articles

Nucleotide sequence of bacteriophage ΦX174 DNA

F. Sanger, G. M. Air*, B. G. Barrell, N. L. Brown†, A. R. Coulson, J. C. Fiddes, C. A. Hutchison III‡, P. M. Slocombe§ & M. Smith¶

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

A DNA sequence for the genome of bacteriophage ΦX174 of approximately 5,375 nucleotides has been determined using the rapid and simple 'plus and minus' method. The sequence identifies many of the features responsible for the production of the proteins of the nine known genes of the organism, including initiation and termination sites for the proteins and RNAs. Two pairs of genes are coded by the same region of DNA using different reading frames.

THE genome of bacteriophage ΦX174 is a single-stranded, circular DNA of approximately 5,400 nucleotides coding for nine known proteins. The order of these genes, as determined by genetic techniques[2-4], is *A–B–C–D–E–J–F–G–H*. Genes *F*, *G* and *H* code for structural proteins of the virus capsid, and gene *J* (as defined by sequence work) codes for a small basic protein that is also part of the virion. Gene *A* is required for double-stranded DNA replication and single-strand synthesis. Genes *B*, *C* and *D* are involved in the production of viral single-stranded DNA: however, the exact function of these gene products is not clear as they may either be involved directly in DNA synthesis or be required for DNA packaging, which is coupled with single-strand production. Gene *E* is responsible for lysis of the host.

The first nucleotide sequences established in ΦX were pyrimidine tracts[5-7] obtained by the Burton and Petersen[8] depurination procedure. The longer tracts could be obtained pure and sequences of up to 10 nucleotides were obtained. More recently Chadwell[9] has improved the hydrazinolysis method to obtain the longer purine tracts. These results are included in the sequence given in Fig. 1.

More extensive ΦX sequences were obtained using partial degradation techniques, particularly with endonuclease IV (refs 10 and 11). Ziff *et al.*[12,13] used this enzyme in conditions of partial hydrolysis to obtain fragments 50–200 nucleotides long which were purified as ^{32}P-labelled material by electrophoresis on polyacrylamide gels. The fragments came from the same region of the genome and the sequence of a 48-nucleotide long fragment (band 6, positions 1,047–1,094) was determined using mainly further degradation with endonuclease IV and partial exonuclease digestions.

Another 50-nucleotide long fragment was obtained by Robertson *et al.*[14] as a ribosome binding site. The viral (or plus) strand DNA of ΦX has the same sequence as the mRNA and, in certain conditions, will bind ribosomes so that a protected fragment can be isolated and sequenced. Only one major site was found. By comparison with the amino acid sequence data it was found that this ribosome binding site sequence coded for the initiation of the gene *G* protein[15] (positions 2,362–2,413).

At this stage sequencing techniques using primed synthesis with DNA polymerase were being developed[16] and Schott[17] synthesised a decanucleotide with a sequence complementary to part of the ribosome binding site. This was used to prime into the intercistronic region between the *F* and *G* genes, using DNA polymerase and ^{32}P-labelled triphosphates[18]. The ribo-substitution technique[16] facilitated the sequence determination of the labelled DNA produced. This decanucleotide-primed system was also used to develop the plus and minus method[1]. Suitable synthetic primers are, however, difficult to prepare and as DNA fragments generated by restriction enzymes are more readily available these have been used for most of the work reported here.

Another approach to DNA sequencing is to make an RNA copy using RNA polymerase with α-^{32}P-labelled ribotriphosphates and then to determine the RNA sequence by more established methods. Blackburn[19,20] used this approach on intact single-stranded ΦX and on fragments obtained by digestion with endonuclease IV or with restriction enzymes. Sedat *et al.*[21] were extending their studies on the larger endonuclease IV fragments and their results, taken in conjunction with the transcription of the DNA fragments[20], amino acid sequence of the F protein[22], and the plus and minus method results, made it possible to deduce a sequence of 281 nucleotides (positions 1,016–1,296, Fig. 1) within the *F* gene[23]. Transcription of *Hin*dII fragment 10, amino acid sequence data in the *G* gene, and the plus and minus method using *Hin*dII fragments 2 and 10 as primers, gave a sequence of 195 nucleotides (positions 2,387–2,582, Fig. 1) at the N terminus of gene *G* (ref. 24).

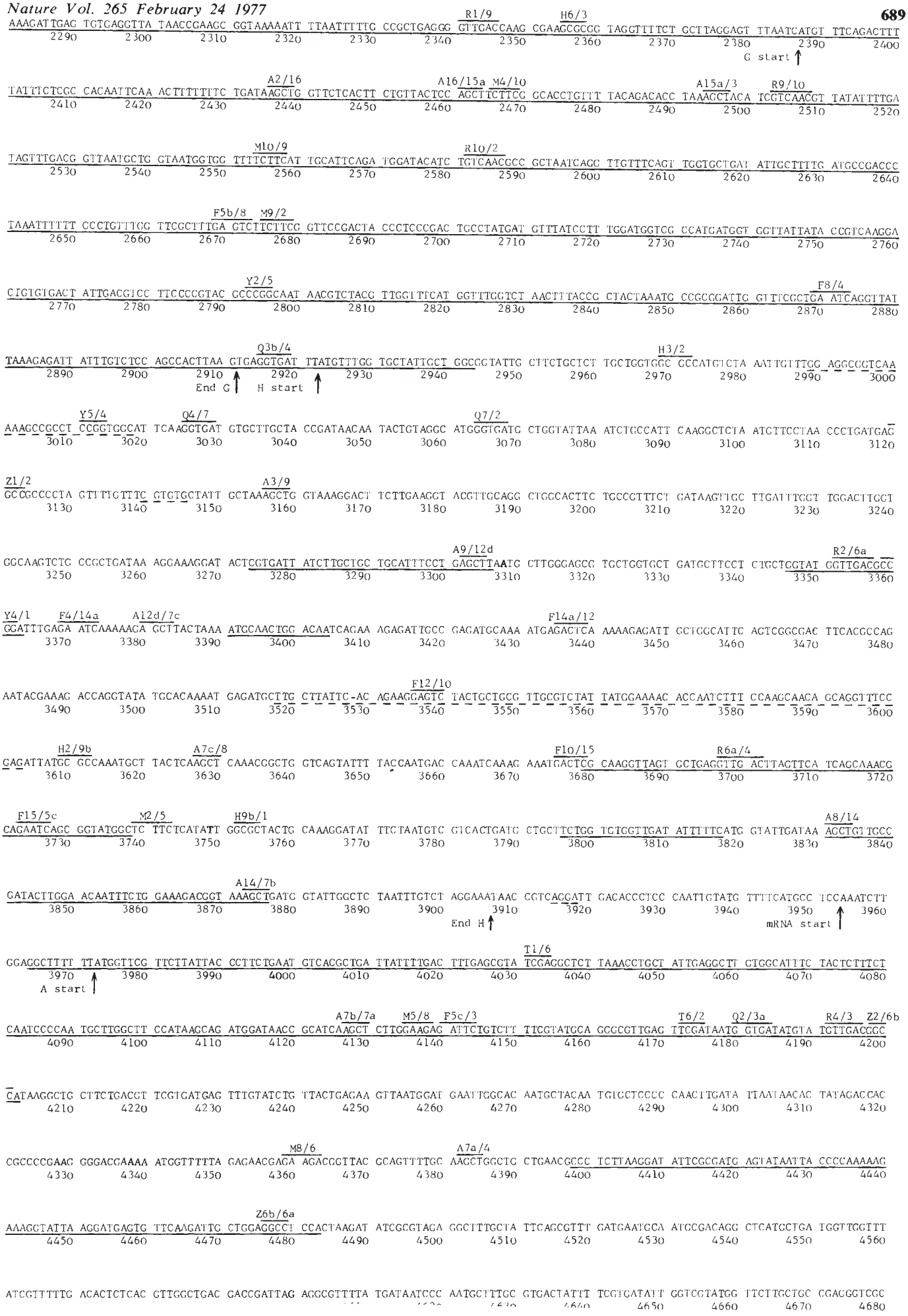
The 'plus and minus' method

Further work on the ΦX sequence has been done using chiefly the plus and minus method primed with restriction fragments. Figure 2 shows the various restriction enzymes used and the fragment maps for each (refs 25–30 and C.A.H., submitted for publication, and N.L.B., C.A.H. and M.S., submitted for publication).

Figure 1 shows the combined results of the sequence work to date. The sequence is numbered from the single cleavage site of the restriction enzyme *Pst*I. As with other methods of sequencing nucleic acids, the plus and minus technique used by itself cannot be regarded as a completely reliable system and occasional errors may occur. Such errors and uncertainties can only

Present addresses: *John Curtin School of Medical Research, Microbiology Department, Canberra City ACT 2601, Australia, †Department of Biochemistry, University of Bristol, Bristol BS8 1TD, UK, ‡Department of Bacteriology and Immunology, University of North Carolina, Chapel Hill, North Carolina 27514, §Max-Planck-Institut für Molekulare Genetik, 1 Berlin 33, FRG, ¶Department of Biochemistry, University of British Columbia, Vancouver BC, Canada V6T 1W5.

Fig. 1 For full description see p. 690.



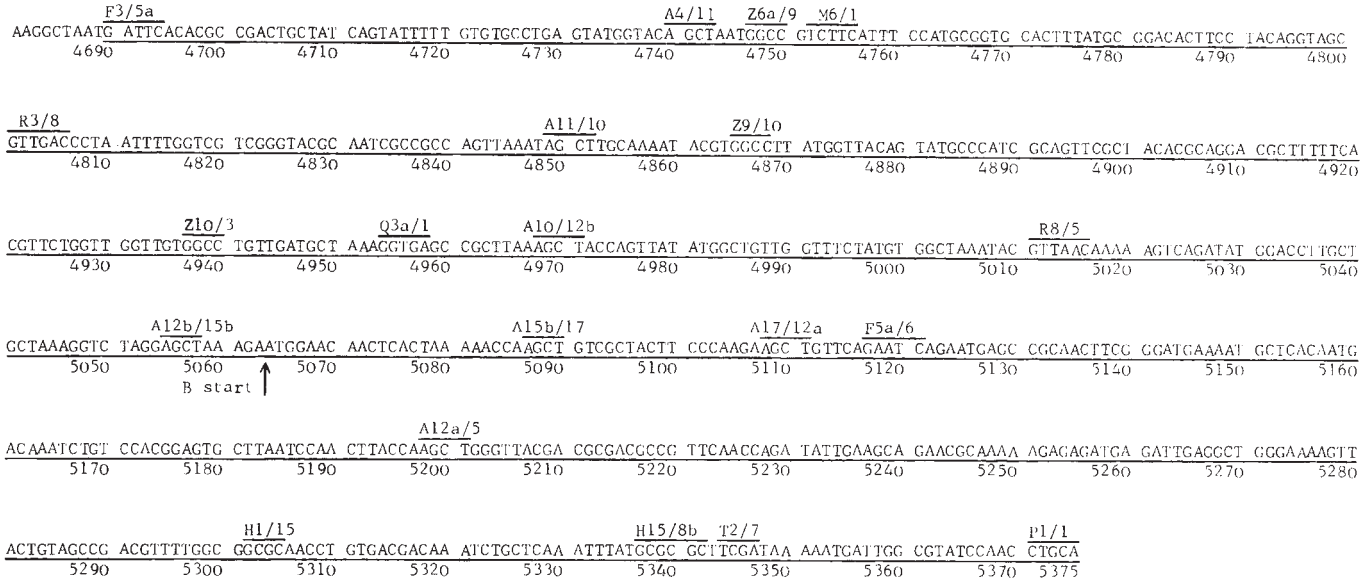


Fig. 1 A provisional nucleotide sequence for the DNA of bacteriophage ΦX174 *am*3 *cs*70. Solid underlining indicates sequences that are fully confirmed; sequences with no underlining probably do not contain more than one mistake per 50 residues. Broken underlining indicates more uncertain sequences. Restriction enzyme recognition sites are indicated (for key to single letter enzyme code see legend to Fig. 2), as are mRNA starts and protein initiation and termination sites. Nucleotides 4,127 to 4,201 have been independently sequenced by van Mansfield *et al.*[58]. The *am*3 codon is at position 587.

be eliminated by more laborious experiments and, although much of the sequence has been so confirmed, it would probably be a long time before the complete sequence could be established. We are not certain that there is any scientific justification for establishing every detail and, as it is felt that the results may be useful to other workers, it has been decided to publish the sequence in its present form.

As template we have used both the viral (plus) and complementary (minus) strands of ΦX. Usually it is possible to determine a sequence with a single primer starting at about 15–100 nucleotides from the appropriate restriction enzyme site. In a particularly good experiment the sequence can be read out to 150–200 nucleotides but the results may become less reliable. Most sequences have been derived by priming on both strands; this allows more confidence than when only one strand could be used.

A useful method for confirming runs of the same nucleotide is depurination of ^{32}P-labelled small restriction enzyme fragments or of products of the DNA polymerase priming experiments (ref. 31 and N.L.B. and M.S., in preparation). The most satisfactory way of confirming the DNA sequences is through amino acid sequence data. As the methods used are entirely unrelated, the results of the two approaches complement each other very well and therefore complete sequences can usually be deduced from incomplete data obtained by each method. The complete sequence of genes *G* (ref. 32), *D* (ref. 33), *J* (ref. 33 and Freymeyer, unpublished) and most of *F* have been obtained in this way.

Many of the sequences in Fig. 1 have been amply confirmed and are regarded as established: these are indicated in the figure by underlining. Some sequences are considered to be reasonably accurate and probably contain no more than one mistake in every 50 nucleotides. Sequences that are particularly uncertain—either because of lack of data or conflicting results—are also indicated in Fig. 1.

In considering the sequence of ΦX174 as a functional unit it is convenient to begin in the region between the *H* and *A* genes and to continue around the DNA in the direction of transcription and translation.

A promoter and terminator

Sinsheimer *et al.*[34,35] and Axelrod[36] have determined the sequences of the 5′ end of three ΦX *in vitro* mRNA species and have located them on the restriction map. These sequences have been identified on the DNA sequence and one of them (AAATCTTGG) is found only at position 3,954 at which an *in vivo* unstable mRNA start has been located[37]. The sequence to the left of this has some characteristics of typical *E. coli* promoters[38] in that five out of the 'ideal' TATPuATPu residues are present. Nearby, to the right of this mRNA initiation, however, is the sequence TTTTTTA which is similar to sequences found at the 3′ ends of a number of mRNAs (see ref. 39) and seems a likely signal for mRNA termination. The presence of a rho-independent termination site in this approximate position has been suggested[36,37], but the relative positions of the initiating and putative termination signals is rather surprising since the terminator for one mRNA would be expected to precede the initiator for the next. One possibility is that the T_6A might be acting as an 'attenuator' involved in the control of mRNA production in a similar manner to that suggested for the tryptophan operon by Bertrand *et al.*[40]. If indeed it were acting as a transcription terminator one would expect a small RNA of 20 nucleotides to be produced, but no such product has yet been detected. Recent work, however, (Rosenberg, unpublished and ref. 41) indicates that termination may require the presence of a base-paired loop structure before the termination site. From the DNA sequence such a loop is probably present before the T_6A sequence, but in mRNA starting from the initiation site at position 3,954 this loop is not formed (Fig. 3). Therefore mRNA that had started at an earlier promoter and extended through the *H* gene would be expected to terminate here, whereas mRNA newly initiated at position 3,954 would not. This could be a way in which the phage has economised on the use of DNA—by having the ends of the two mRNAs overlapping.

The A protein

Where the amino acid sequence is available there is no problem in relating the DNA sequence to its coding properties, but it is

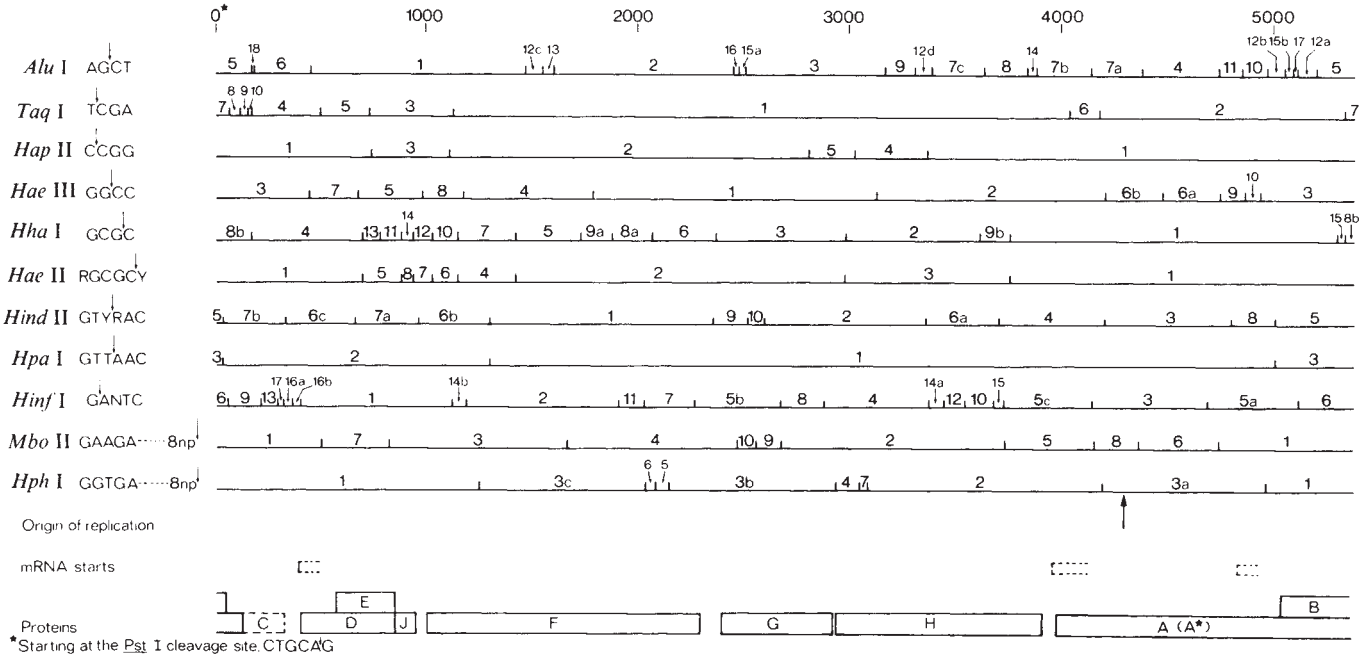


Fig. 2 Fragment maps of restriction enzymes used in the sequence analysis of ΦX174 *am*3 RFI DNA. Fragment maps of ΦX174 have been prepared for *Hin*dII (R), *Hae*III (Z) and *Hpa*I + II by Lee and Sinsheimer[25], *Hin*HI and *Hap*II (Y) by Hayashi and Hayashi[26], and for *Alu*I (A) by Vereijken *et al.*[27] and for *Pst*I (P) by Brown and Smith[30]. B.G.B., G.M.A., C.A.H. and D. Jaffe prepared the *Hin*fI (F) map, C.A.H. the *Hph*I (Q) map, and Jeppesen *et al.*[28] the *Hha*I (H), *Alu*I, *Hae*II and *Hap*II maps by using a rapid method depending on priming with DNA polymerase. A rapid two-dimensional hybridisation technique has been developed by C.A.H. (submitted for publication) and recently used for mapping *Mbo*II (M) (N.L.B., C.A.H., and M.S., submitted for publication) and *Taq*I (T)[29]. *Hha*I and *Hin*fI maps have also been prepared by Baas *et al.*[52].

more difficult to do so in the absence of such data, as is the case for the A protein. One way of identifying the reading phase of the DNA is from the distribution of nonsense codons. Over a sufficiently long sequence that is known to be coding for a single protein there is usually one phase that contains no nonsense codons, and this is identified as the reading phase. This requires completely accurate determination of the DNA sequences however: omission of a single nucleotide may give completely erroneous results. Another approach is possible in the case of ΦX. The results with the *F* and *G* genes[23,24,32] showed an unexpectedly high frequency of codons ending in T. Therefore in a coding region there is a tendency for every third nucleotide to be a T and it is then possible to define the reading phase. Figure 4 illustrates how this characteristic was used to help determine the reading phase for the A protein and to identify its initiation codon at position 3,973. In a similar way the distribution of Ts may be used to identify errors in the DNA sequence, provided that such errors occur only infrequently.

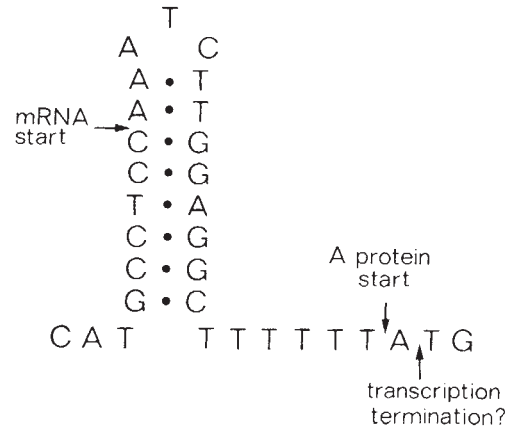
A different approach to identifying the initiation site and reading phase in a coding sequence is by looking for a characteristic 'initiation sequence'. Shine and Dalgarno have shown that a common feature of ribosome binding sites is a number of nucleotides (at least three) preceding the ATG that are capable of forming base pairs with a sequence at the 3′ end of 16S rRNA[42,43]. All of the known initiation sites in ΦX174 that have been identified by direct amino acid sequencing (for the F, G, H, J and D proteins) satisfy this criterion (see Table 2) and the fact that the sequence preceding the ATG in position 3,973 also has this characteristic supports its identification as the initiation site for the A protein.

If, as has been suggested[37], some mRNA from the previous promoter does extend beyond the hairpin structure, initiation of A protein synthesis may be controlled by the inclusion of the region complementary to the 16S rRNA in the hairpin loop. This could explain the presence of two types of mRNA covering the *A* cistron, as suggested by Hayashi *et al.*[37]—one unstable and active and the other stable but inactive. The former would be intiated at the A promoter, the latter at an earlier promoter and result from 'read-through' at the terminator. The postulated reading frame for the A protein was confirmed by sequencing amber mutants mapping in the N-terminal region of gene *A*. *am*86 proved to be a C→T change at position 4,108 and *am*33 a C→T change at position 4,372. These both result in formation of an amber codon (TAG) in the same reading frame as the proposed initiating ATG and the sequence continues to the termination codon at position 133. The A protein, which is the largest coded by ΦX174, is thus 512 amino acids long with a molecular weight of 56,000, in good agreement with SDS gel estimations (see refs 4 and 44). The A* protein, with a molecular weight of about 35,000, is believed to result from an internal translational start in the *A* gene, in the same reading phase[45]. From consideration of possible ribosome binding sequences[42,43] the ATG in position 4,657 seems to be the most likely initiation site for the A* protein.

The origin of replication

The origin of ΦX viral strand DNA synthesis has been located in gene *A*, in restriction fragment Z6b (ref. 46). This origin, while

Fig. 3 Potential secondary structure at the *A* mRNA start.



	Ts in codon position (1)	(2)	(3)
3,910 C C G . T C A . G G A . T T G . A C A . C C C . T C C . C A A . T T G . T A T .	5	2	1
3,940 G T T . T T C . A T G . C C T . C C A . A A T . C T T . G C A . G G C . T T T .	3	4	3*
	2	5	5
3,970 T T T . A T G . G T T . C G T . T C T . T A T . T A C . C C T . T C T . G A A .	3	5	7*
	5	3	7
4,000 T G T . C A C . G C T . G A T . T A T . T T T . G A C . T T T . G A G . C G T .	5	0	7*
	4	2	7

*These figures refer to the last five codons of the previous line and the first five of the next line.

Fig. 4 Identification of the initiation codon for the A protein. Sequences of 30 nucleotides in the region in which the initiation was expected were written down and arbitrarily marked off in triplets. The number of T residues in the first position in each triplet was then counted and listed, and similarly the number of Ts in the second and third positions. The marked preference for Ts in position 3 in the last two lines, as compared with the first two lines, suggests that they are coding for protein and that the triplets are correctly marked off. The most likely initiation codon for the A protein is the ATG in position 3,973.

coding for part of the A protein, probably corresponds to the position of the plus strand nick made by the same protein[44]. Gaps in this region that are found in replicating double-stranded (RF) DNAs are probably related to the position of the nick. Eisenberg *et al.*[47] have investigated such gaps by depurination analysis and identified, in particular, the product C_6T. The sequence CTC_5 is found in position 4,285 (Fig. 1) and the location of the origin in this region agrees precisely with the results of Baas *et al.*[46]. It is not possible at present to identify the actual position of the origin nick. The region shows no apparent secondary structure or symmetrical sequences, although there is an AT-rich region (4,298–4,307) between two GC-rich regions which might be of significance. Such a region is found near the origin of replication of SV40 DNA (ref. 48).

B promoter

The second of the mRNA 5′ sequences (AUCGC)[34] has been mapped in restriction fragment R8 (Fig. 2), which starts about 300 nucleotides on from the proposed A* initiation. The sequence ATCGC is found at positions 4,832 and 4,888 in Fig. 1. The only way we can choose between them at the moment is that the second is preceded by the sequence TACAGTA (position 4,877), which is more akin to sequences found in known promoters[38] than are sequences preceding the other possible mRNA start. Irrespective of which of these sequences is used, the mRNA has a long 'leader' sequence (232 or 176 nucleotides) before the next proposed initiation codon (gene *B*).

Table 1 ΦX174 coding capacity

Gene	Protein molecular weight from SDS gels*	Number of nucleotides (Fig. 1)	Protein molecular weight from sequence information
A	55,000–67,000	1,536	56,000
(*A**)	35,000		
B	19,000–25,000	(360)†	13,845‡
C	7,000		
D	14,500	456	16,811‡
E	10,000–17,500	(273)†	9,940
J	5,000	114	4,097‡
F	48,000	1,275	46,400
G	19,000	525	19,053‡
H	37,000	984	35,800
Non-coding and *C*		485	
Total		5,375	

*See ref. 4.
†Values in parenthesis are overlapping sequences and therefore not included in the addition to obtain the total length of DNA.
‡These values are calculated from the amino acid sequence (in the case of B deduced from the nucleotide sequence). The others are derived using the formula

$$\text{Protein molecular weight} = \frac{\text{No. of nucleotides}}{3 \times 0.00915}$$

The B protein

From a study of the ribonuclease T_1 digestion products of the ribosome binding sites of ΦX mRNAs[49], it was possible to identify an initiating ATG in position 5,064. From the genetic map[2,3], this would be expected to be gene *B* but, as discussed above, the A protein coding sequence extends right through this region, past the *Pst* site at residue 1 in Fig. 1, and terminates at residue 133. The initiating codon contained in the ribosome-protected sequence is, however, out of phase with the A protein reading frame. The proposed B protein coding sequence is one nucleotide to the left of the A protein phase, and continues until a termination codon occurs at position 49. Therefore the B protein coding sequence is totally contained within the *A* gene. These reading frames have been confirmed by sequencing mutants in genes *A* and *B* (*am*16, N.L.B. and M.S., in preparation; *am*18, *am*35, *ts*116 (ref. 50). Since the B protein has not been purified no protein sequence data is available. The complete amino acid sequence can be predicted from the DNA sequence however. The protein is 120 amino acids long with a molecular weight of 13,845 (including the N-terminal Met). The molecular weight estimates of the B protein obtained by SDS–gel electrophoresis are mostly greater than this (see review, ref. 4), but the electrophoretic mobility varied with gel concentration and cross linker. Such anomalous behaviour suggests that there may be, for instance, carbohydrate attached to the B protein.

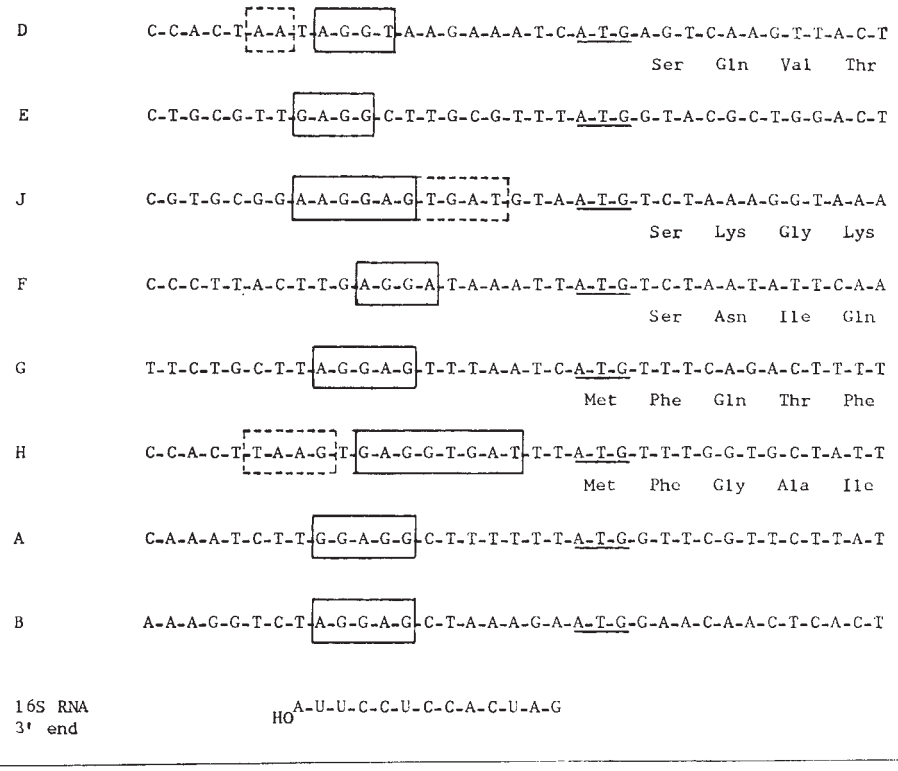
The C protein

The next known gene product, protein C, maps between genes *B* and *D*. Examination of the DNA sequence in this region indicates that the most probable initiating ATG overlaps the termination codon, TGA, of gene A in the sequence ATGA at position 134. A possible termination codon for gene *C* could then be at position 391, although the sequence and phasing is not yet confirmed through this region. There is another possible protein initiation codon (position 51, overlapping the B protein termination codon) which would result in a slightly shorter gene product terminating at nucleotide 219. For the C protein, however, we favour the 'A terminator' start, since only this reading frame contains a CAA sequence, which by a C → T alteration could give the ochre 6 mutant. Ochre 6 is a gene *C* mutant produced by the decay of ^{3}H-cytosine[51] and has been mapped in fragments A6 and F9 (ref. 52); that is, between nucleotides 170 and 205 (Fig. 1).

Sequence following the *D* promoter

The mRNA 5′ sequence which maps before the *D* gene (GAUGC)[34] is found at position 358 in Fig. 1. The sequence preceding the messenger start has only four of the TATPuATPu nucleotides[38]. Thirty-two nucleotides after the mRNA initiation is the ATG (position 390) that initiates D protein synthesis. The amino acid sequence of the D protein has been determined

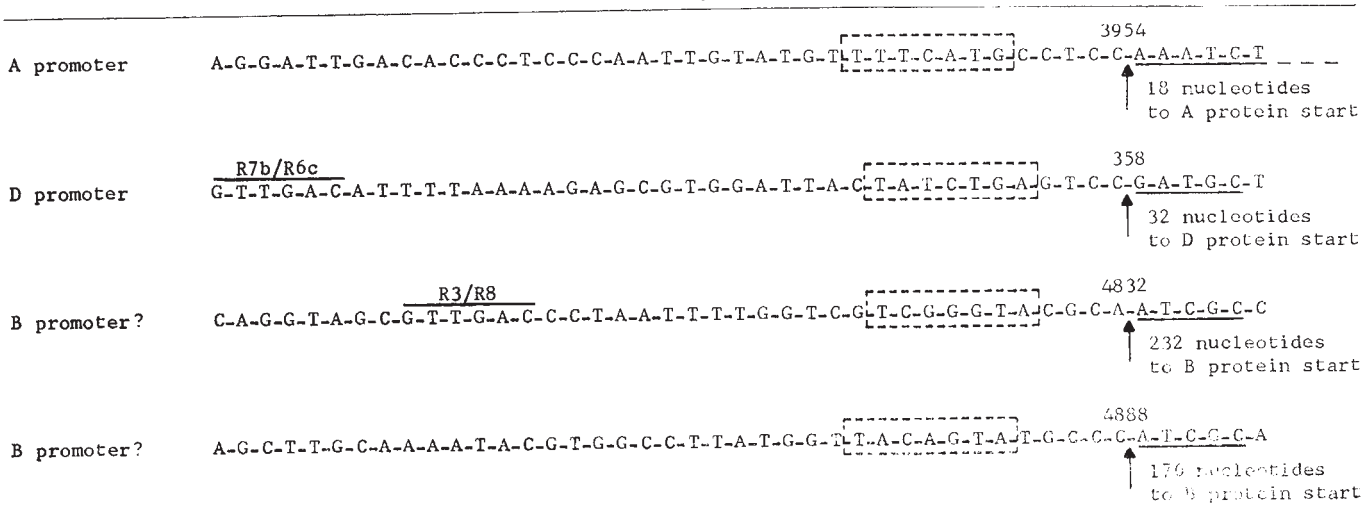
Table 2 Initiation sequences of ΦX174 coded proteins



Where the protein start has been independently confirmed by protein sequencing data the amino acid sequences are indicated. The other initiation regions were identified as described in the text. Sequences complementary to the 3′ end of 16S rRNA (refs 42, 43) are boxed; broken lines indicate further complementarity if some nucleotides are looped out or not matched. Ribosome binding to mRNA has been demonstrated in these regions for genes *J*, *F*, *G* and *B* (ref. 49).

almost completely, and nucleotide and amino acid sequences can be correlated to the termination codon at position 846 (ref. 33). The D protein, which is involved in capsid assembly, is 151 amino acids in length, with a molecular weight of 16,811. The *D* termination codon overlaps the initiation codon for gene *J* in the sequence TAATG. A similar structure has also been found by Platt and Yanofsky[53] in the tryptophan operon. The DNA sequence following this initiation codon matches the amino acid sequence of the small basic protein (37 amino acids) of the virion determined by D. Freymeyer, P. R. Shank, T. Vanaman, C.A.H. and M. H. Edgell (personal communication). Benbow *et al.*[2,3] suggested that the mutation *am6* was located in a gene *J*, coding for the small protein of the virion, and mapping immediately before gene *F*. Although marker rescue experiments indicate that *am6* is not in this region[54], the DNA sequence shows that there is a gene coding for the virion protein and we

Table 3 Promoter sequences in ΦX174



mRNA initiation sequences[34–36] are underlined. Boxed regions indicate sequences that may correspond to the TATPuATPu sequence found in other promoters[38], taking into account the distance from the mRNA starts.

have defined this as gene *J* (ref. 33). Since the *J* initiation codon overlaps the *D* termination codon we had to look elsewhere for gene *E*, which genetic mapping[2,3] had placed between them. Amber mutants in gene *E* (*am*3, *am*27, *am*34 and *am*N11) were located by the marker rescue technique and sequenced. All were found to be within the *D* coding sequence, with the mutant amber codons one nucleotide to the right of the *D* reading frame[33]. Thus the *E* coding sequence is completely contained within the *D* coding region but in a different reading frame. The proposed initiation and termination codons for the E protein are at nucleotides 568 and 840, respectively[33], giving a protein 91 amino acids in length with a molecular weight of about 9,900 (including the N-terminal methionine).

The F protein

Following the *J* gene is an intercistronic region of 39 nucleotides before initiation of the F protein. There is no known function of this apparently untranslated sequence, although the presence of a hairpin structure (positions 969–984) suggests that it could be the site of the *in vivo* messenger termination signal [37] mapped in this region. The F protein is initiated by the ATG at position 1,001. This is the capsid component of the virion, and almost all the amino acid sequence is known[22,24]. There are regions in this gene where the DNA sequence is not completely established, but the protein is about 424 amino acids in length, giving a molecular weight of $\simeq$46,300.

The G protein region

The termination signal for the F protein (position 2,276) is followed by an unusually long untranslated sequence of 111 nucleotides until the G protein initiation codon[31]. This region contains a looped structure which was postulated to have some functional role, as yet unknown, in the single-stranded DNA or the mRNA.

Initiation of the G protein at position 2,387 is followed by a sequence of 425 nucleotides until termination at position 2,912, giving a spike protein of molecular weight 19,053. The nucleotide and amino acid sequences of this gene and product are known[24,32].

The H protein

The initiation codon for the H protein (position 2,923) was identified first on the basis of the distribution of T nucleotides between the three reading phases, and later confirmed by amino acid sequence analysis. Amino acid sequence data on the H protein is minimal but the five peptide sequences known do correspond to the amino acid sequence, deduced from the DNA sequence by using the high frequency of third position T to help in assigning a reading frame to any given region. The DNA sequence is not entirely confirmed but it is possible to write a reasonably accurate amino acid sequence for the H protein. The protein terminates at nucleotide 3,907, in agreement with carboxypeptidase results, giving a spike protein of molecular weight $\simeq$ 35,600 (326 amino acids). The amino acid sequence at the N terminus seems to be particularly rich in hydrophobic residues, which is consistent with its suggested function as the 'pilot' protein that reacts with the bacterial membrane[55,56]. After H protein termination there are 66 nucleotides before initiation of the A protein at position 3,973.

Coding capacity of the ΦX174 genome

The most striking feature of the ΦX DNA sequence is the way in which the various functions of the genome are compressed within the 5,375 nucleotides. Since the identification of ΦX gene products[2,4] it has been clear that proteins of the accepted molecular weights could not be separately coded on the available length of DNA. However, with the presence of two pairs of overlapping genes (*B* within *A* (ref. 50), *E* within *D* (ref. 33)) the genome has more coding capacity than had been originally supposed on the assumption that each gene was physically separate. Table 1 summarises the molecular weights of the known ΦX-coded proteins. There are other potential initiation sites for polypeptide synthesis (for example, in genes *A*, *F*, *G* and *H*) and further genetic work may clarify whether there are in fact other ΦX genes as yet unidentified.

Initiation of protein synthesis

Table 2 lists the protein initiation sequences for genes *A*, *B*, *D*, *E*, *J*, *F*, *G* and *H*. It can be noted that there are no extra precursor sequences in proteins D, J, F, G or H at either the N or C terminus. There seems to be no relationship between the degree of complementarity to the 16S rRNA and the amount of protein synthesised, and we see no other features in the sequence that could explain different efficiencies of translation except where genes overlap.

Transcription of ΦX174

The sequences preceding known mRNA starts[34–36] are shown in Table 3. Other studies on promoter sequences[38] have suggested certain features that they may have in common. Although some of these features are present in the sequences preceding the ΦX transcription initiations others are not, and at present it is difficult to suggest what signal on the DNA determines a promoter site or the efficiency with which it initiates RNA synthesis. It is interesting to note that a polymerase binding site found by Chen *et al.*[57], but not associated with any *in vitro* or *in vivo* mRNA starts, mapped near the region where there is the sequence TATGATG characteristic of promoters[38] (positions 2,705–2,711).

Table 4 Codons used in ΦX174

Phe	TTT	39	Ser	TCT	35	Tyr	TAT	36	Cys	TGT	12
	TTC	26		TCC	9		TAC	15		TGC	10
Leu	TTA	19		TCA	16	Ter	TAA	3	Ter	TGA	5
	TTG	26		TCG	14		TAG	0	Trp	TGG	16
Leu	CTT	36	Pro	CCT	34	His	CAT	16	Arg	CGT	40
	CTC	15		CCC	6		CAC	7		CGC	29
	CTA	3		CCA	6	Gln	CAA	27		CGA	4
	CTG	24		CCG	21		CAG	34		CGG	8
Ile	ATT	45	Thr	ACT	40	Asn	AAT	37	Ser	AGT	9
	ATC	12		ACC	18		AAC	25		AGC	5
	ATA	2		ACA	13	Lys	AAA	47	Arg	AGA	6
Met	ATG	42		ACG	19		AAG	31		AGG	1
Val	GTT	53	Ala	GCT	64	Asp	GAT	44	Gly	GGT	38
	GTC	14		GCC	17		GAC	35		GGC	28
	GTA	10		GCA	12	Glu	GAA	27		GGA	13
	GTG	11		GCG	12		GAG	34		GGG	3

The totals are derived from sequences in Fig. 1 which are fully confirmed, that is, 377 codons in gene *A*, 120 in gene *B*, 152 in gene *D*, 91 in gene *E*, 38 in gene *J*, 344 in gene *F*, 175 in gene *G* and 49 in gene *H*. Out of a total of 1,346 codons 42.9% terminate in T. The percentages in the different genes are: *A*, 37.1 (non-overlapping region 47.1; overlapping region 15.8); *B*, 34.2; *D*, 42.1; *E*, 14.3; *J*, 47.4; *F*, 52.0; *G*, 54.3; *H*, 49.0. The initiating ATG is included in all cases.

The use of codons in ΦX174

Table 4 shows the codons used in regions where the nucleotide sequence is fully confirmed. It is clear that the pattern established by early observations on non-random use of codons[23,24] is continued now that more information is available. In particular, the preference for T at the third position of the codon is marked throughout the genome, as shown in Table 4. In regions of overlapping genes, one of the pair tends to continue the 'third T' trend (*D* and *B*), thus excluding the other (*E* and *A*). This may give some indication of the order in which overlapping genes evolved[33,50]. Another interesting feature is the very low occurrence of codons starting AG, particularly in non-overlapping regions. The base composition of the sequence of ΦX174 DNA shown in Fig. 1 is: *A*, 23.9%; *C*, 21.5%; *G*, 23.3% and *T*, 31.2%. This is in good agreement with previously determined values (see ref. 4).

We thank D. McCallum and R. Staden for carrying out the computer data storage and analysis of the sequence.

Note added in proof: J. E. Sims and D. Dressler (personal communication) have independently determined the sequence in positions 263–375 and 4,801–4,940. Their results agree with those given in Fig. 1. They have also identified the 'B' mRNA start as being at position 4,888.

Received November 30; accepted December 24, 1976.

1 Sanger, F. & Coulson, A. R. *J. molec. Biol.* **94**, 441–448 (1975).
2 Benbow, R. M., Hutchison, C. A. III, Fabricant, J. D. & Sinsheimer, R. L. *J. Virol.* **7**, 549–558 (1971).
3 Benbow, R. M., Zuccarelli, A. J., Davis, G. C. & Sinshiemer, R. L. *J. Virol.* **13**, 898–907 (1974).
4 Denhardt, D. T. *CRC Crit. Rev. Microbiol.* **4**, 161–222 (1975).
5 Hall, J. B. & Sinsheimer, R. L. *J. molec. Biol.* **6**, 115–127 (1963).
6 Ling, V. *Proc. natn. Acad. Sci. U.S.A.* **69**, 742–746 (1972).
7 Harbers, B., Delaney, A. D., Harbers, K. & Spencer, J. H. *Biochemistry* **15**, 407–414 (1976).
8 Burton, K. & Petersen, G. B. *Biochem. J.* **75**, 17–27 (1960).
9 Chadwell, H. A. Thesis, University of Cambridge (1974).
10 Sadowski, P. D. & Bakyta, I. *J. biol. Chem.* **247**, 405–412 (1972).
11 Ling, V. *FEBS Lett.* **19**, 50–54 (1971).
12 Ziff, E. B., Sedat, J. W. & Galibert, F. *Nature new Biol.* **241**, 34–37 (1973).
13 Galibert, F., Sedat, J. W. & Ziff, E. B. *J. molec. Biol.* **87**, 377–407 (1974).
14 Robertson, H. D., Barrell, B. G., Weith, H. L. & Donelson, J. E. *Nature new Biol.* **241**, 38–40 (1973).
15 Air, G. M. & Bridgen, J. *Nature new Biol.* **241**, 40–41 (1973).
16 Sanger, F., Donelson, J. E., Coulson, A. R., Kössel, H. & Fischer, D. *Proc. natn. Acad. Sci. U.S.A.* **70**, 1209–1213 (1973).
17 Schott, H. *Makromolek. Chem.* **175**, 1683–1693 (1974).
18 Donelson, J. E., Barrell, B. G., Weith, H. L., Kössel, H. & Schott, H. *Eur. J. Biochem.* **58**, 383–395 (1975).
19 Blackburn, E. H. *J. molec. Biol.* **93**, 367–374 (1975).
20 Blackburn, E. H. *J. molec. Biol.* **107**, 417–432 (1976).
21 Sedat, J. W., Ziff, E. B. & Galibert, F. *J. molec. Biol.* **107**, 391–416 (1976).
22 Air, G. M. *J. molec. Biol.* **107**, 433–444 (1976).
23 Air, G. M. *et al. J. molec. Biol.* **107**, 445–458 (1976).
24 Air, G. M., Blackburn, E. H., Sanger, F. & Coulson, A. R. *J. molec. Biol.* **96**, 703–719 (1975).
25 Lee, A. S. & Sinsheimer, R. L. *Proc. natn. Acad. Sci. U.S.A.* **71**, 2882–2886 (1974).
26 Hayashi, M. N. & Hayashi, M. *J. Virol.* **14**, 1142–1152 (1974).
27 Vereijken, J. M., van Mansfeld, A. D. M., Baas, P. D. & Jansz, H. S. *Virology* **68**, 221–233 (1975).
28 Jeppesen, P. G. N., Sanders, L. & Slocombe, P. M. *Nucl. Acids Res.* **3**, 1323–1339 (1976).
29 Sato, S., Hutchison, C. A. III & Harris, J. I. *Proc. natn. Acad. Sci. U.S.A.* (in the press).
30 Brown, N. L. & Smith, M. *FEBS Lett.* **65**, 284–287 (1976).
31 Fiddes, J. C. *J. molec. Biol.* **107**, 1–24 (1976).
32 Air, G. M., Sanger, F. & Coulson, A. R. *J. molec. Biol.* **108**, 519–533 (1976).
33 Barrell, B. G., Air, G. M. & Hutchison, C. A. III *Nature* **264**, 34–41 (1976).
34 Smith, L. H. & Sinsheimer, R. L. *J. molec. Biol.* **103**, 699–735 (1976).
35 Grohmann, K., Smith, L. H. & Sinsheimer, R. L. *Biochemistry* **14**, 1951–1955 (1975).
36 Axelrod, N. *J. molec. Biol.* **108**, 753–779 (1976).
37 Hayashi, M., Fujimura, F. K. & Hayashi, M. *Proc. natn. Acad. Sci. U.S.A.* **73**, 3519–3523 (1976).
38 Pribnow, D. *Proc. natn. Acad. Sci. U.S.A.* **72**, 784–788 (1975).
39 Rosenberg, M., de Crombrugghe, B. & Musso, R. *Proc. natn. Acad. Sci. U.S.A.* **73**, 717–721 (1976).
40 Bertrand, K. *et al. Science* **189**, 22–26 (1975).
41 Sugimoto, K., Sugisaki, H., Okamoto, T. & Takanami, M. *J. molec. Biol.* (in the press).
42 Shine, J. & Dalgarno, L. *Proc. natn. Acad. Sci. U.S.A.* **71**, 1342–1346 (1974).
43 Steitz, J. A. & Jakes, K. *Proc. natn. Acad. Sci. U.S.A.* **72**, 4734–4738 (1975).
44 Henry, T. J. & Knippers, R. *Proc. natn. Acad. Sci. U.S.A.* **71**, 1549–1553 (1974).
45 Linney, E. & Hayashi, M. *Nature* **249**, 345–348 (1974).
46 Baas, P. D., Jansz, H. S. & Sinsheimer, R. L. *J. molec. Biol.* **102**, 633–656 (1976).
47 Eisenberg, S., Harbers, B., Hours, C. & Denhardt, D. T. *J. molec. Biol.* **99**, 107–123 (1975).
48 Subramanian, K. N., Dhar, R. & Weissman, S. M. *J. biol. Chem.* (in the press).
49 Ravetch, J. V., Model, P. & Robertson, H. D. *Nature* **265**, 698–702 (1977).
50 Smith, M. *et al.* (submitted to Nature).
51 Funk, F. & Sinsheimer, R. L. *J. Virol.* **6**, 12–19 (1970).
52 Baas, P. D., van Heusden, G. P. H., Vereijken, J. M., Weisbeek, P. J. & Jansz H. S. *Nucl. Acids Res.* **3**, 1947–1960 (1976).
53 Platt, T. & Yanofsky, C. *Proc. natn. Acad. Sci. U.S.A.* **72**, 2399–2403 (1975).
54 Weisbeek, P. J., Vereijken, J. M., Baas, P. D., Jansz, H. S. & Van Arkel, G. A. *Virology* **72**, 61–71 (1976).
55 Jazwinski, S. M., Lindberg, A. A. & Kornberg, A. *Virology* **66**, 283–293 (1975).
56 Kornberg, A. *DNA Synthesis* (W. H. Freeman, San Francisco, 1974).
57 Chen, C. Y., Hutchison, C. A. III & Edgell, M. H. *Nature new Biol.* **243** 233–236 (1973).
58 van Mansfeld, A. D. M., Vereijken, J. M. & Jansz, H. S. *Nucl. Acids Res.* **3**, 2827–2843 (1976).

DNA sequence of a region of the ΦX174 genome coding for a ribosome binding site

Nigel L. Brown* & Michael Smith†

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

The DNA region corresponding to a newly identified ribosome binding site in ΦX174 DNA is sequenced and mapped in relation to the physical map of ΦX174. Assignments of the binding site to a specific gene are discussed, and the possibility of a second case of overlapping genes in ΦX174 is considered.

RECENT studies on the DNA of bacteriophage ΦX174 have defined sequences corresponding to the ribosome binding sites of genes *D*, *J*, *F* and *G* (refs 1–3 and B. G. Barrell, personal communication). A ribosome binding site that does not correspond to any of the above sequences has now been defined by studies on a ribosome-protected fragment of a transcript *in vitro* of ΦX174 RF DNA (ref. 4). Inspection of preliminary DNA sequence data obtained from ΦX174 (F. Sanger, *et al.*, unpublished) for potential ribosome binding sites, and comparison of the sequence data with T_1-oligonucleotides from the unassigned RNA fragment[4], allowed identification of the DNA sequence corresponding to this RNA fragment. We report here the complete DNA sequence of a region of ΦX174 *am*3 RF DNA that corresponds to this new ribosome binding site. Because the DNA sequence contains accurately mapped cleavage sites for restriction endonucleases, the ribosome binding site can be located precisely on the physical map of the ΦX genome[5], and therefore located on the genetic map[6–8].

Present addresses: *Department of Biochemistry, University of Bristol, Bristol BS8 1TD, UK. †Department of Biochemistry, University of British Columbia, Vancouver BC, Canada V6T 1W5.

Determination of the DNA sequence

The ribosome binding site lies in that portion of *Hin*dII fragment 5 that overlaps *Hin*fI fragment 5a (nucleotides 5,000–5,100 approx, Fig. 1). The DNA sequence of this region was determined using the 'plus–minus' technique of Sanger and Coulson[9] (Fig. 2). It was quantitatively confirmed by characterisation of the pyrimidine oligonucleotides obtained by primed synthesis *in vitro* of DNA labelled with α-^{32}P-dATP or α-^{32}P-dGTP using viral or complementary strand as template[10,11] (Fig. 3, Table 1).

The sequence between the *Hin*dII 8/5 cleavage site and the *Hin*fI 5a/6 cleavage site contains 111 nucleotides. It also contains the *Alu*I cleavage sites between fragments 12b, 15b, 17 and 12a (Fig. 1). The sequence determination allowed the