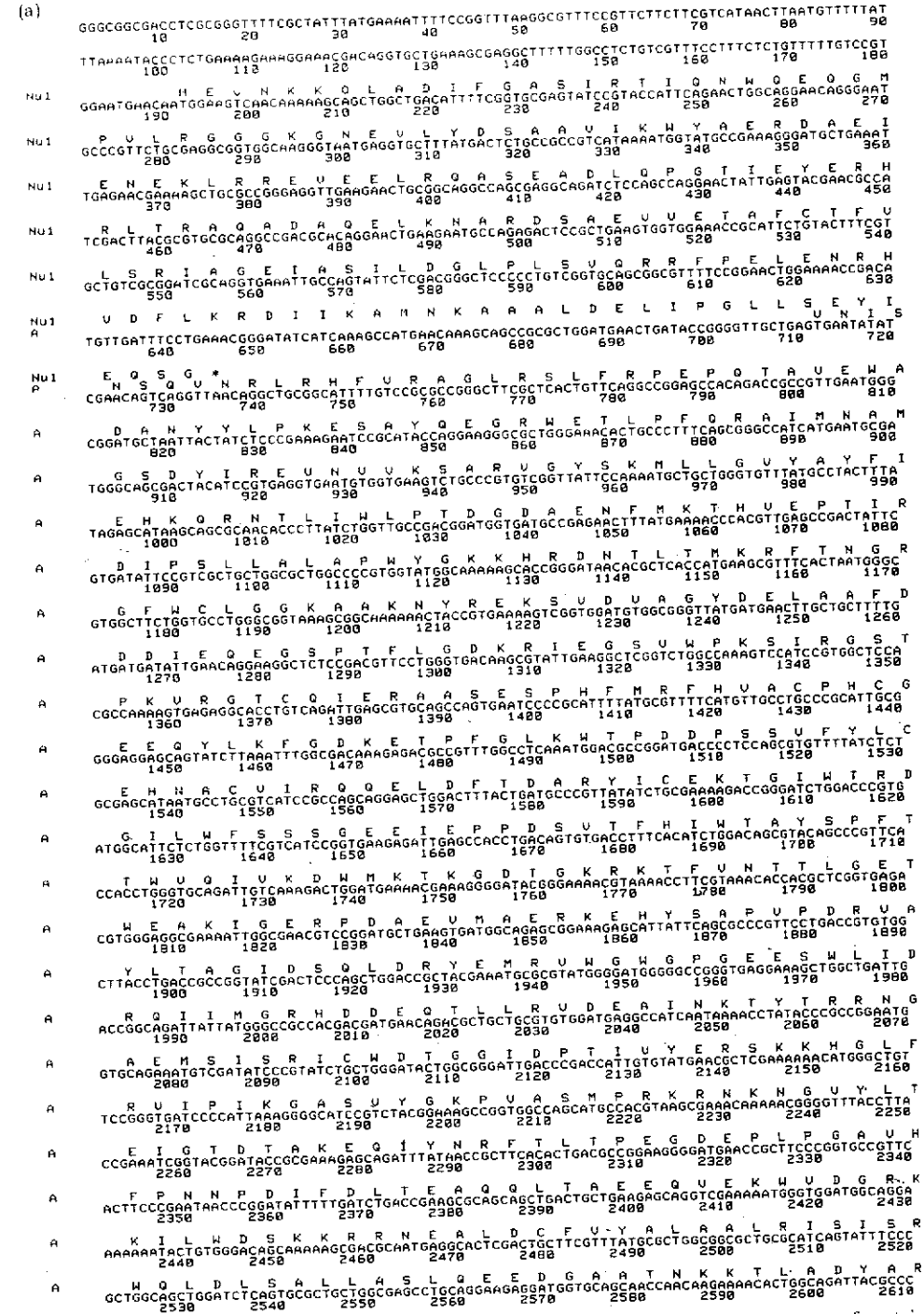
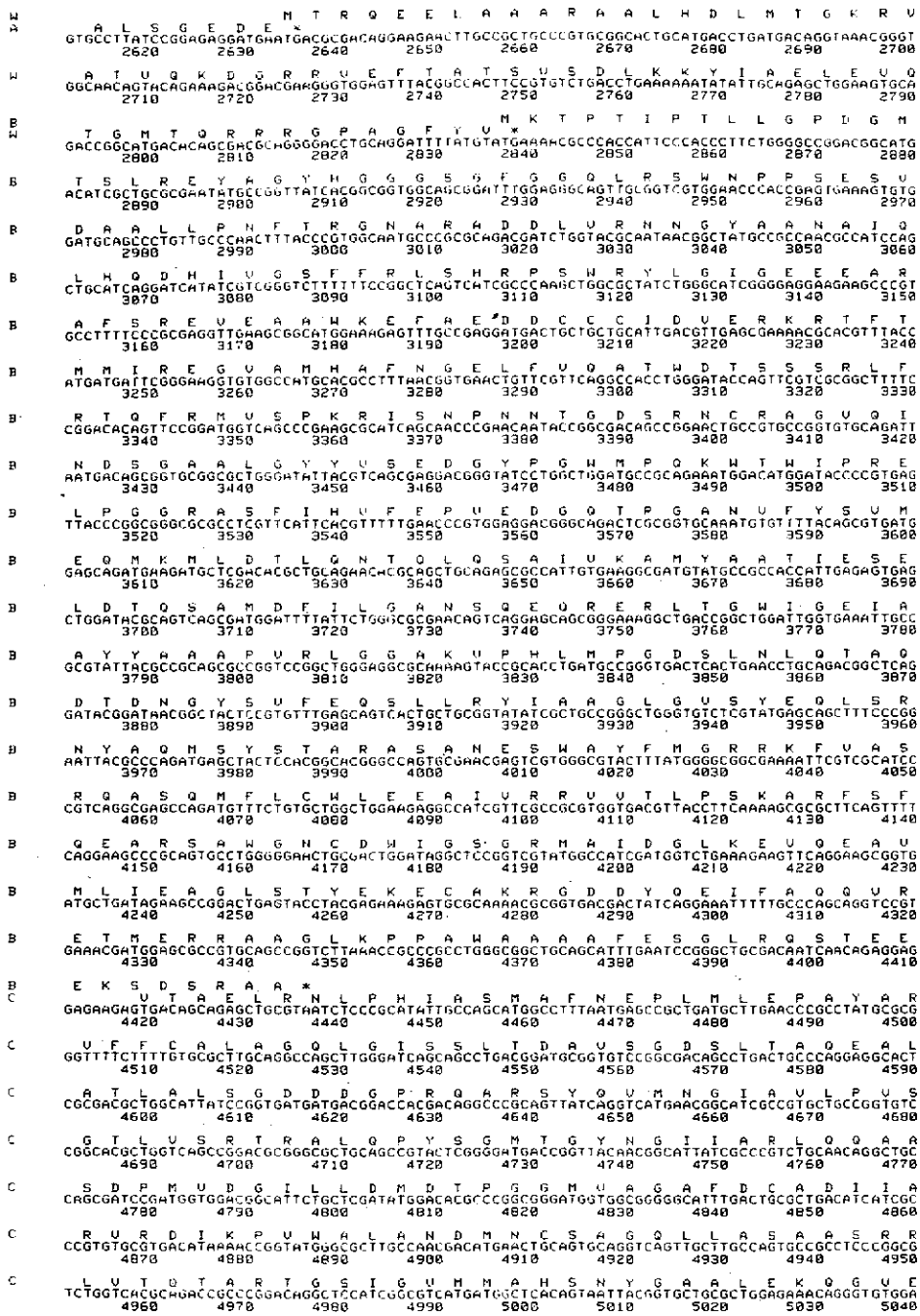
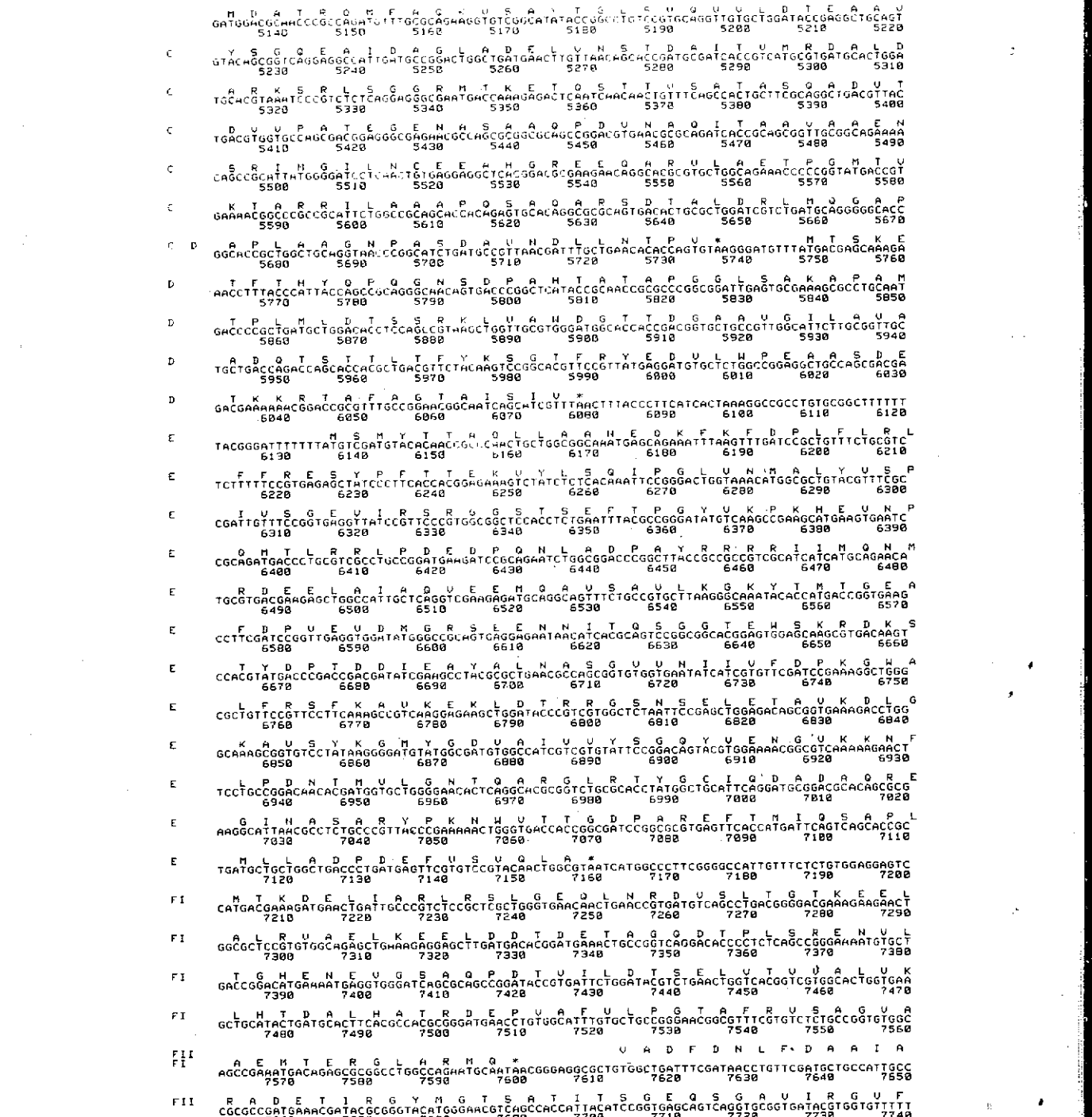
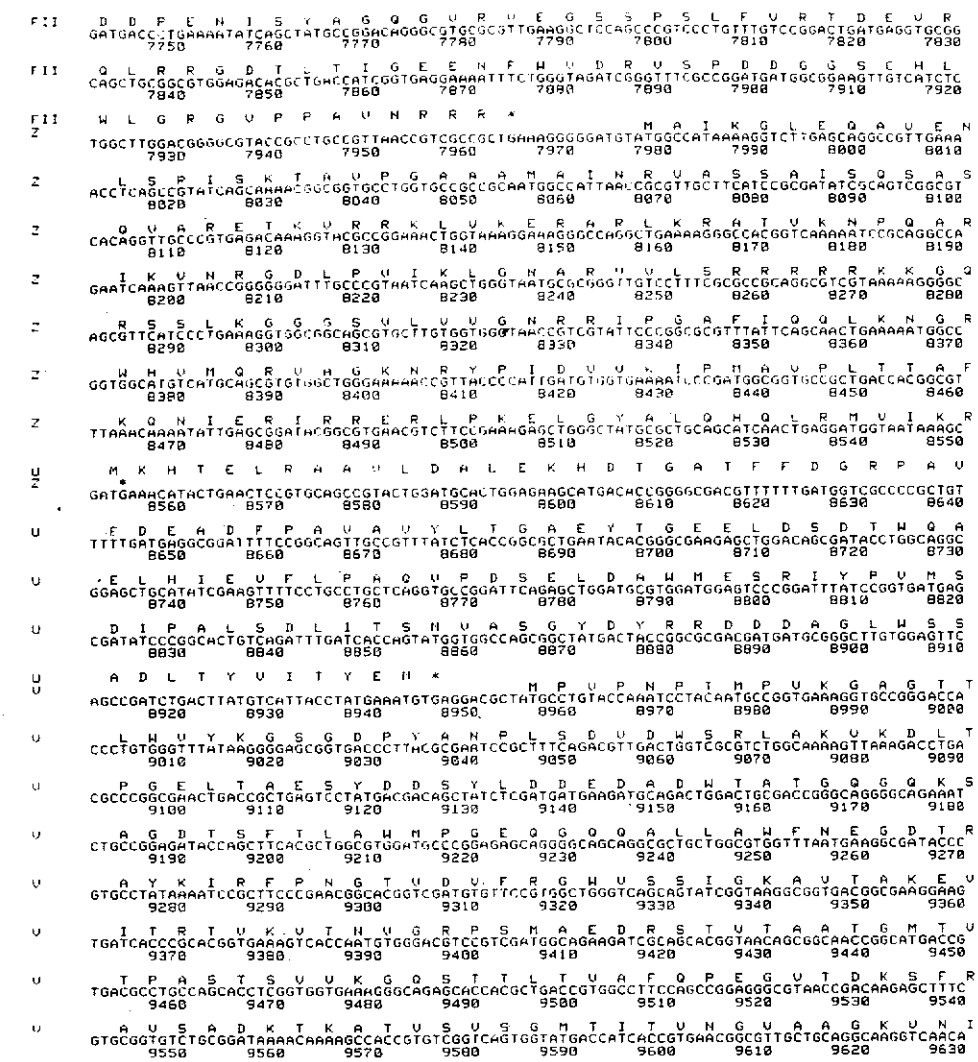


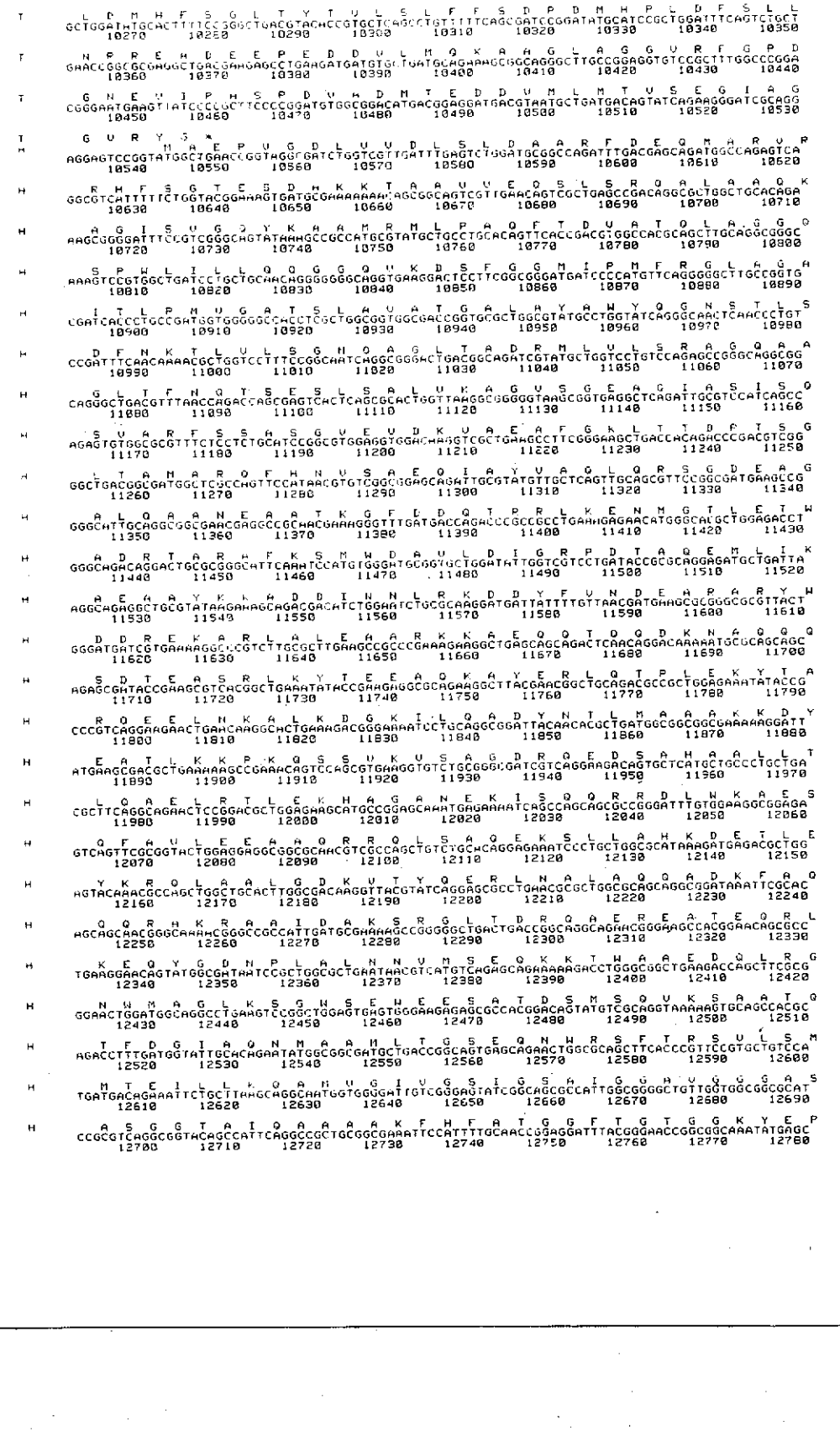
FIG. 2. Nucleotide sequence of lambda DNA and the most probable amino acid sequences of proteins it codes for. The sequence given is for *λcI*ndltsS57*S*am7; mutations from the wild type are at 37,589 (*cI*ndI), 37.742 (*cI*tsS57) and 45,351 to 45.353 (*S*am7). The sequence is in 3 sections according 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

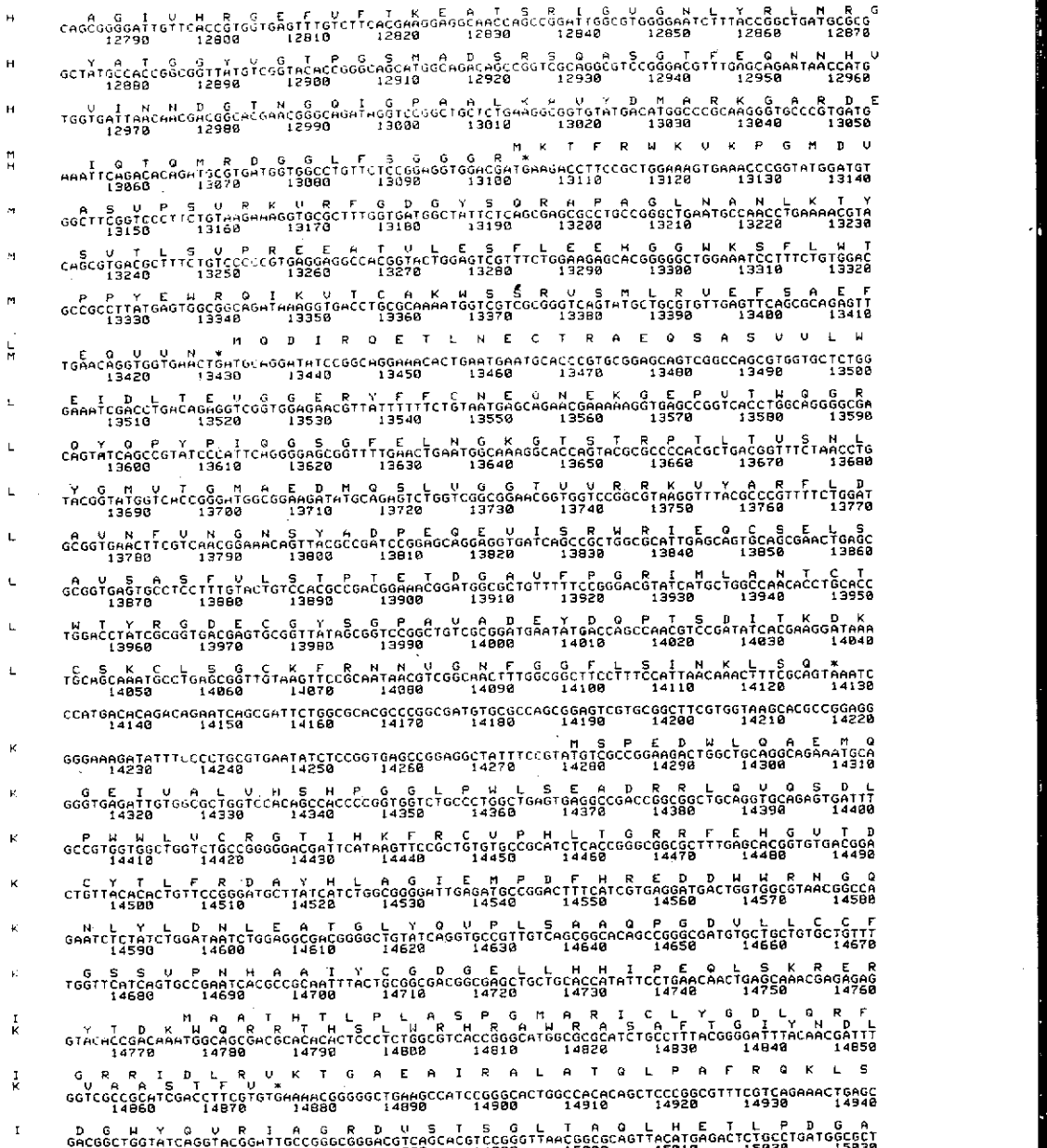


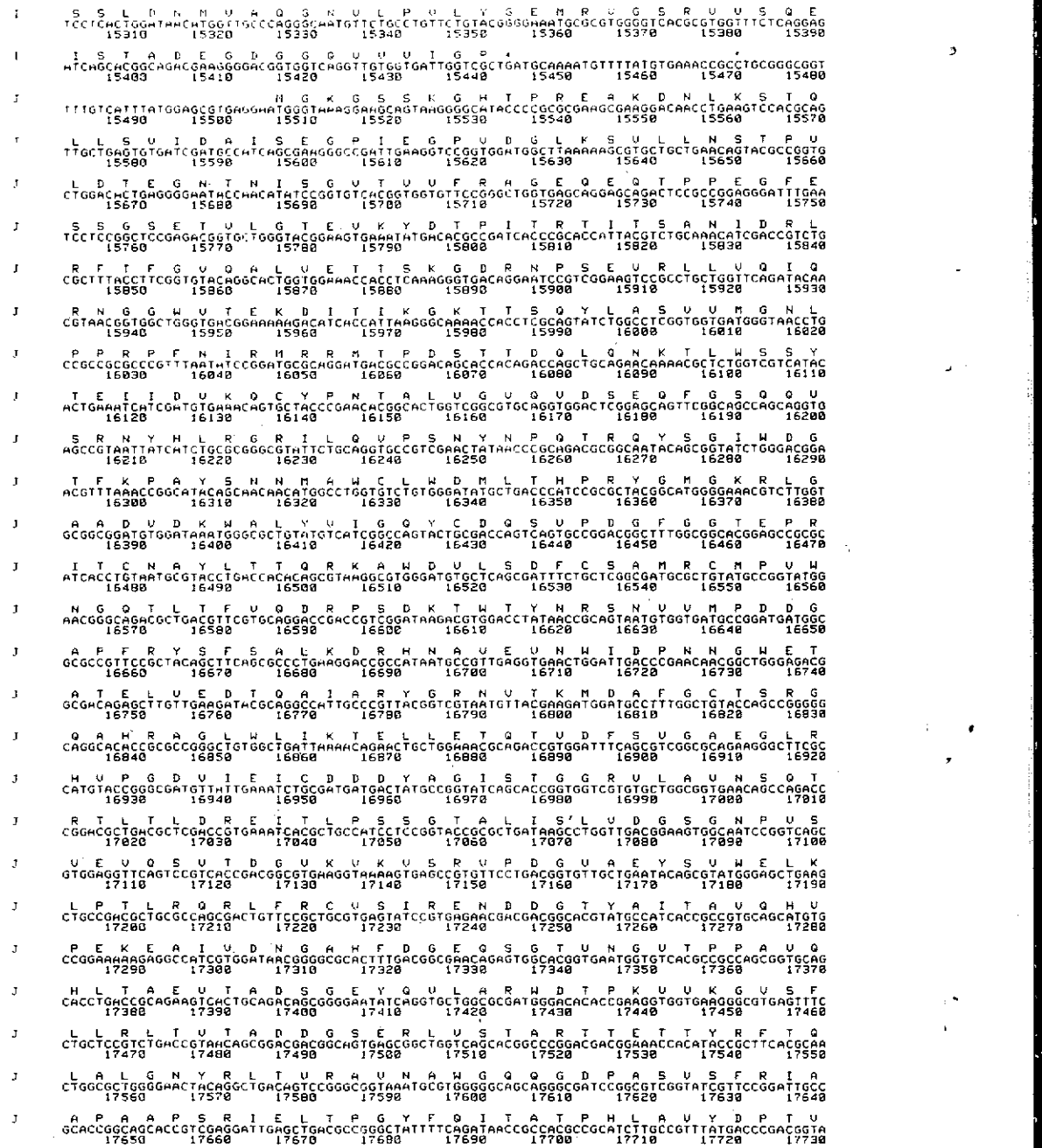
products or reading frames (*orf*) are identified in the left-hand margin. Some of the more important control regions are also marked above the sequences. The start points of the main transcription products are indicated as follows: s_R, start site for p_R RNA; s_E, for p_E RNA; $s_{R'}$, for $p_{R'}$ 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).

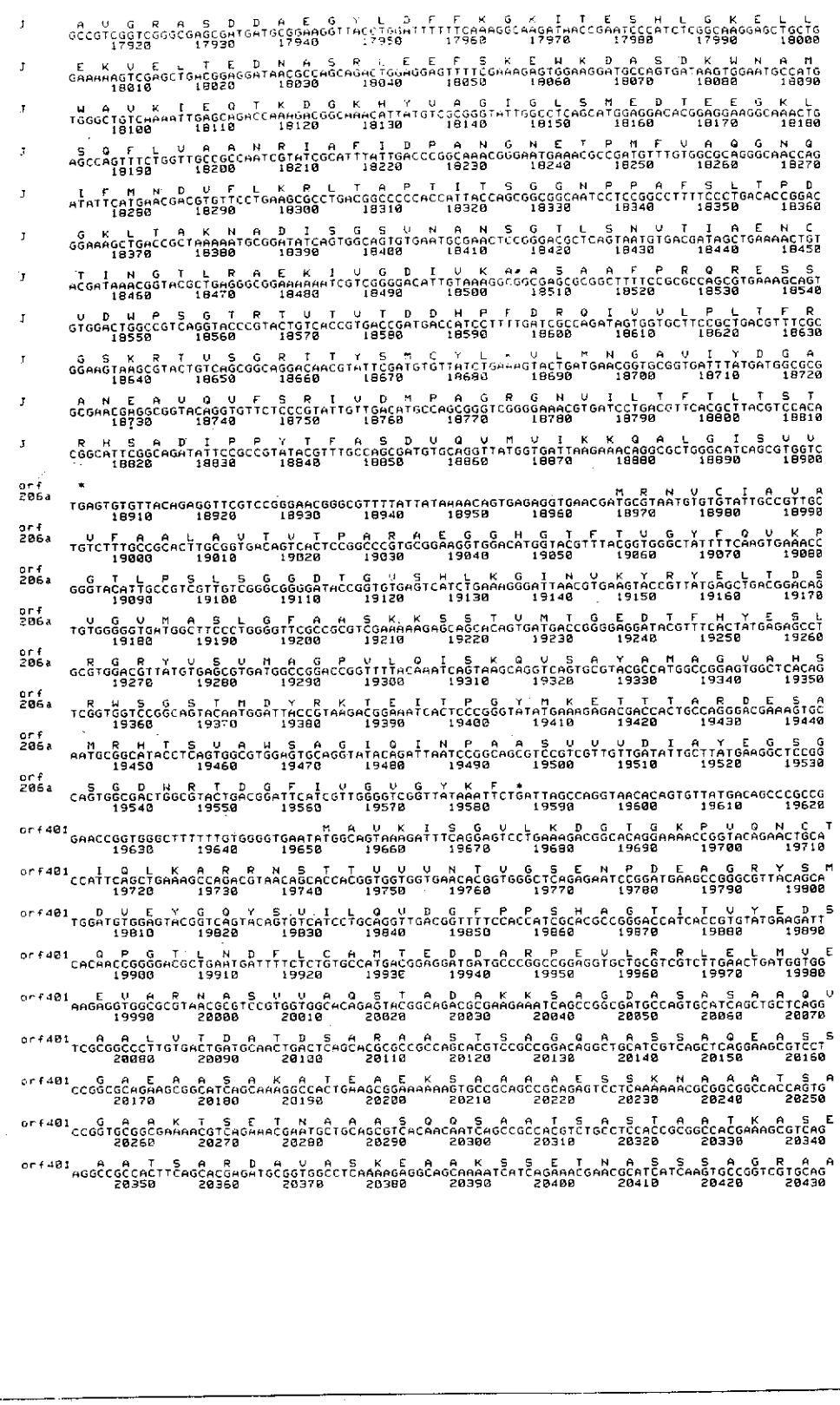


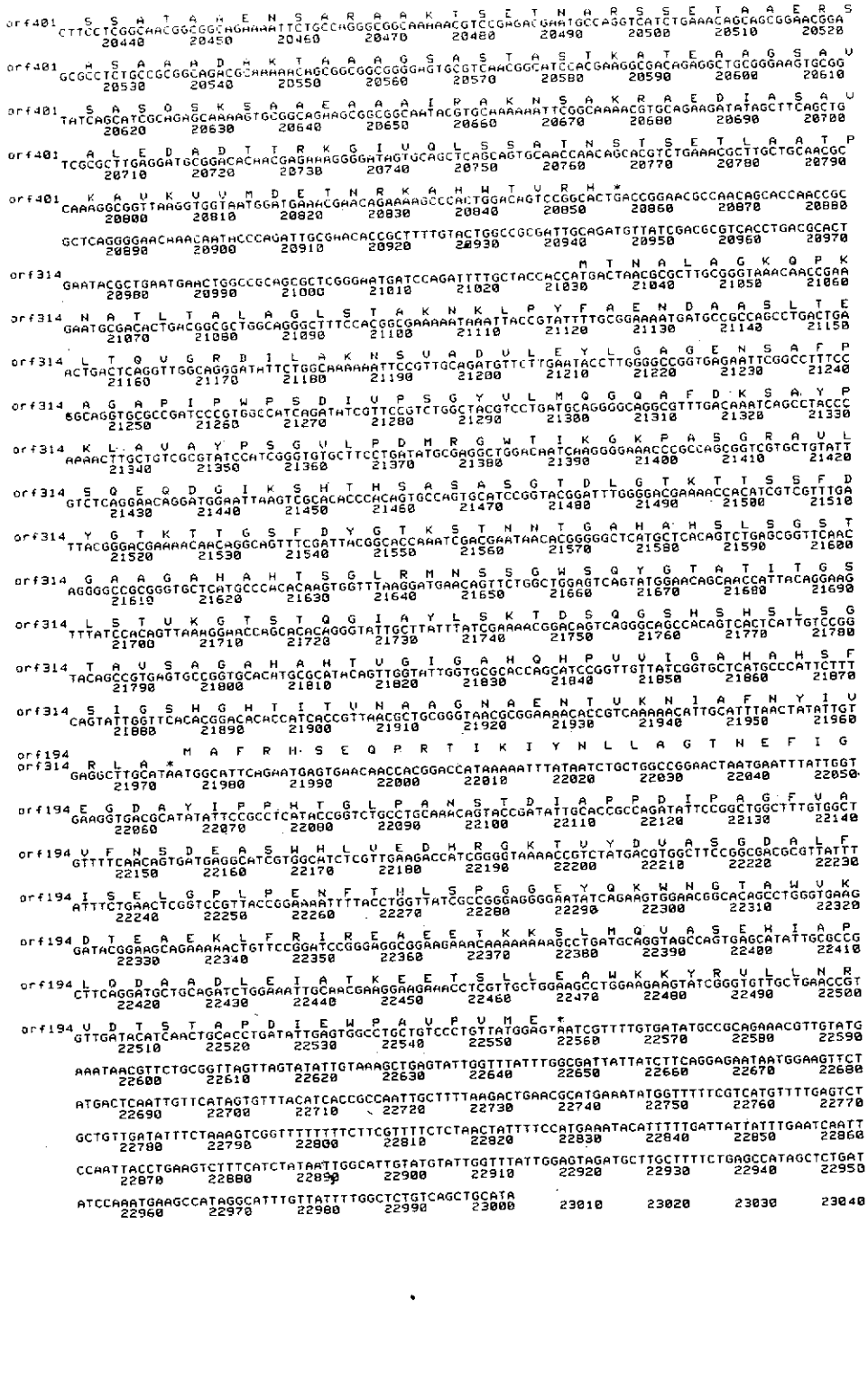












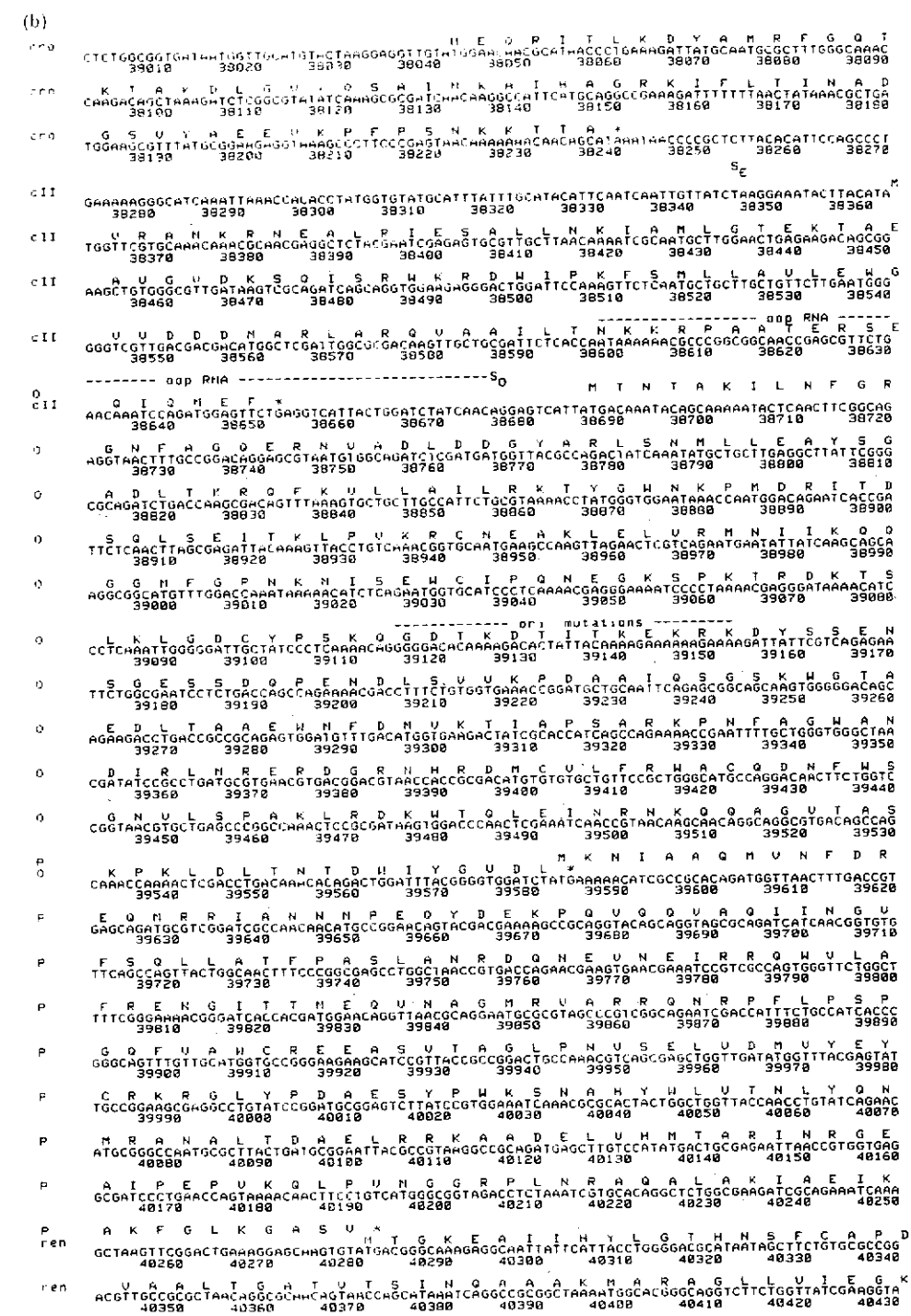
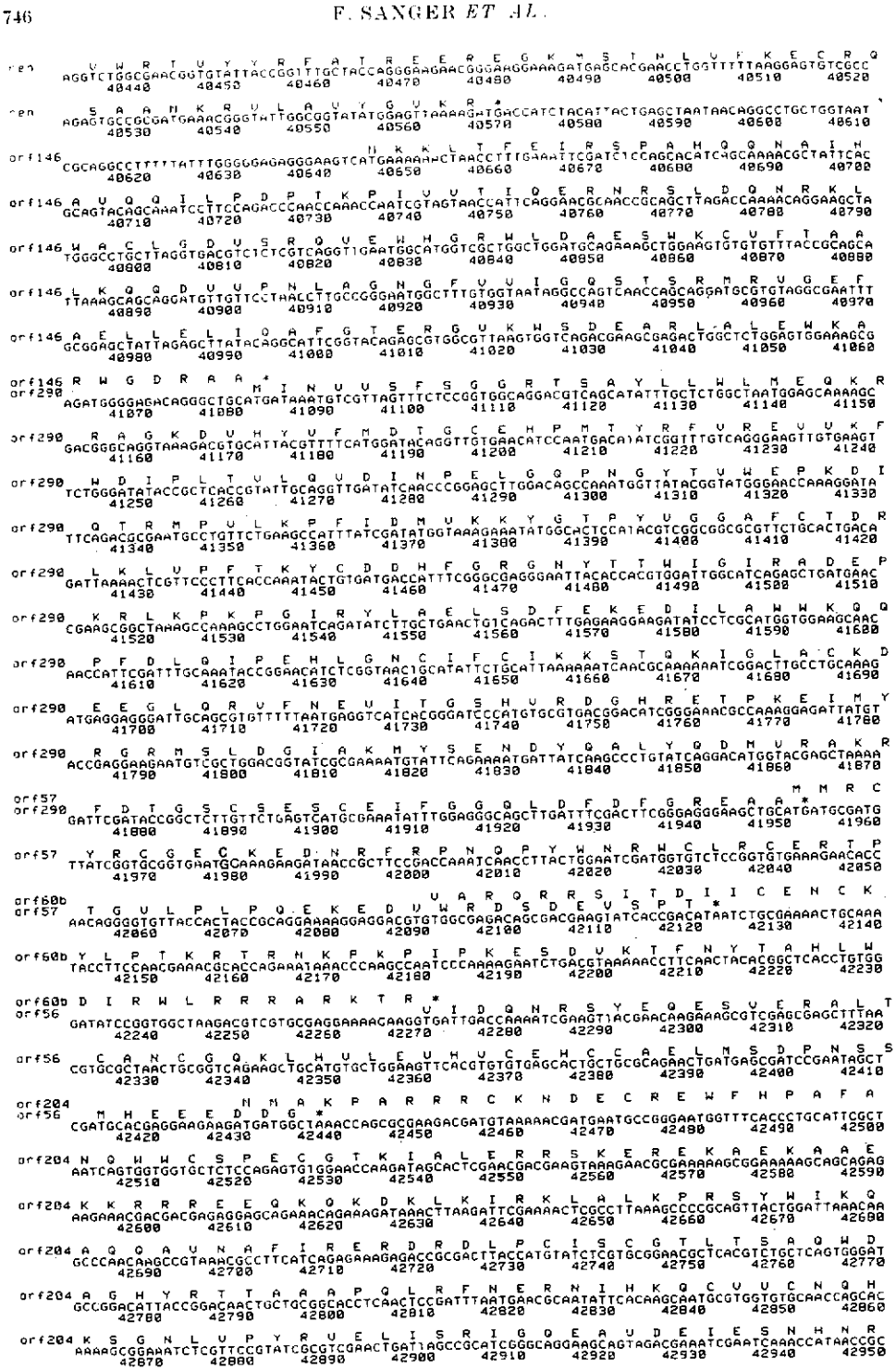
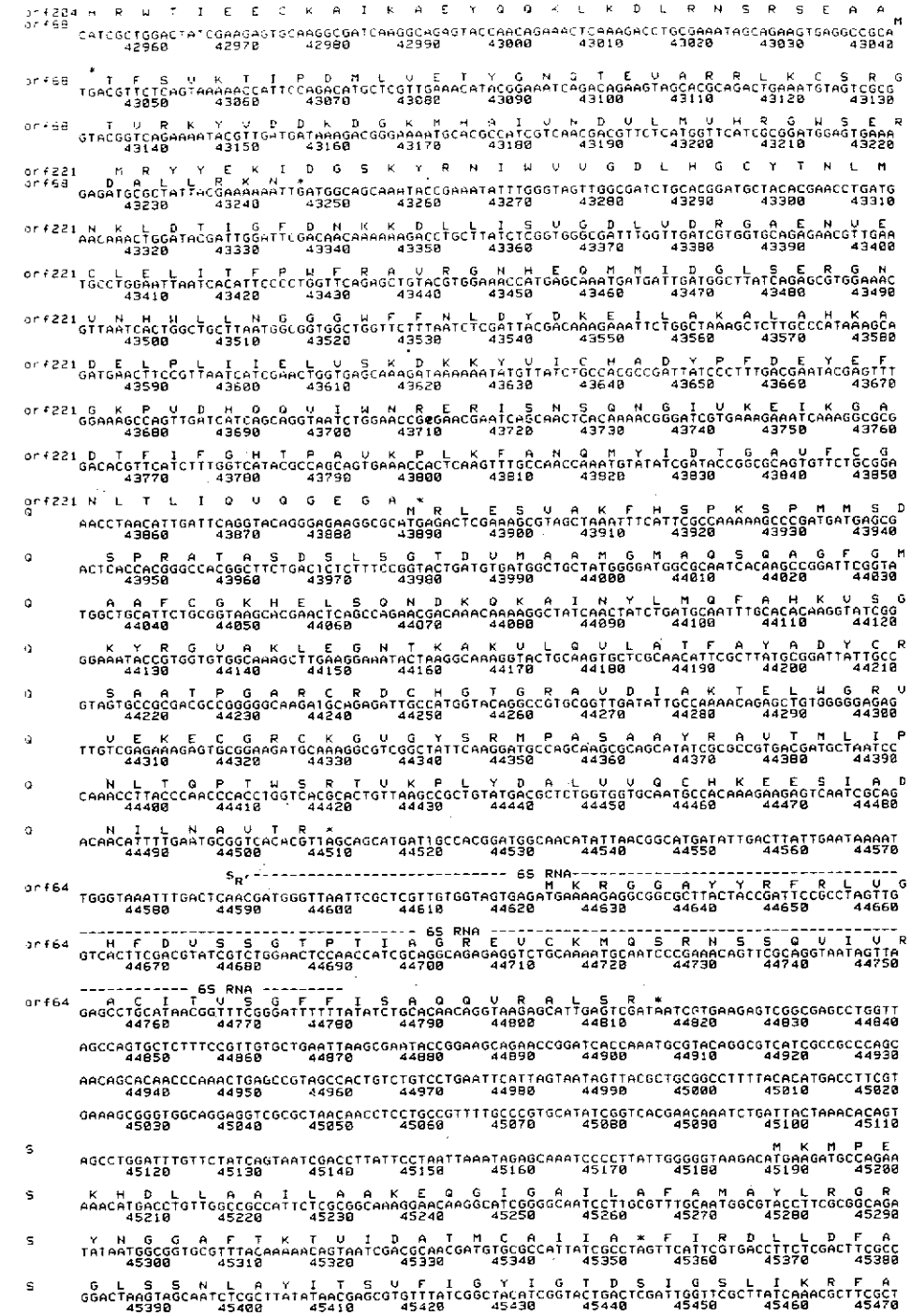
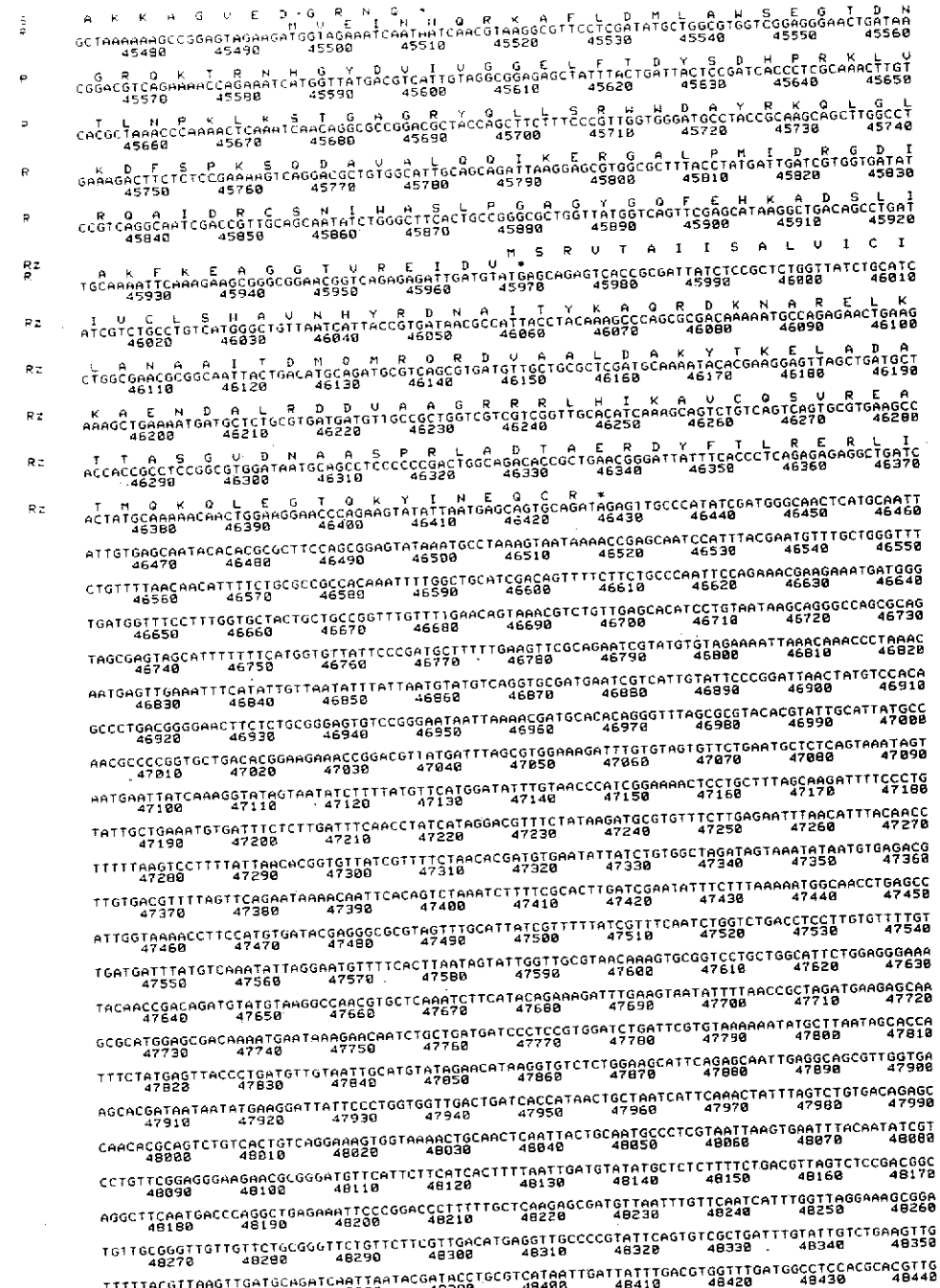


FIG. 2 (b).







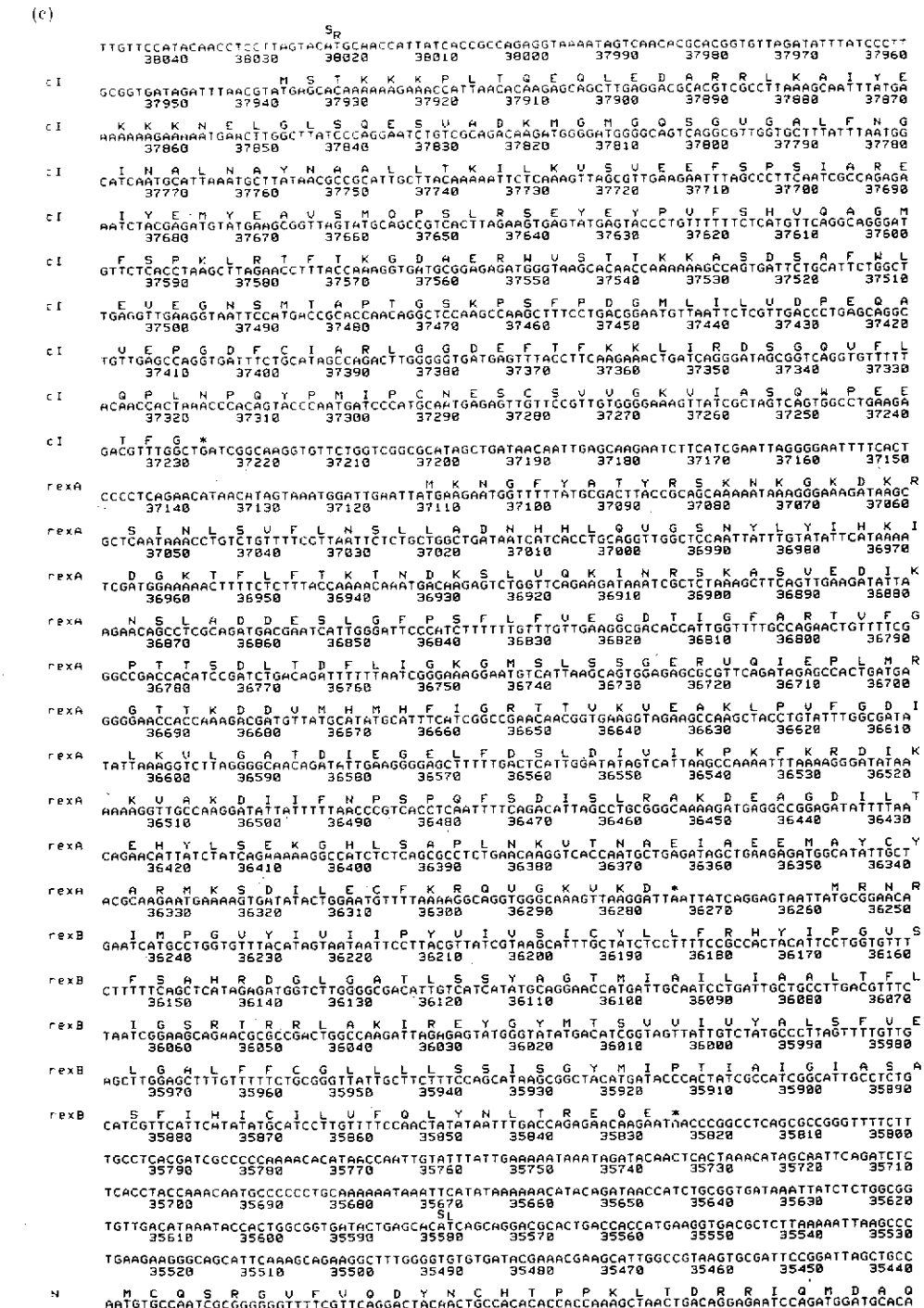
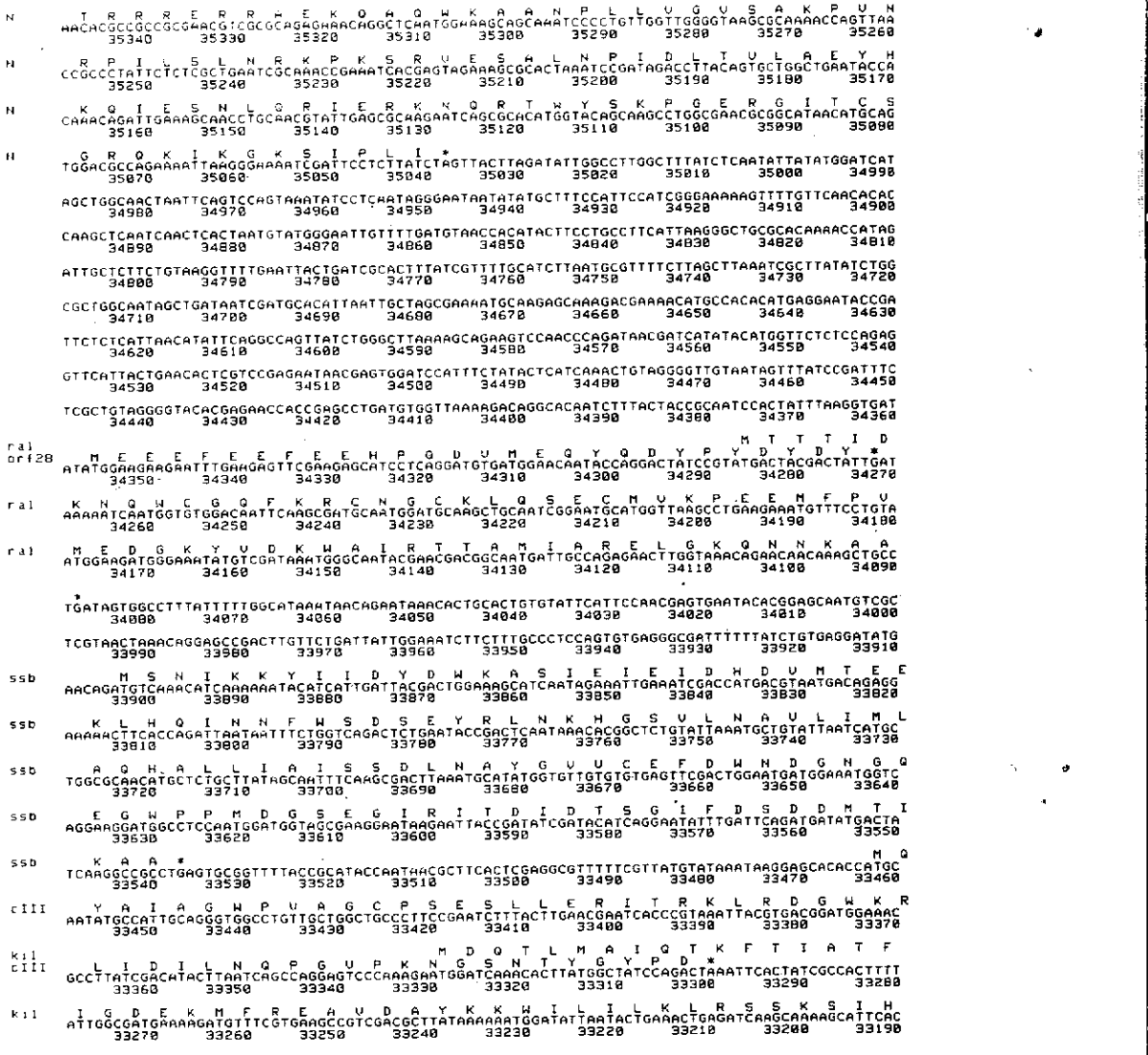
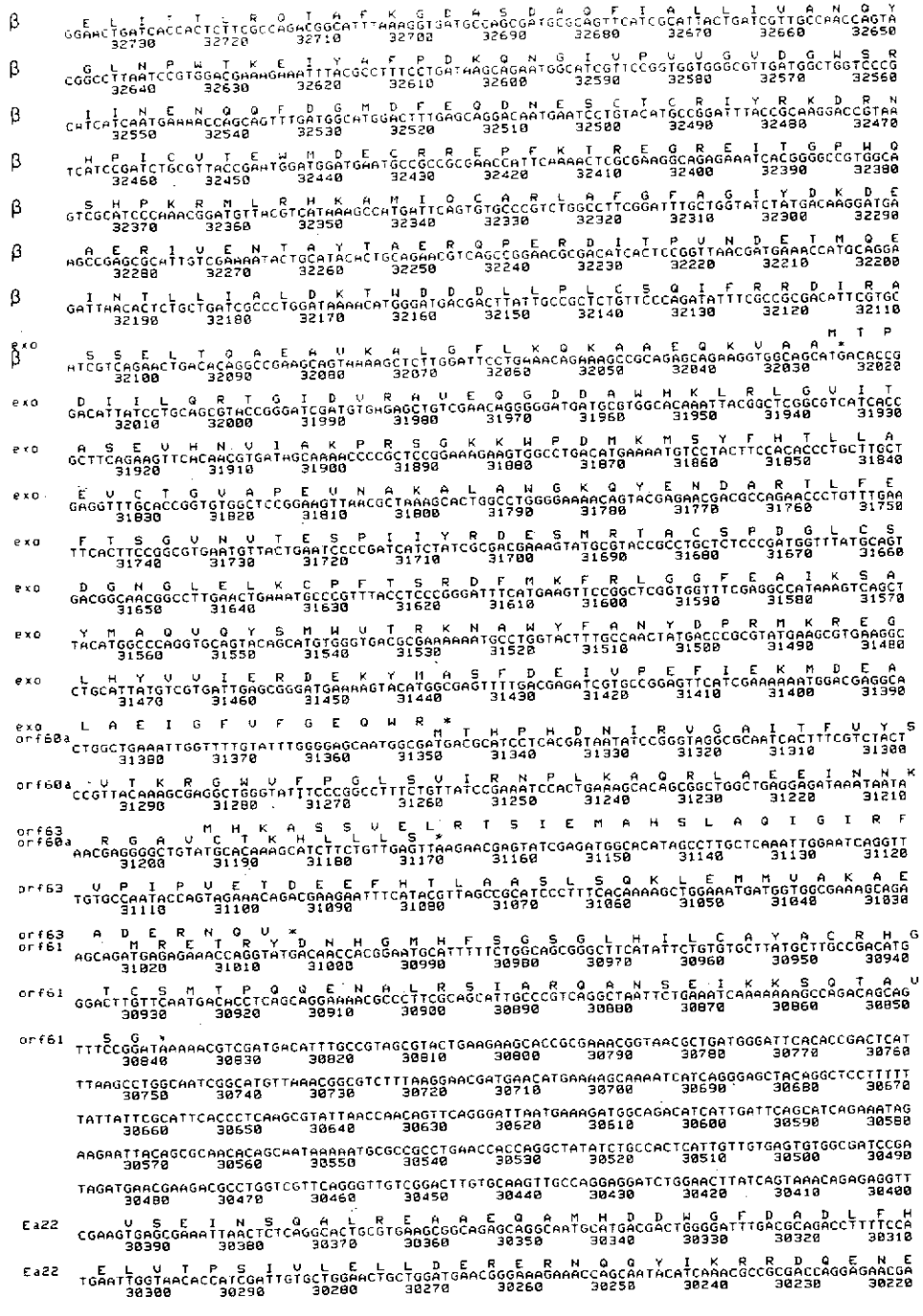
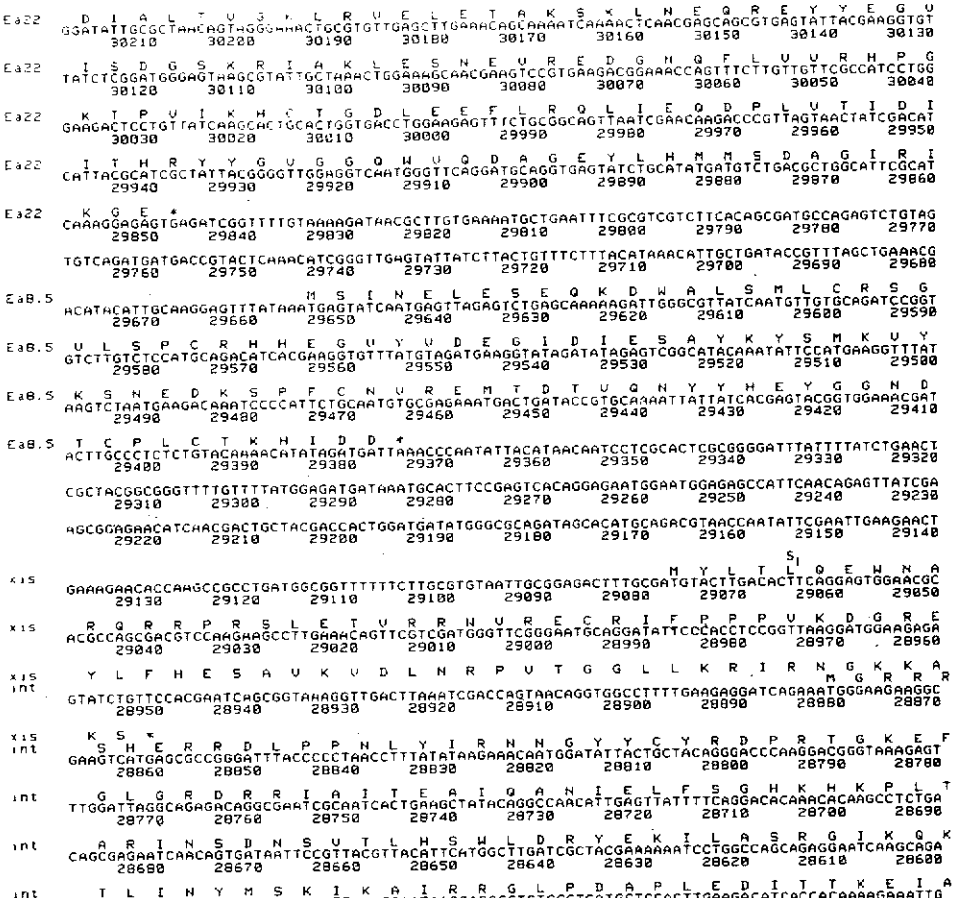
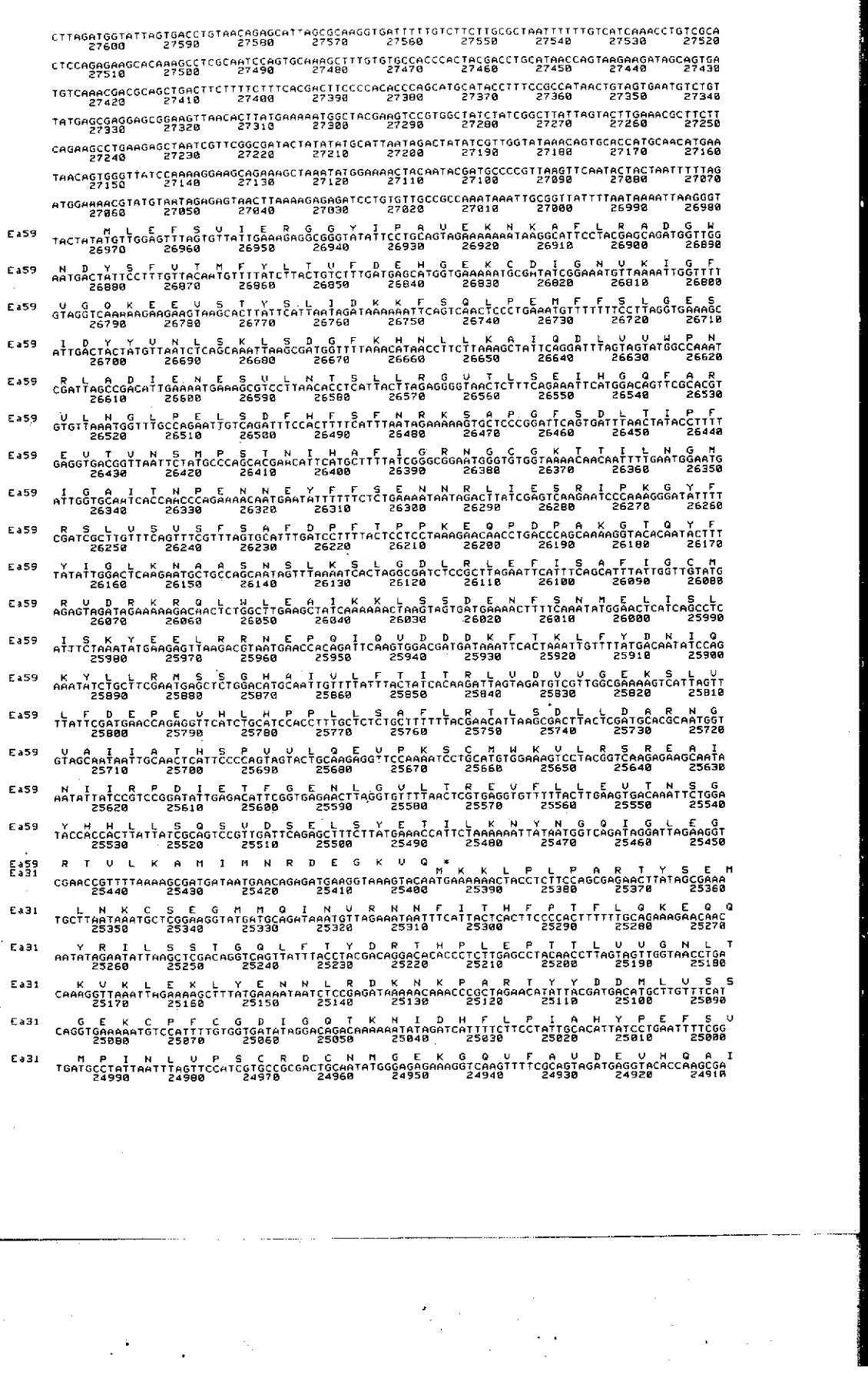


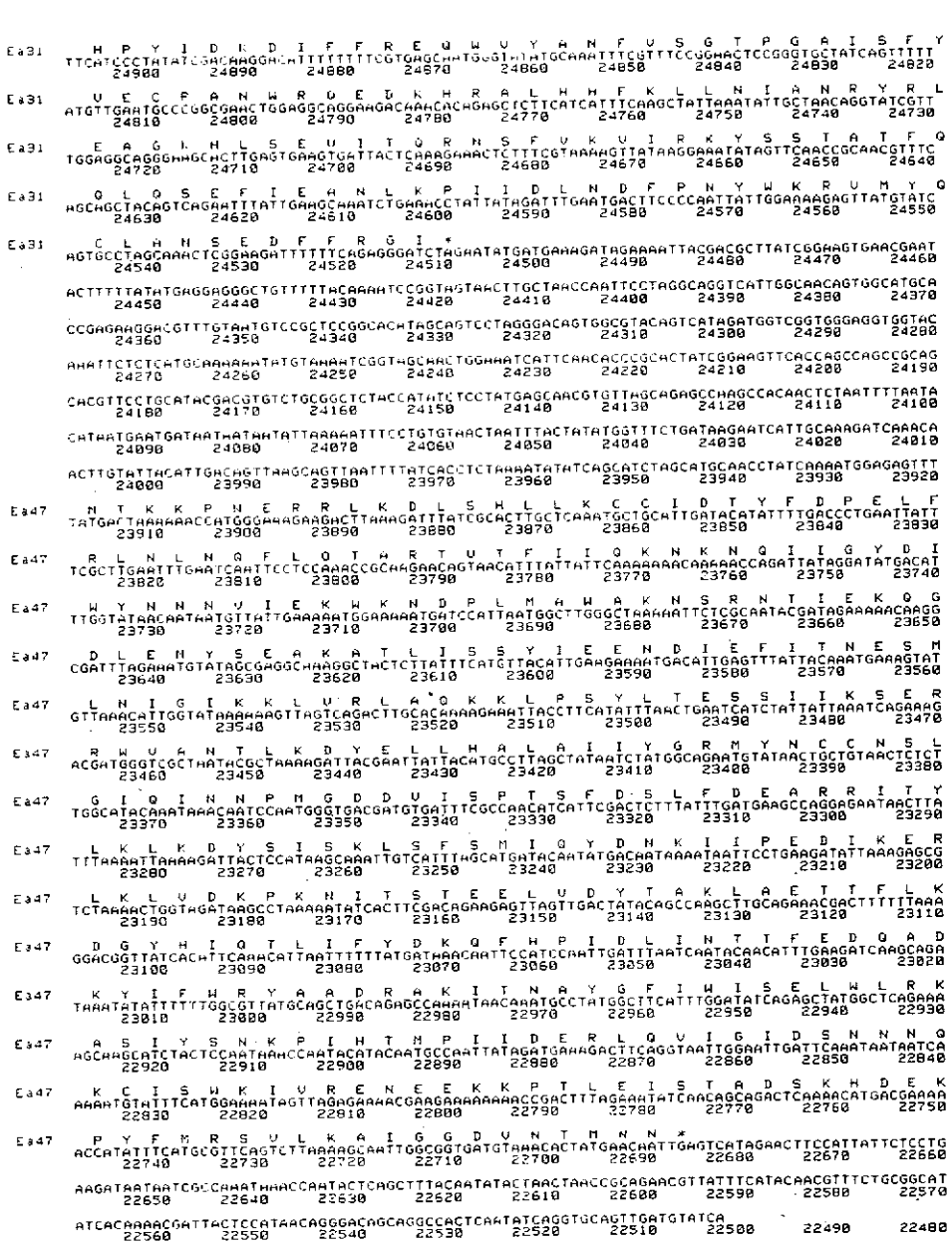
FIG. 2 (c).






EaB2
EaB.5
xis
int
att site





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

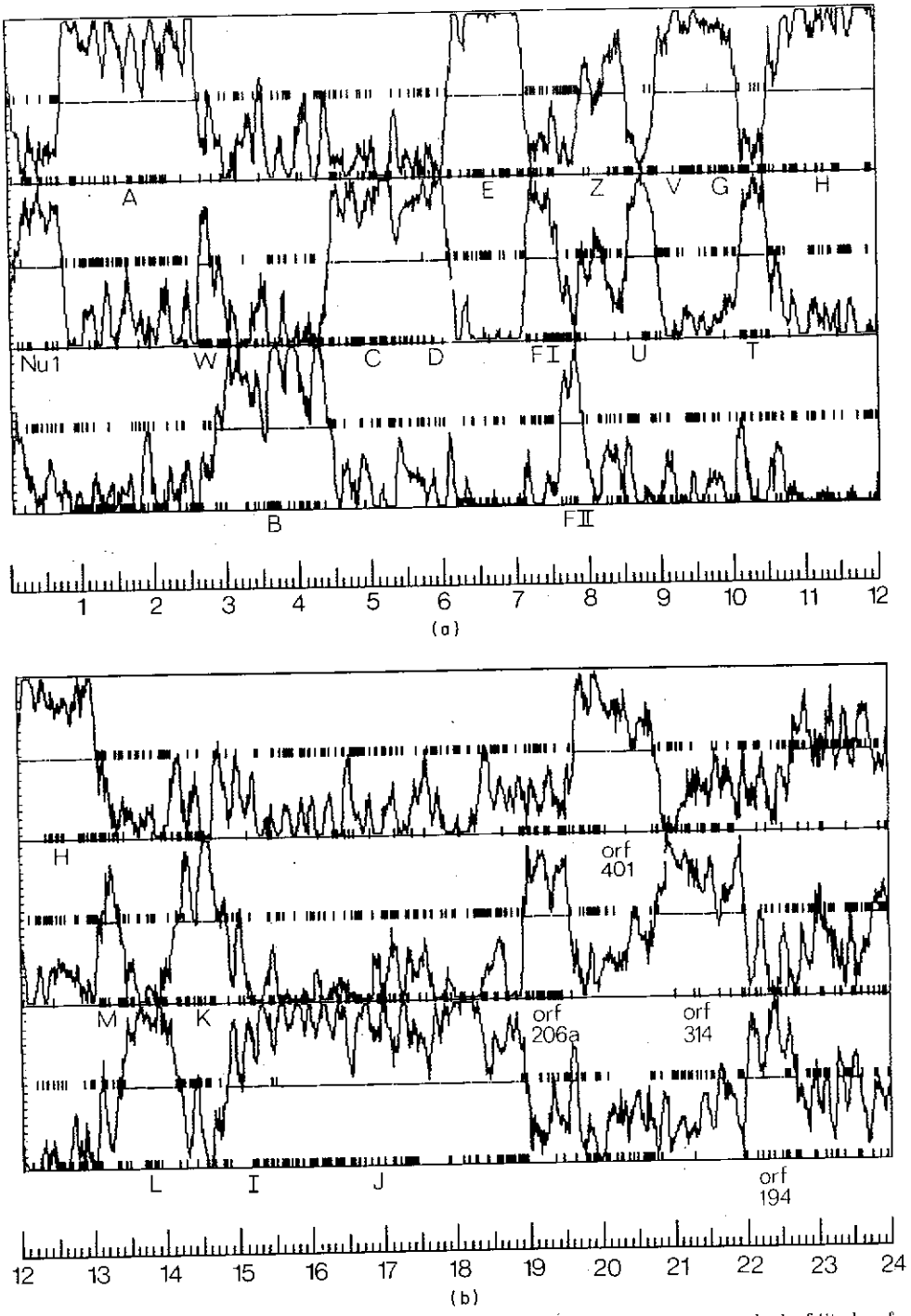
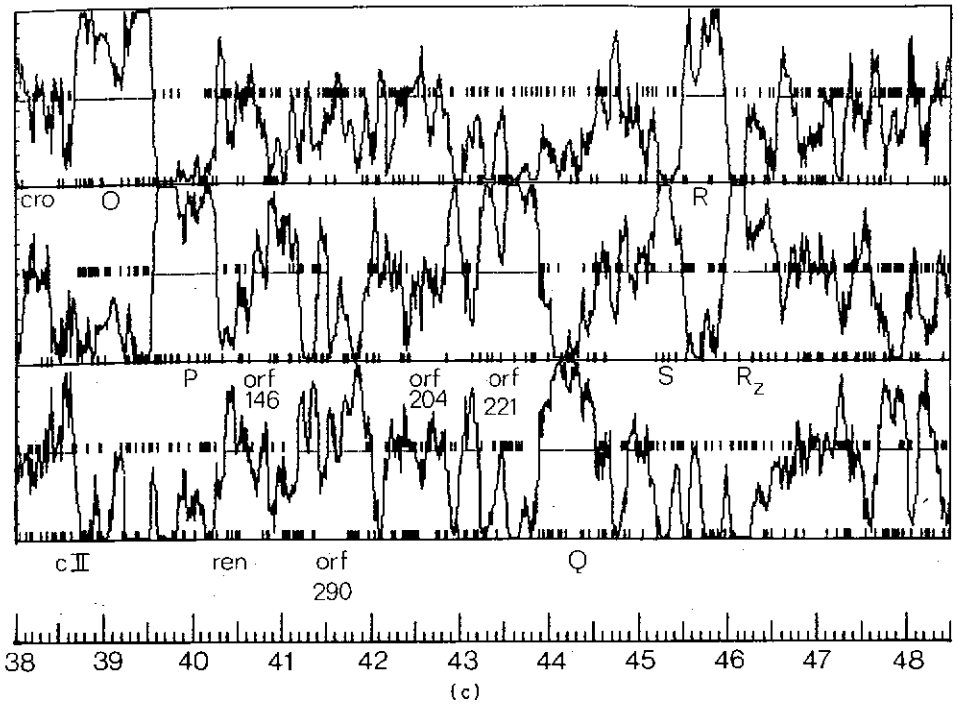
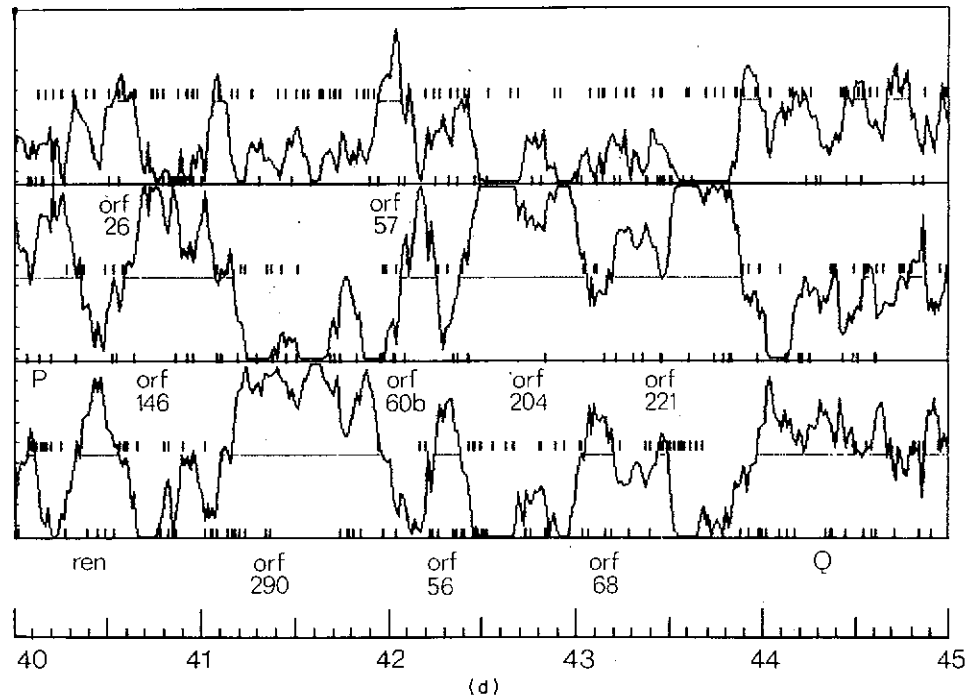


FIG. 3. Gene predictions for the DNA of lambda using the codon preference method of Staden & McLachlan (1982), which is one function of ANALYSEQ, a general sequence analysis program. The x axis represents sections of the lambda sequence (Figs 1 and 2) and the probability of coding is plotted in the y 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.

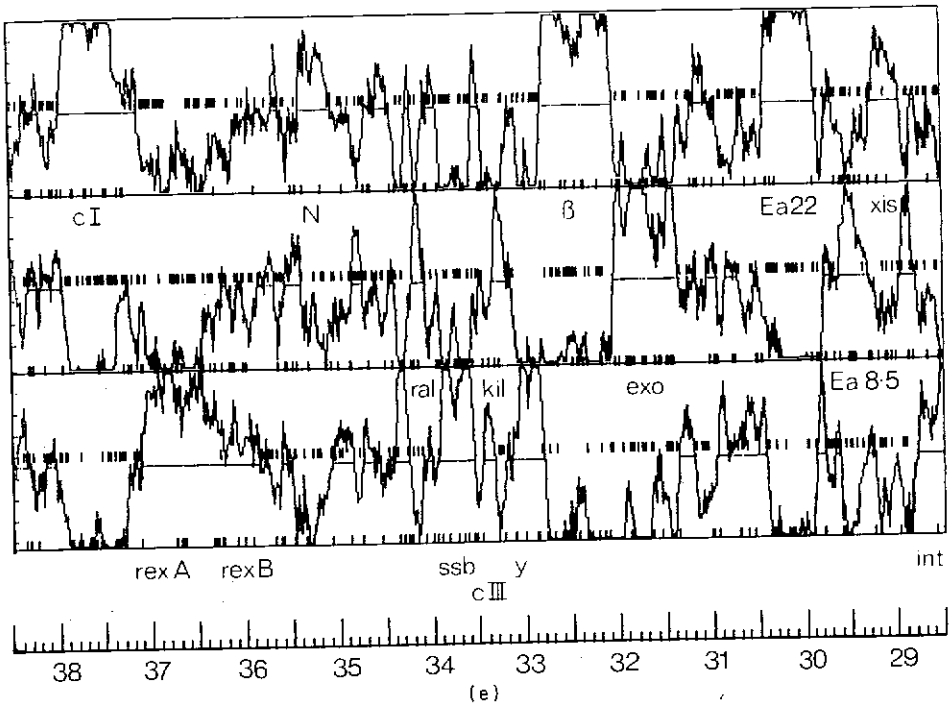


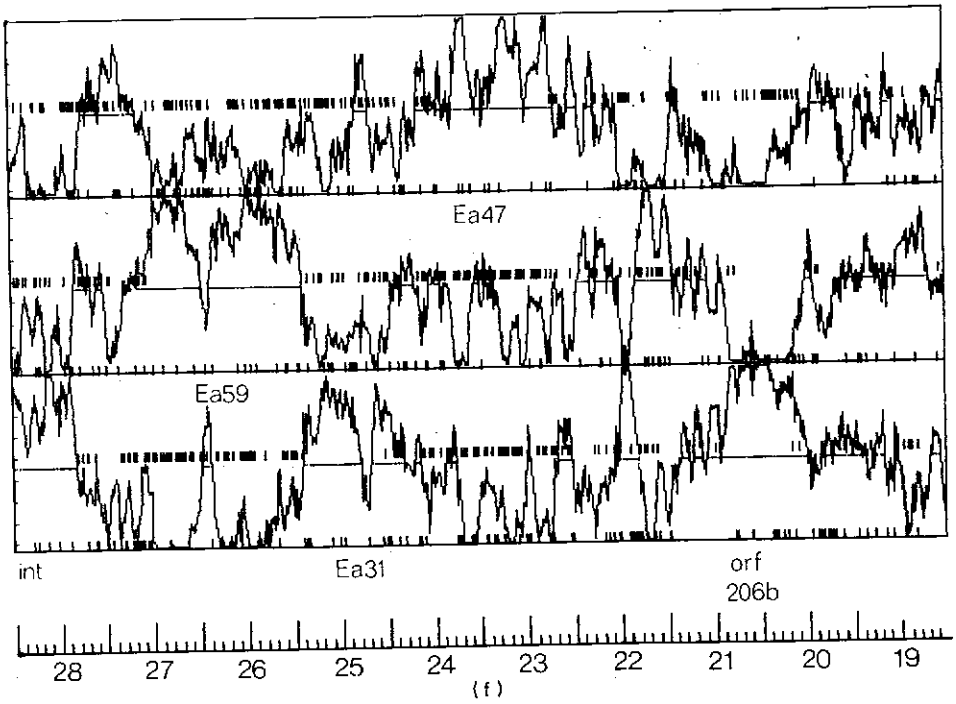
(c)



(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





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: From sequence	Proposed start: From ANALYSEQ	Shine & Dalgarno sequence†	Status (also see the text)
Nu1	191	210	G-G-A(7)ATG	One of 2 possible
A	711	770	G–G-G-T(8)GTG	Uncertain
W	2633	2660	G-G-A-G(8)ATG	Established by amber mutation
B	2836	2880	A-G-G-A(8)ATG	One of 2 possible
C	4418	4430	A-G-G-A-G-G(8)GTG	Best of 6 possibilities
D	5747	—	A-A-G-G(6)ATG	Amino acid sequence
E	6135	6150	G-G-A(7)AGT	Amino acid sequence
FI	7202	7170	G-G-A-G-G(7)ATG	Most probable of 2
FII	7612	7620	G-G-A-G-G(6)GTG	Most probable of 2
Z	7977	7960	A-A-G-G–G-G(6)ATG	Most probable of 3
U	8552	8550	G-G-A(13)ATG	Only possible
V	8955	8970	G-A-G-G(7)ATG	Amino acid sequence
G	9711	—	G-G-A-G(9)ATG	Most probable of 2
T	10,115	10,110	G-G(6)GTG	Uncertain
H	10,542	10,520	A-G-G-A-G-G(10)ATG	Amino acid sequence
M	13,100	13,090–13,160	G-G-A-G-G-T(8)ATG	Most probable of 5
L	13,429	13,410	A-G-G-G-T-T(9)AGT	Only probable
K	14,276	14,250	G-G-A-G-G(12)ATG	Uncertain
I	14,773	14,860	G-A-G-G-T(13)ATG	Most probable of 3
J	15,505	—	G-G-A-G(9)ATG	Most probable
orf206a	18,965	18,960	G-A-G-G-T(8)ATG	Most probable of 3
orf401	19,650	19,660	G-G–G(6)ATG	Most probable
orf314	21,029	20,845	A-G(12)ATG	Most probable (see the text)
orf194	21,937	21,990	G-A-G-G(10)ATG	Only probable
cro	38,041	37,980	A-A-G-G-A-G-G-T(6)ATG	Schwarz *et al.* (1978); Roberts *et al.* (1977)
cII	38,360	38,370	A-A-G-G-A(11)ATG	Schwarz *et al.* (1978); Roberts *et al.* (1977)
O	38,686	38,680	A-G-G-A-G(6)ATG	Schwarz *et al.* (1978)
P	39,582	39,560	G–G-G-T(9)ATG	Schwarz *et al.* (1980)
ren	40,280	40,370	A-A-G-G-A-G(8)ATG	Only probable one
nin region			See the text and Table 4	
Q	43,886	43,910	G-G-A-G(8)ATG	Daniels & Blattner (1982); Petrov *et al.* (1981)
orf64	44,621	—	G–G-G-T(9)ATG	Daniels & Blattner (1982)
S	45,186	45,130	G-G–G-G-T(8)ATG	Daniels & Blattner (1982)
R	45,493	45,490	G-G-A-G(7)ATG	Amino acid sequence (Imada & Tsugita, 1971)
R_Z	45,966	45,960	G-A(7)ATG	Most probable
cI	37,940	37,950	‡	Ptashne *et al.* (1976)
rexA	37,114	37,100	G-G-A(7)ATG	Only possible Landsmann *et al.* (1982)
rexB	36,259	—	A-G-G-A-G(6)AGT	One of 2 possible Landsmann *et al.* (1982)
N	35,438	35,390	G-G-A(10)ATG	Franklin & Bennett (1979)
ral	34,287	34,210	A-G-G-A(8)ATG	Ineichen *et al.* (1981)
ssb	33,904	33,930	A-G-G-A(9)ATG	Ineichen *et al.* (1981)
cIII	33,463	33,460	A-A-G-G-A-G(7)ATG	See the text
kil	33,330	33,340	A-G-G-A-G(10)ATG	See the text
γ	33,112	33,100	A-G-G-A-G(8)ATG	See the text

TABLE 3 (*continued*)

Gene	Proposed start: From sequence	Proposed start: From ANALYSEQ	Shine & Dalgarno sequence†	Status (also see the text)
β	32,810	32,800	G-A-G-G(12)AGT	Only probable
exo	32,028	32,040	A-A-G-G–G-G(6)ATG	Most probable
Ea22	30,395	30,430	G-A-G-G(8)GTG	Only probable
Ea8·5	29,655	29,600	A-A-G-G-A-G(8)ATG	Only probable
xis	29,078	29,250	G-G-A-G(9)ATG	Hoess *et al.* (1980)
int	28,882	28,780	A-G-G-A(6)ATG	Hoess *et al.* (1980)
Ea59	26,973	27,040	A-A-G-G(7)ATG	Only probable
Ea31	25,399	25,410	A-A-G-G(8)ATG	One of 2 possible
Ea47	23,918	23,750	G-G-A-G(7)ATG	Most probable
orf206b	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 *cI* 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§
A	*W*	10/11	0	ATGA
W	*B*	11/15	8	ATGA
B	*C*	14/14	0	GTG(14)TAA
Z	*U*	13/16	2	ATGA
G	*T*	6/6	0	GTG(13)TGA
T	*H*	8/8	6	ATG(2)TGA
H	*M*	8/9	2	ATGA
M	*L*	11/14	3	TGATG
K	*I*	14/19	0	ATG(97)TGA
orf314	*orf194*	7/12	0	TAATG
O	*P*	7/8	3	ATGA
P	*ren*	11/12	3	ATGA
ren	*orf26*	9/11	0	ATGA
orf26	*orf146*	14/14	2	ATGA
orf146	*orf290*	25/29	4	ATGA
orf290	*orf57*	10/10	4	ATGATG‖
orf57	*orf60b*	13/13	3	GTG(28)TAA
orf60b	*orf56*	11/12	0	GTGA
orf56	*orf204*	10/10	0	ATGATG(2)TAA‖
orf204	*orf68*	15/18	4	ATGA
orf68	*orf221*	14/16	0	ATG(17)TGA
orf221	*Q*	10/10	3	ATGA
S	*R*	16/19	0	ATG(11)TAA
R	R_Z	9/11	3	ATGA
β	*exo*	14/17	1	ATGA
exo	*orf60a*	10/12	2	ATGA
orf60a	*orf63*	21/25	4	ATG(22)TAA
orf63	*orf61*	12/14	0	ATG(14)TGA
xis	*int*	12/14	0	ATG(17)TGA
Ea59	*Ea31*	14/16	4	ATGA

† 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, 1981*b*). 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 CUA 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 (Murialdo & 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, 1981*a*,*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 (Murialdo & 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	AGG^{A}_{G}	UCA	
Nu1	1·10	0	0	1·65	2·75	1·10	0·55	
A	0·62	1·16	0	0·31	1·09	0·62	0·47	1
W	0	0	0	4·3	1·45	2·9	0	
B	0·56	0·38	0·19	0·56	2·4	0·38	0·93	
C	0·24	0·24	0·48	0·48	1·7	0	0·73	
Nu3‡	0	0	0·57	0·57	0	0	1·7	230
D	0	0	0	0	0·9	0	0	870
E	0	0	0·88	0	0	0	0·58	470
FI	0	0	1·5	0	0	0·76	0·76	360
FII	1·7	0	0	0	4·2	0	1·7	
Z	2·1	0	1·6	0·52	3·1	3·1	1·6	
U	0	0	2·3	0	1·5	0	2·3	375
V	0	0	0	0	0	0	0·41	265
G	0	0	0	0	2·1	0·71	2·1	18
T	0	0	0	0·69	1·4	0	1·4	
H	0·12	0	0·59	0·35	1·6	0·70	0·59	12
M	0·92	0	0	0·92	1·8	0·92	0	
L	0	0	0	0·43	0·86	0	0	44
K	0	0	0·50	1·0	3·0	0·50	1·5	14
I	0·45	1·3	0·45	0·45	1·3	0·90	2·2	
J	0·88	0·09	0·26	0·26	1·1	0·71	0·35	13
orf206a	0·49	0·49	0·97	0	0·97	0·49	0·49	
orf401	0·75	0	0·75	0·25	0·75	1·2	5·7	
orf314	0	1·3	0·32	0·32	0	0·96	1·6	
orf194	0·52	1·5	0·52	0	2·5	0·52	0·52	
	AUA	UUA	CAA	CGA	CGG	AGG^{A}_{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 (Murialdo & 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).

Nu1

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.

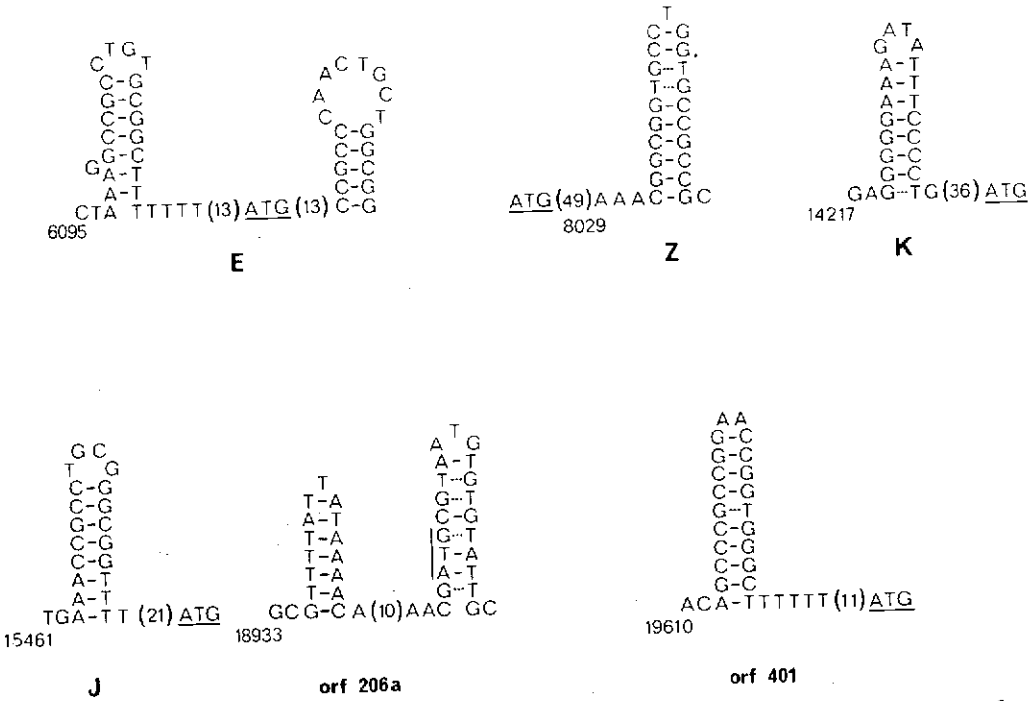


FIG. 4. Possible looped structures found near the initiating sites of some of the genes in the left arm of lambda. The initiating ATG is indicated by underlining and its relative position is shown by the numbers of nucleotides in parentheses. Base-pairing is indicated by dashes and sequence hyphens have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

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 prefered 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).

orf101

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, GCN, 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.

```
                                V  L  L  V  A  L  L  S  C  T  I  P  F  L  V  V  S  A  S  S  S  A  T
CGTTGCAGCAAGCGTTTCAGACGTGCTGTTGGTTGCACTGCTGAGCTGCACTATCCCCTTTCTCGTTGTGTCCGCATCCTCAAGCGCGAC
     20780     20770     20760     20750     20740     20730     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  D  T
AGCTGAAGCTATATCTTCTGCACGTTTTGCCGAATTTTTGCACGTATTGCCGCCGCTTCTGCCGCACTTTGCTCTGCGATGCTGATAC
     20690     20680     20670     20660     20650     20640     20630     20620     20610

  A  L  P  A  A  S  V  A  F  V  D  A  V  D  A  L  P  A  A  A  V  F  A  S  A  A  A  E  A  L
CGCACTTCCCGCAGCCTCTGTCGCCTTCGTGGATGCCGTTGACGCACTCCCCGCCGCCGCTGTTTTTGCGTCTGCCGCGGCAGAGGCGCT
     20600     20590     20580     20570     20560     20550     20540     20530     20520

  R  S  A  A  V  S  D  D  L  A  F  V  S  D  V  F  A  A  L  A  E  F  S  A  A  V  A  E  E  A
CCGTTCCGCTGCTGTTTCAGATGACCTGGCATTCGTCTCGGACGTTTTTGCCGCCCTGGCAGAATTTTCTGCCGCCGTTGCCGAGGAAGC
     20510     20500     20490     20480     20470     20460     20450     20440     20430

  A  R  P  A  L  D  D  A  F  V  S  D  D  F  A  A  S  F  E  A  T  A  S  R  A  E  V  A  A  S
TGCACGACCGGCACTTGATGATGCGTTCGTTTCTGATGATTTTGCTGCCTCTTTTGAGGCCACCGCATCTCGTGCTGAAGTGGCGGCCTC
     20420     20410     20400     20390     20380     20370     20360     20350     20340

  D  A  F  V  A  A  V  E  A  D  V  A  A  D  C  C  D  A  A  A  F  V  S  D  V  F  A  A  P  A
TGACGCTTTCGTGGCCGCGGTGGAGGCAGACGTGGCGGCTGATTGTTGTGACGCTGCAGCATTCGTTTCTGACGTTTTCGCCGCACCGGC
     20330     20320     20310     20300     20290     20280     20270     20260     20250

  L  V  A  A  A  F  F  E  D  S  A  A  A  A  L  F  S  A  S  V  A  F  A  D  A  A  S  A  P  E
ACTGGTGGCCGCCGCGTTTTTTGAGGACTCTGCGGCTGCGGCACTTTTTTCCGCTTCAGTGGCCTTTGCTGATGCCGCTTCTGCGCCGGA
     20240     20230     20220     20210     20200     20190     20180     20170     20160

  D  A  S  *
GGACGCTTCCTGAGCTGACGATGCAGC
     20150     20140     20130
```

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.

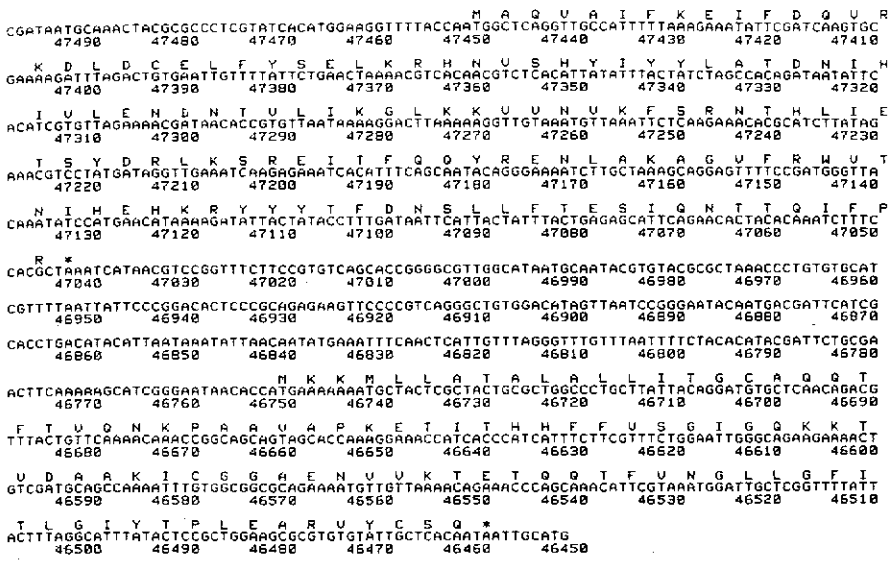


FIG. 6. Nucleotide sequence of the *r* strand (i.e. the opposite strand to that shown in Fig. 2(b)) in the region 46,452 to 47,500 and the amino acid sequence of 2 open reading frames.

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 ATG 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.* (1980*a*,*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; Murialdo & 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; Murialdo & 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). Murialdo & 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*, *I* 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, L. & 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. (1980*a*). *J. Mol. Biol.* **138**, 715–729.

Court, D., de Crombrugghe, B., Adhya, S. & Gottesman, M. (1980*b*). *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. & Murialdo, 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. (1981*a*). *J. Mol. Biol.* **146**, 1–21.

Ikemura, T. (1981*b*). *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., Serpinski, 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.