Proc. R. Soc. Lond. B. **191**, 317–333 (1975)
Printed in Great Britain

The Croonian Lecture, 1975

Nucleotide sequences in DNA

By F. Sanger, F.R.S.

Medical Research Council Laboratory of Molecular Biology, Cambridge

(*Delivered* 15 *May* 1975 – *Received* 15 *May* 1975)

[Plates 22–25]

DNA, the chemical component of the gene, plays a central role in biology and contains the whole information for the development of an organism, coded in the form of sequences of the four nucleotide residues. The lecture describes the development and application of some methods that can be employed to deduce sequences in these very large molecules. Special attention has been applied to a rapid simple method in which DNA polymerase is primed with specific oligonucleotide primers, thus making it possible to study small sections of radioactively labelled DNA.

The techniques have been applied to the single-stranded DNA of bacteriophage ϕX 174, and two sequences of about 250 nucleotides long have been deduced and related to the amino acid sequences of the proteins for which they code.

Introduction

Although the whole of the properties of living matter are controlled by the unique sequences of nucleotide residues in the DNA of the genes, the study of these sequences has proved a formidable task and it is only recently that some methods have become available. Interest in this field has therefore centred largely on the development of techniques rather than on the interpretation of results (which are still relatively few), and in this lecture I shall describe some of the work in this direction from our laboratory.

Earlier work on nucleic acid sequencing had been done with RNA and we had developed a number of techniques that could be applied to relatively small RNA molecules (Sanger, Brownlee & Barrell 1965; Brownlee & Sanger 1969). In this work special attention was given to the development of rapid and simple fractionation techniques using ionophoresis and chromatography on modified papers and thin-layer systems. As such methods can only be carried out efficiently on a small scale, RNA labelled with ^{32}P was used as being a highly sensitive method for the detection and estimation of the nucleotides. The methods were initially developed using small RNA molecules (transfer RNAs and the 5*S* RNA (Brownlee, Sanger & Barrell 1968)). With further refinements it was shown that they could be applied to a bacteriophage RNA containing about 3000 residues, thus making possible the

study of nucleic acid containing genetic information (Adams, Jeppesen, Sanger & Barrell 1969; Sanger 1971; Min Jou, Haegeman, Ysebaert & Fiers 1972).

The main obstacle to the extension of these techniques to DNA was the large size of the simplest DNA molecules. Whereas it had been possible to use small RNAs of 75–120 residues in the development of methods for RNA sequencing, the smallest suitable DNA molecules are the small bacteriophage and viral DNAs, which contain over 5000 residues. Most of our work has been done with the single-stranded circular DNA of bacteriophage ϕX 174 because much is already known

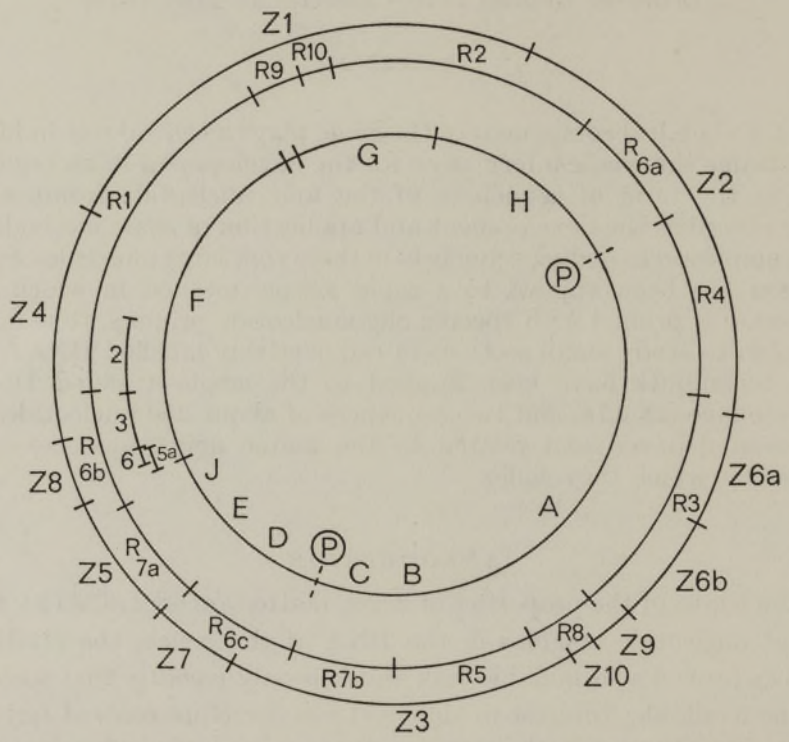

FIGURE 1. Diagram showing the genetic map (inner circle) of ϕX DNA and the cleavage map produced by the restriction enzymes R. *Hae* (Z) and R. *Hin*$_d$ (R), and by endonuclease IV (2, 3, 6 and 5a). The gene lengths correspond to the molecular masses of the F and G products. Otherwise the exact lengths are uncertain as there is some discrepancy between the sizes of the proteins and of the DNA. (P) marks the probable position of promoter sites. (Chen, Hutchison & Edgell 1973.)

about its genetic properties, largely from the work of Sinsheimer and his group (Sinsheimer 1968; Benbow, Zuccarelli, Davis & Sinsheimer 1974) and also as work was in progress on the amino acid sequences of some of the proteins coded by the DNA.

Figure 1 shows the genetic map of ϕX DNA. There are nine genes corresponding to nine proteins. Genes F and G, which are the only two that will be considered in this lecture, correspond to two of the component structural proteins of the virus and their amino acid sequence is being studied by Air (Air & Bridgen 1973).

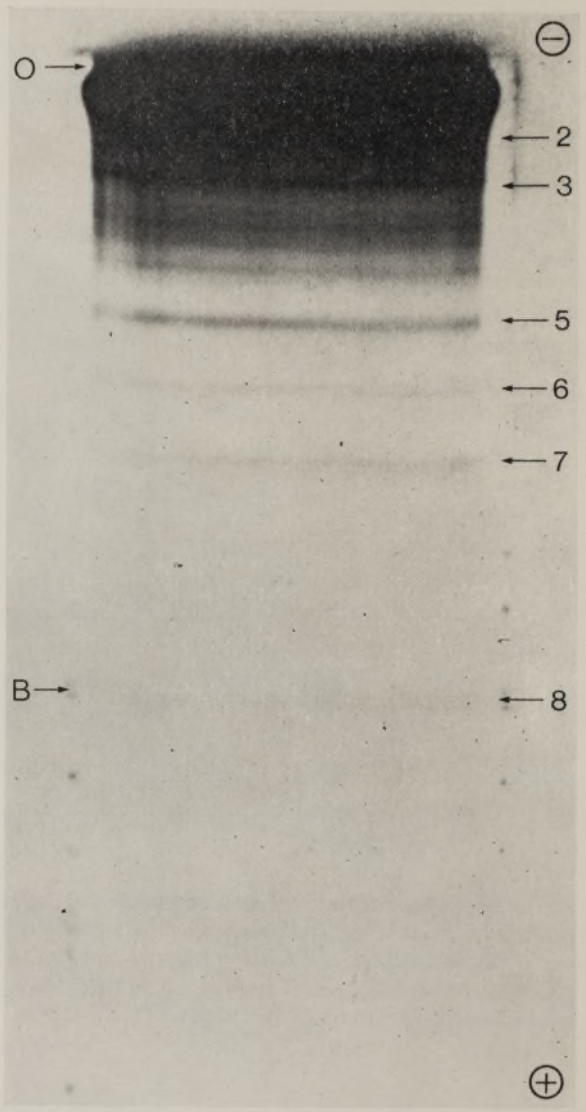

FIGURE 2. Radioautograph of an acrylamide gel ionophoretic fractionation of a partial endonuclease IV digest of ϕX DNA (Ziff *et al.* 1973).

(*Facing p.* 319)

Direct sequencing of DNA by partial degradation

The most widely used approach to sequencing both for RNA and proteins, is the method of partial degradation, and this principle was first applied to studies on ^{32}P-labelled ϕX DNA. Degradation is usually carried out enzymically and very much depends on the availability of suitable enzymes. That most widely used for RNA is ribonuclease T_1, which splits specifically at rGp† residues. This unique specificity made it particularly suitable for both complete and partial digestion. Although many deoxyribonucleases (DNAase) are known, there is no suitable one having a unique specificity for one residue, and it was thought that this would make DNA sequencing particularly difficult. However there are in fact DNAases of greater specificity and these are proving very useful.

One such enzyme is endonuclease IV (from bacteriophage T4), originally isolated by Sadowski & Hurwits (1969), who showed that it would split specifically on the 5′ side of C residues. However it does not split all such residues and gives oligonucleotides, on the average, of between 10 and 20 residues. 'Limit' digests of ϕX DNA were extremely complex, but Ziff, Sedat & Galibert (1973) (and Galibert, Sedat & Ziff 1974) showed that if ^{32}P-labelled ϕX DNA was subjected to partial digestion and the products fractionated by ionophoresis on acrylamide gel a number of pure oligonucleotides could be isolated (figure 2, plate 22). One of these (band 6) was about 50 residues long and this was a suitable small piece of DNA on which to work out methods for detailed sequencing.

Another ^{32}P-labelled fragment of about 50 residues from ϕX was isolated by Robertson, Barrell, Weith & Donelson (1973) as a ribosomal binding site. Ribosomes will normally bind to the initiation sites for protein chains on a messenger RNA and this principle has been used to study such sites in viral RNAs (Hindley & Staples 1969; Steitz 1969). ϕX DNA, being single-stranded, is presumably not unlike a messenger RNA and under suitable conditions a complex is formed. After digestion of the complex with pancreatic DNAase the protected fragment could be isolated.

With these two small fragments of DNA it was possible to develop methods for detailed sequencing. These methods involved in particular the further use of endonuclease IV and partial digestion of the smaller products obtained with exonucleases (Galibert *et al.* 1974; Robertson *et al.* 1973).

J. W. Sedat, E. B. Ziff & F. Galibert (personal communication) have extended these studies to some of the larger products obtained from the partial endonuclease IV digest (figure 2). It was found that all of the fragments studied originated from the same part of the ϕX DNA and their relative order was shown to be as in figure 1. Endonuclease IV is a single-strand specific nuclease and presumably this part of the molecule has less folding than the remainder of the DNA, which remains

† Abbreviations: As this lecture is predominantly concerned with DNA, the symbols C, T, G and A will be used for the deoxyribonucleosides, and rC, rU, rG and rA for the ribonucleosides.

largely intact under the conditions of partial digestion. The complete sequence of bands 6, 5 and 5A, the almost complete sequence of band 3, and partial data on band 2 were determined by this method. Concurrent with this work G. M. Air (personal communication) was studying the amino acid sequence of the viral coat protein corresponding to gene F and from the sequences involved it became clear that the DNA in these fragments was coding for the amino acid sequence near to the *N*-terminus of this protein (see figure 1).

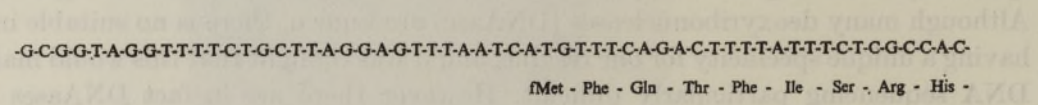

FIGURE 3. Nucleotide sequence of a ribosomal binding site in ϕX DNA (Robertson *et al.* 1973) and *N*-terminal amino acid sequence of the gene G protein (Air & Bridgen 1973).

Figure 3 shows the nucleotide sequence for the ribosomal binding site determined by Robertson *et al.* (1973). This sequence contains the triplet A T G, which could code for the initiating formylmethionine tRNA, and, from the nucleotide sequence following and the genetic code, an amino acid sequence could be predicted for the *N*-terminus of a protein chain initiating at this site. Air & Bridgen (1973) were studying the sequences of proteins from ϕX and found that this *N*-terminal sequence did in fact occur in the protein of the gene G. The function of this piece of DNA and its position on the genetic map were thus defined.

Restriction enzymes

A group of enzymes that are proving extremely useful for studies on DNA are the restriction enzymes. These are extremely specific, recognizing a sequence of 4-6 nucleotide residues in double-stranded DNA. Thus one of these enzymes (*Hin*$_d$ II) from *Haemophilus influenzae* recognizes the sequence G-T-Y-R-A-C- (where Y is a pyrimidine and R a purine nucleotide) and splits the replicative form of ϕX into 13 well-defined fragments (Edgell, Hutchison & Sclair 1972) which can be purified by ionophoresis on acrylamide. The relative order of these fragments and their positions on the genetic map have been determined and are shown in figure 1. There are a number of other such enzymes with different specificities and these are proving very useful for the defined degradation of DNA, and therefore for sequence analysis.

SEQUENCING DNA BY COPYING METHODS

One of the main difficulties in using the direct approach to DNA sequencing by partial degradation procedures is the large amounts of ^{32}P that have to be used to obtain sufficient labelled DNA. With ϕX DNA this was possible, but with larger DNA molecules it would probably be impracticable. One way of overcoming the problem is to use unlabelled DNA and to copy it with RNA polymerase or DNA

polymerase, using ^{32}P-labelled triphosphates. The potential of copying procedures for nucleotide sequences was first demonstrated by C. Weissmann and his colleagues (Billeter *et al.* 1969) in their ingenious studies on bacteriophage Qβ RNA using the virus's own replicase. Not only is this an effective way of preparing radioactive nucleic acid, but, by using a situation in which copying starts at a unique site, it is possible to use the 'pulse-labelling' technique and thus prepare defined stretches of labelled nucleic acid of various lengths, which greatly helps in the sequencing.

Copying ϕX DNA with RNA polymerase

DNA may be copied with RNA polymerase or with DNA polymerase. In the former case ^{32}P-labelled RNA is produced and methods were already available for studying RNA sequences. This approach was applied to ϕX DNA by Blackburn (1975). Initially the whole single-stranded ϕX was copied and the RNA product digested with ribonuclease T_1. The large products, of about 15–25 residues in length, could be isolated and sequenced. While this approach gave certain information about sequences, it was not at first possible to identify the position of the fragments on the genetic map or their biological significance. More meaningful results were obtained by copying defined fragments of the ϕX. Thus bands 6 and 3 from the endonuclease IV partial digest were studied in this way (Blackburn, 1975, and personal communication). In the former case the results confirmed, and in the latter confirmed and supplemented, the sequences determined by direct partial degradation of the ^{32}P-labelled DNA. Fragments from restriction enzyme digests of ϕX replicative form could also be studied in this way. In this case copies of both strands were obtained. Although this increased the complexity of the digests it also helped in the deduction of the sequence by 'overlapping' corresponding products from the two strands. Much of the sequence of fragment Hin_d^{10} (see figure 1) was deduced with RNA polymerase (Air *et al.* 1975).

Copying DNA with DNA polymerase

Much of our own recent work has been concerned with the development of methods for DNA sequencing with DNA polymerase. As shown below, the approach has several advantages – particularly the possiblity of using pulse-labelling techniques. For these it is necessary to devise a situation in which copying starts at one unique site of the DNA molecule. Figure 4 shows the specificity requirements for DNA polymerase action. Triphosphates are added sequentially to the unique 3′ terminus of the primer making the complementary copy of the template. Wu & Kaiser (1968) were the first to use this approach to DNA sequencing in their studies on the 'sticky ends' of bacteriophage λ DNA. These 'sticky ends' are formed by a single-stranded extension of 12 nucleotide residues from the otherwise double-stranded molecule, thus providing a substrate for DNA polymerase as

illustrated in figure 4. By using radioactive triphosphates it was possible to label uniquely the complementary sequence for these sticky ends and to determine their sequences (Wu & Taylor 1971).

Our own initial studies in sequencing with DNA polymerase made use of synthetic oligonucleotides to provide the unique 3′ terminus for chain extension. Such an oligonucleotide had to bind specifically to a unique site on the template

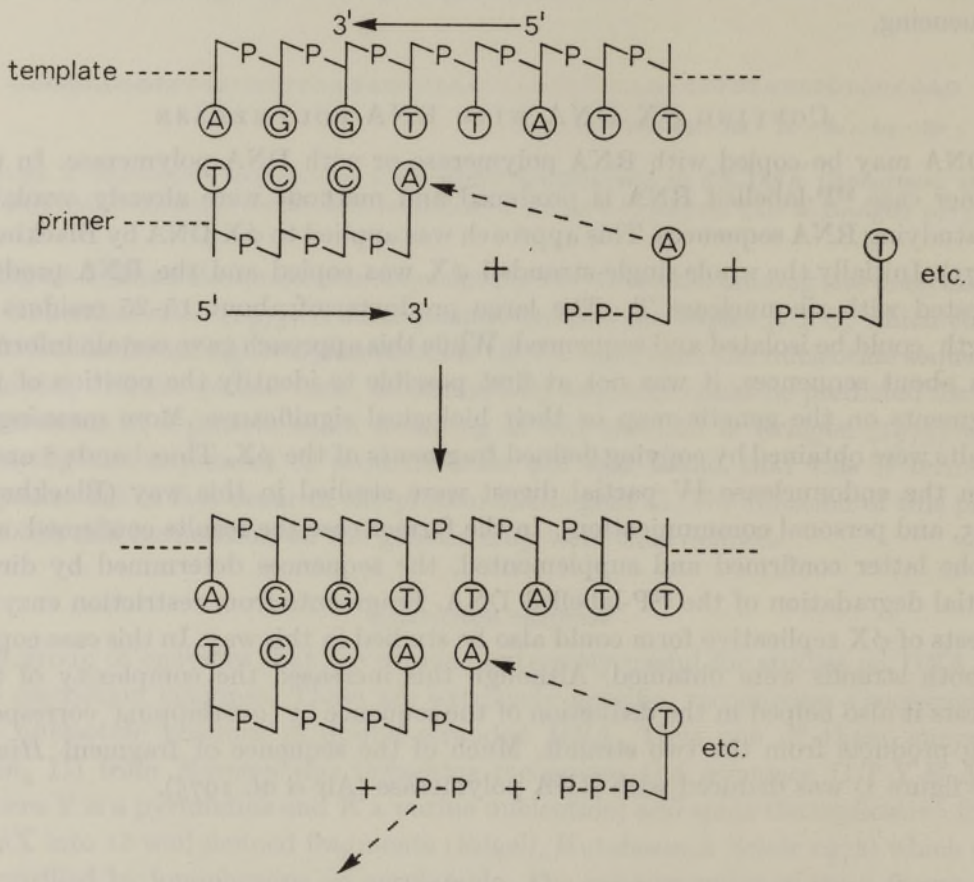

FIGURE 4. Diagram summarizing the specificity requirements for DNA polymerase.

DNA by hybridization. Two oligonucleotides – an octamer and a decamer – have been synthesized by Kössel and his co-workers (Schott, Fischer & Kössel 1973; Schott 1974) and used on ϕX DNA to develop the techniques. The decamer was designed to initiate copying in the ribosomal binding site sequence discussed above. This site contained the sequence -T-T-T-T-A-T-T-T-C-T- (figure 3), and the decamer synthesized had the complementary sequence A-G-A-A-A-T-A-A-A-A. When added in considerable excess it hybridized sufficiently to the template to act as a primer for the synthesis of specific lengths of radioactive DNA having a sequence complementary to a known position in the ϕX template. Two main approaches have been used in deducing sequences copied in this way.

The ribosubstitution method

Berg, Fancher & Chamberlin (1963) showed that if polymerization with DNA polymerase is carried out in the presence of Mn (replacing the normally used Mg) ribonucleotides can be incorporated into the DNA chain. Thus, by using rCTP and the other three nucleotides as deoxyribotriphosphates, a DNA chain is formed in which all the C residues are in the ribo form. The advantage of this method was that these ribo bonds could now be split specifically by using alkali or a suitable ribonuclease. One therefore has specific methods for degrading the products, which greatly facilitates sequence analysis. Although this approach was proposed over ten years ago it is only recently that suitable systems have become available to demonstrate its practicability (Van de Sande, Loewen & Khorana 1972; Salser, Fry, Brunk & Poon 1972; Sanger *et al.* 1973).

In our experience, both rCTP and rGTP could be used for the ribosubstitution in separate incubations, so that two sets of oligonucleotides could be obtained and this facilitated sequence deduction by 'over-lapping'. Figure 5, plate 23 shows an example of how this method was used to confirm and extend the sequence at the ribosomal binding site (J. E. Donelson & B. G. Barrell, personal communication). The decanucleotide primer and ϕX DNA were incubated with DNA polymerase in the presence of rCTP and the three other deoxytriphosphates (one of which was ^{32}P-labelled) in such a way that the length of the newly formed chains was 20–50 nucleotides. If this product was then digested with pancreatic ribonuclease and the products fractionated by a two-dimensional technique, a small number of fragments were obtained as in figure 5, VI. These could be eluted and their sequences determined by further degradation procedures. In order to deduce the relative order of these fragments a sample of the incubation mixture was fractionated before hydrolysis (figure 5, I). These products all contained the decanucleotide sequence and extended to various lengths. They also contained rC residues. They were considerably larger than the products of hydrolysis so that an appropriately modified fractionation system was employed. Each spot (figure 5, I, *a-e*) was eluted, hydrolised and the products fractionated (figure 5, II, 5, V). From the gradually increasing complexity of these fingerprints the relative order of the digestion products was shown to be c1, c2, c3, c4, c5. From experiments such as this it was possible to deduce a sequence of 41 residues extending from the decamer primer. One important conclusion from this work was that the method did work and that the DNA polymerase was copying faithfully in spite of the somewhat 'unphysiological' conditions used.

This method was originally worked out by using an octanucleotide primer on bacteriophage f1 DNA and a sequence of about 80 residues was deduced (Sanger *et al.* 1973). More recently the sequence in a repressor binding site of bacteriophage λ has been determined (Maniatis, Ptashne, Barrell & Donelson 1974).

The 'plus and minus' method

One of the main problems about nucleotide sequencing, which becomes increasingly forbidding as larger molecules are studied, is the amount of work that is required to obtain and establish results. Thus in the partial degradation procedure an initial partial digestion will give many fragments each of which may be eluted, redigested and fractionated, yielding more fragments whose sequence is then to be determined by further digestions. One problem that frequently arises is that because of the continual proliferation of the products one may end up with insufficient amounts of the small oligonucleotides to carry out the final analyses. For this reason we have paid considerable attention to simplifying, and if possible short-circuiting, procedures; in particular to the development of methods in which sequence information can be obtained directly from fractionation procedures, and so avoiding a final analytical step. In general this is more easily accomplished with nucleic acids than with proteins because the former have fewer different residues. For instance it was shown that the composition of products of ribonuclease T_1 suggests of RNA could frequently be deduced from their position on the fingerprint obtained using a two-dimensional fractionation procedure (Sanger *et al.* 1965). Another technique of this type is a modification of the partial exonuclease digestion method introduced by Ling (1972) to deduce the sequences of depurination products of DNA. In this method the partial exonuclease digest is fractionated on a two-dimensional system in which fractionation in one dimension depends on size and in the other on composition. From the pattern of spots obtained it is possible to deduce the sequence, as loss of a C residue from a nucleotide will lead to displacement in one direction and loss of a T residue in the other. The method has now been extended to include oligonucleotides containing all four mononucleotides (Rensing & Schoenmakers 1973; Ziff *et al.* 1973; Bambara, Jay & Wu 1974; Maniatis *et al.* 1974). Although confirmation of such results is frequently necessary, the method has proved extremely useful.

Another such method that promises to accelerate DNA sequence determination is the 'plus and minus' method (Sanger & Coulson 1975). It involves the primed synthesis with DNA polymerase as discussed above. In studies with the oligo-

---

DESCRIPTION OF PLATE 23

FIGURE 5. Determination of a sequence in the ribosomal binding site of ϕX DNA using primed synthesis with the decanucleotide A-G-A-A-A-T-A-A-A-A and the ribosubstitution technique (J. E. Donelson & B. G. Barrell, personal communication). I, Radioautograph of the two-dimensional fractionation of the product obtained by the limited action of DNA polymerase on ϕX DNA in the presence of the primer and rCTP. The spots marked A, B, C, D and E were eluted, digested with pancreatic ribonuclease, and the products fractionated on the two-dimensional system. Radioautographs of these fractionations are shown in II, III, IV, V and VI respectively. The sequences of the oligonucleotides C1–C5 are shown in VI. For the two-dimensional fractions the first dimension was by ionophoresis at pH 3.5 on cellulose acetate and the second by homochromatography (Brownlee & Sanger 1969). (A 7% 'homomix' was used in I and a 3% in the other fractionations.)

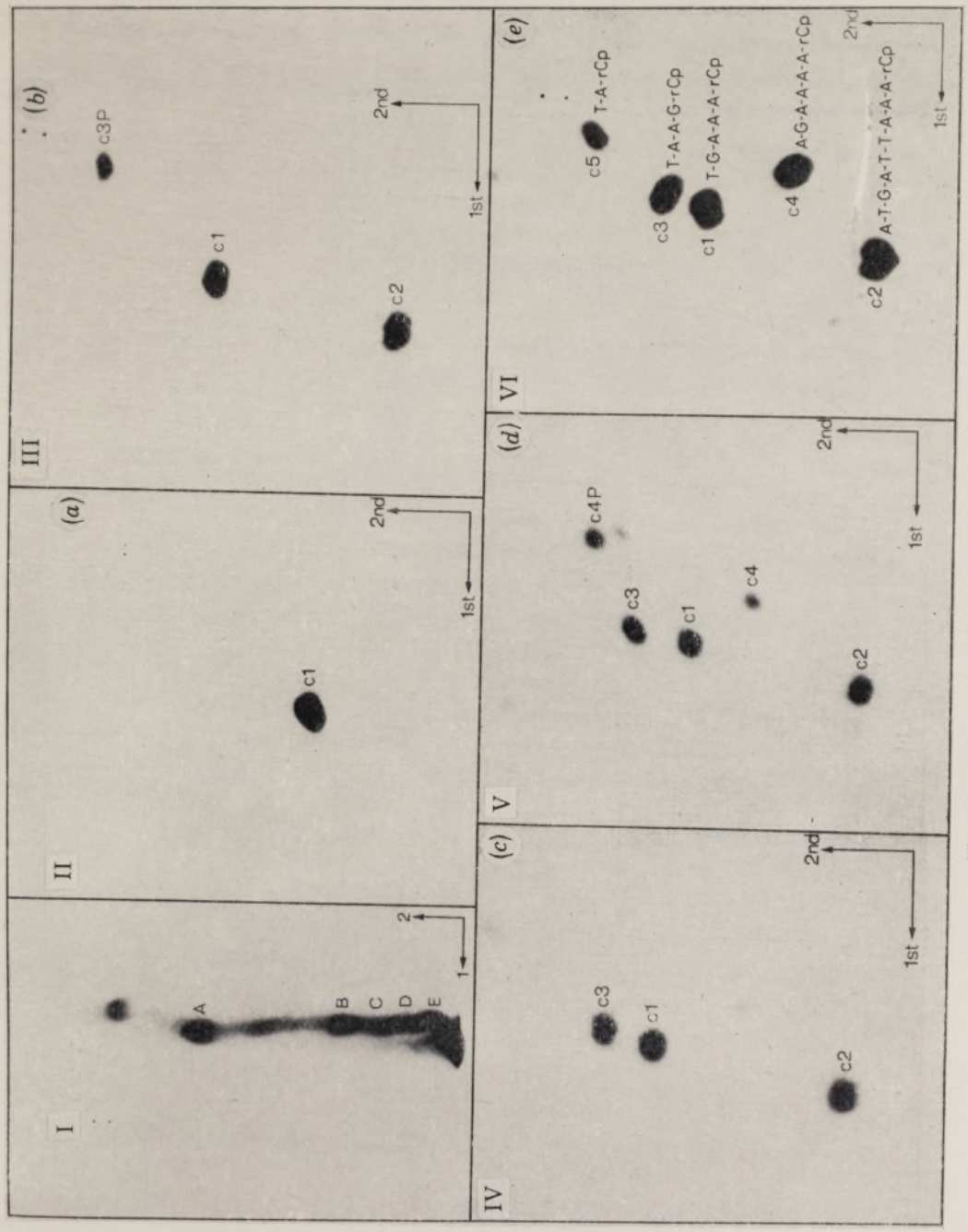

FIGURE 5. For description see facing page.

(*Facing p.* 324)

nucleotide primers it was frequently observed that when incubation was carried out with a low concentration of a radioactive triphosphate (for instance ^{32}P-dATP) and the products fractionated by ionophoresis on acrylamide gel, a series of well-defined bands was obtained in which elongation by the polymerase had stopped

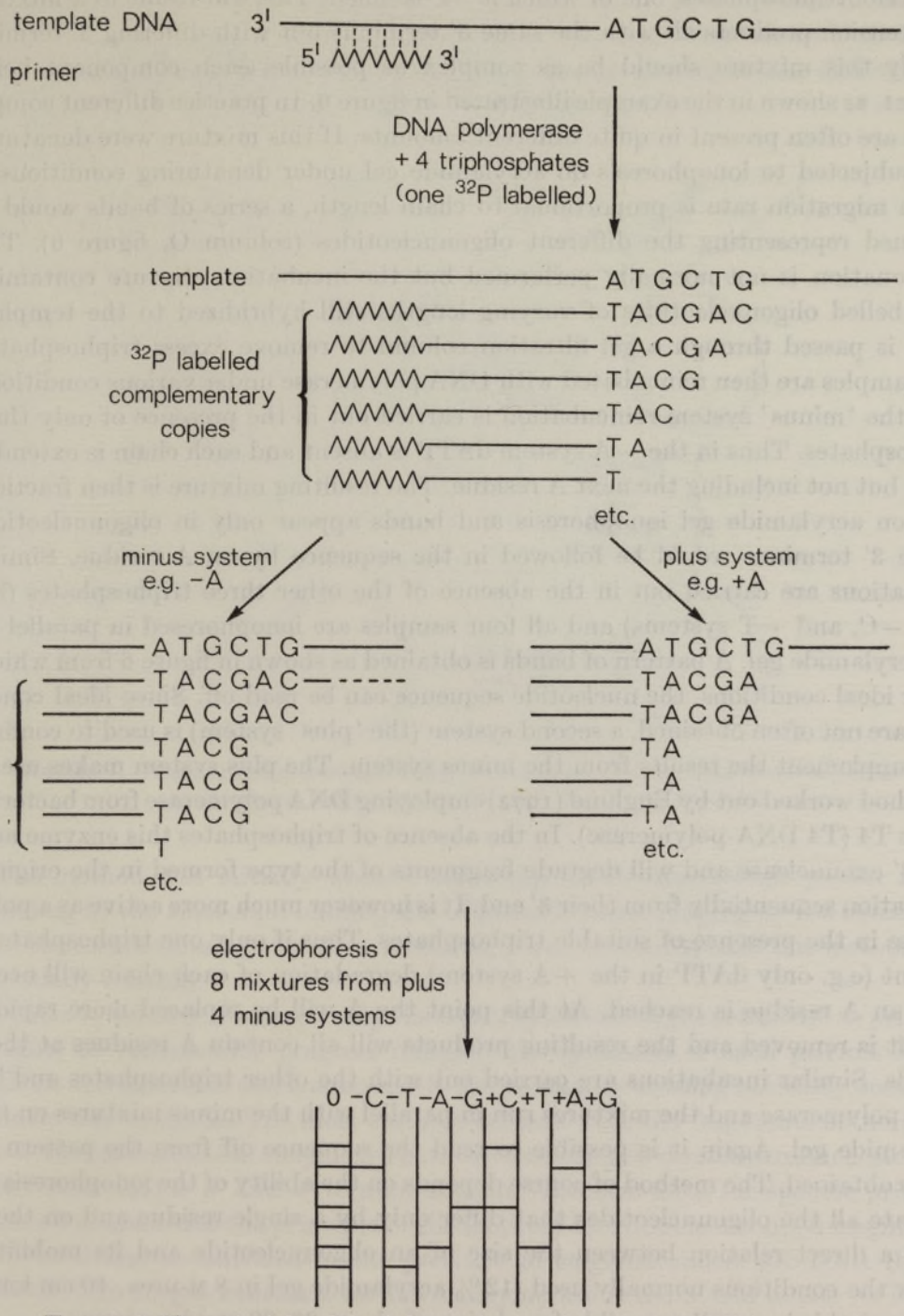

FIGURE 6. The principle of the 'plus and minus' method for DNA sequencing.

before the A residues. This suggested the possibility of determining the relative distribution of residues from the patterns of bands obtained by such fractionations.

The general principle of the method is shown in figure 6. DNA polymerase is first allowed to extend the primer for a limited time in the presence of all four deoxyribotriphosphates, one of which is ^{32}P-labelled. This will result in a mixture of extension products all with the same 5′ terminus but with differing 3′ termini. Ideally this mixture should be as complex as possible, each component being present, as shown in the example illustrated in figure 6. In practice different components are often present in quite different amounts. If this mixture were denatured and subjected to ionophoresis on acrylamide gel under denaturing conditions in which migration rate is proportional to chain length, a series of bands would be obtained representing the different oligonucleotides (column O, figure 6). This fractionation is not normally performed but the incubation mixture containing ^{32}P-labelled oligonucleotides of varying lengths still hybridized to the template DNA is passed through a gel filtration column to remove excess triphosphates, and samples are then reincubated with DNA polymerase under various conditions.

In the 'minus' system reincubation is carried out in the presence of only three triphosphates. Thus in the −A system dATP is absent and each chain is extended up to but not including the next A residue. The resulting mixture is then fractionated on acrylamide gel ionophoresis and bands appear only in oligonucleotides whose 3′ terminus would be followed in the sequence by an A residue. Similar incubations are carried out in the absence of the other three triphosphates (the −G, −C, and −T systems) and all four samples are ionophoresed in parallel on the acrylamide gel. A pattern of bands is obtained as shown in figure 6 from which, under ideal conditions, the nucleotide sequence can be read off. Since ideal conditions are not often obtained, a second system (the 'plus' system) is used to confirm and supplement the results from the minus system. The plus system makes use of a method worked out by Englund (1972) employing DNA polymerase from bacteriophage T4 (T4 DNA-polymerase). In the absence of triphosphates this enzyme acts as a 3′ exonuclease and will degrade fragments of the type formed in the original incubation sequentially from their 3′ end. It is however much more active as a polymerase in the presence of suitable triphosphates. Thus if only one triphosphate is present (e.g. only dATP in the +A system) degradation of each chain will occur until an A residue is reached. At this point the A will be replaced more rapidly than it is removed and the resulting products will all contain A residues at their 3′ ends. Similar incubations are carried out with the other triphosphates and T4 DNA polymerase and the mixtures run in parallel with the minus mixtures on the acrylamide gel. Again it is possible to read the sequence off from the pattern of bands obtained. The method of course depends on the ability of the ionophoresis to separate all the oligonucleotides that differ only by a single residue and on there being a direct relation between the size of an oligonucleotide and its mobility. Under the conditions normally used (12% acrylamide gel in 8 M-urea, 40 cm long, 20 V/cm) this is usually possible for chains of about 20–80 residues long.

From the example illustrated in figure 6 it may be seen that each oligonucleotide is represented by two bands – one in the plus system, which defines the 3′ terminus, and one in the minus system, which defines the next residue in the sequence. Thus in the example given in figure 6, in the position of the smallest product there is a band in the −T column and in that of the next there are two bands in the +T and −A columns. This defines a dinucleotide sequence T-A where T is the 3′ terminus of this oligonucleotide. Similarly the next product gives the dinucleotide A-C which overlaps with the T-A, giving the sequence T-A-C.

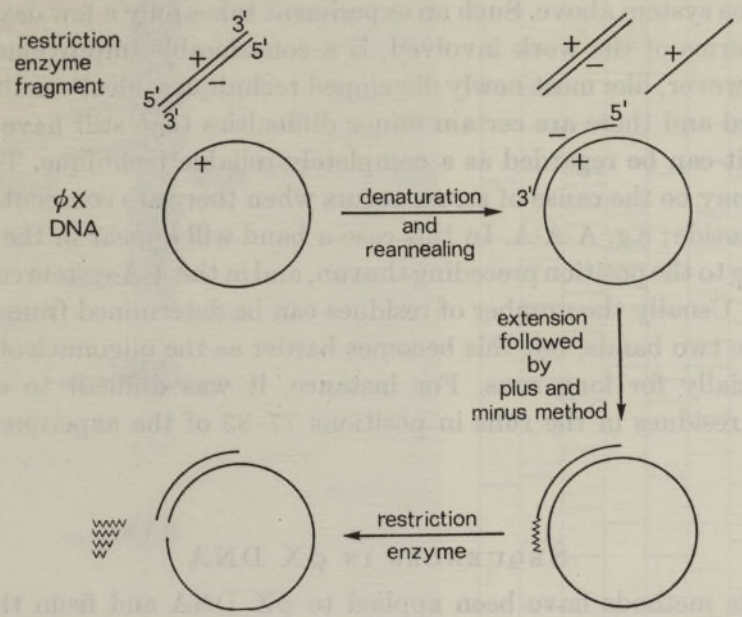

FIGURE 7. Diagram summarizing the use of restriction enzyme fragments as primers to determine sequences by the plus and minus method.

The method was initially worked out using the decanucleotide primer corresponding to the ribosomal binding site of protein G and, judging by the agreement with the results obtained by other methods, it was concluded that this method was essentially reliable.

Hitherto we have discussed only the use of synthetic oligonucleotides as primers. In spite of considerable progress recently, the synthesis of such primers is still difficult and laborious and consequently the possibility of using naturally occurring primers was explored. The most suitable primers were the fragments obtained by the action of restriction enzymes. These are well-defined double-stranded pieces of DNA with unique 5′ and 3′ termini and are therefore ideal for specific priming. Figure 7 summarizes the principle of the method. The double-stranded fragment is denatured and re-annealed in the presence of single-stranded φX DNA (the + strand). The unique substrate for DNA polymerase is the 3′ end of the − strand of the restriction enzyme fragment. This is extended as described above and the

eight separate incubations in the plus and minus systems carried out. If however these products were subjected to ionophoresis they would move very slowly since they still contain the primer and are therefore usually too large. However the site at which the extension started is also a site for digestion by the original restriction enzyme, so that the newly formed DNA can be liberated as suitable small products by digestion with this enzyme and fractionated on the acrylamide gel.

Figures 8 and 9, plates 24 and 25, show some results obtained by this method. Figure 8, in which *Hin*$_d$ fragment 1 was used as a primer, was a particularly clear one on which a sequence of residues could be read off, essentially from the results with the minus system above. Such an experiment takes only a few days and consequently, in terms of the work involved, is a considerable improvement on other methods. However, like most newly developed techniques, ideal conditions are not often obtained and there are certain minor difficulties that still have to be overcome before it can be regarded as a completely reliable technique. The main difficulty that may be the cause of errors occurs when there are consecutive 'runs' of a given nucleotide; e.g. A-A-A. In this case a band will appear in the −A system corresponding to the position preceding the run, and in the +A system corresponding to the last A. Usually the number of residues can be determined from the distance between these two bands, but this becomes harder as the oligonucleotides become larger – especially for long runs. For instance, it was difficult to estimate the number of A residues in the runs in positions 77–83 of the experiment shown in figure 9.

Sequences in ϕX DNA

The various methods have been applied to ϕX DNA and from the combined results two main sequences have been deduced and related to the amino acid sequences of the proteins for which they code.

The F protein cistron

When the sequences of the endonuclease IV digestion products studied by Ziff *et al.* (1973) were compared with the amino acid sequences of peptides obtained from the F protein by G. M. Air (personal communication), it was clear that the two were related to one another according to the genetic code, and the position of the endonuclease IV fragments on the genetic map was established (figure 1). J. W. Sedat, E. B. Ziff and F. Galibert (personal communication) established the complete sequence of fragments 6 and 5A and an almost complete sequence of fragment 3. Further data on fragment 3 were obtained by transcribing it with RNA polymerase (E. M. Blackburn, personal communication) and by applying the plus and minus method by using *Hin*$_d$ fragment 1 as primer. The 3′ end of the minus strand of this fragment is near the 3′ end (+ strand) of endonuclease IV fragment 3, and so priming with *Hin*$_d$ fragment 1 gives the complementary sequence to that found in endonuclease IV fragment 3 (figure 1). This experiment is shown in

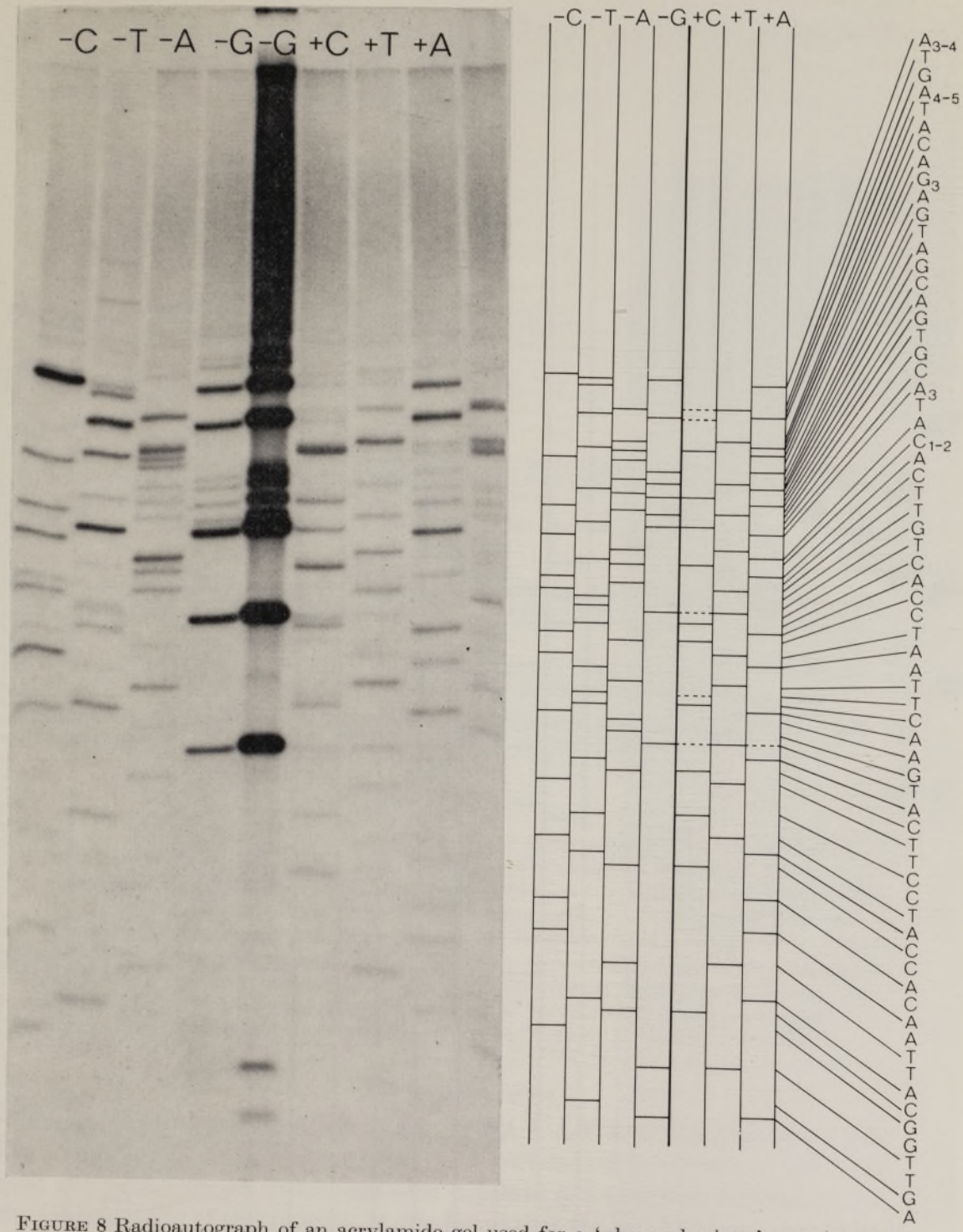

FIGURE 8 Radioautograph of an acrylamide gel used for a 'plus and minus' experiment in which Hin_d fragment 1 was used as a primer on ϕX DNA, and a diagram illustrating the interpretation and the sequence deduced. The very dark centre sample labelled −G contained five times as much material as the other samples. The +G sample was unsatisfactory in this experiment and was not used in the interpretation.

(*Facing p.* 328)

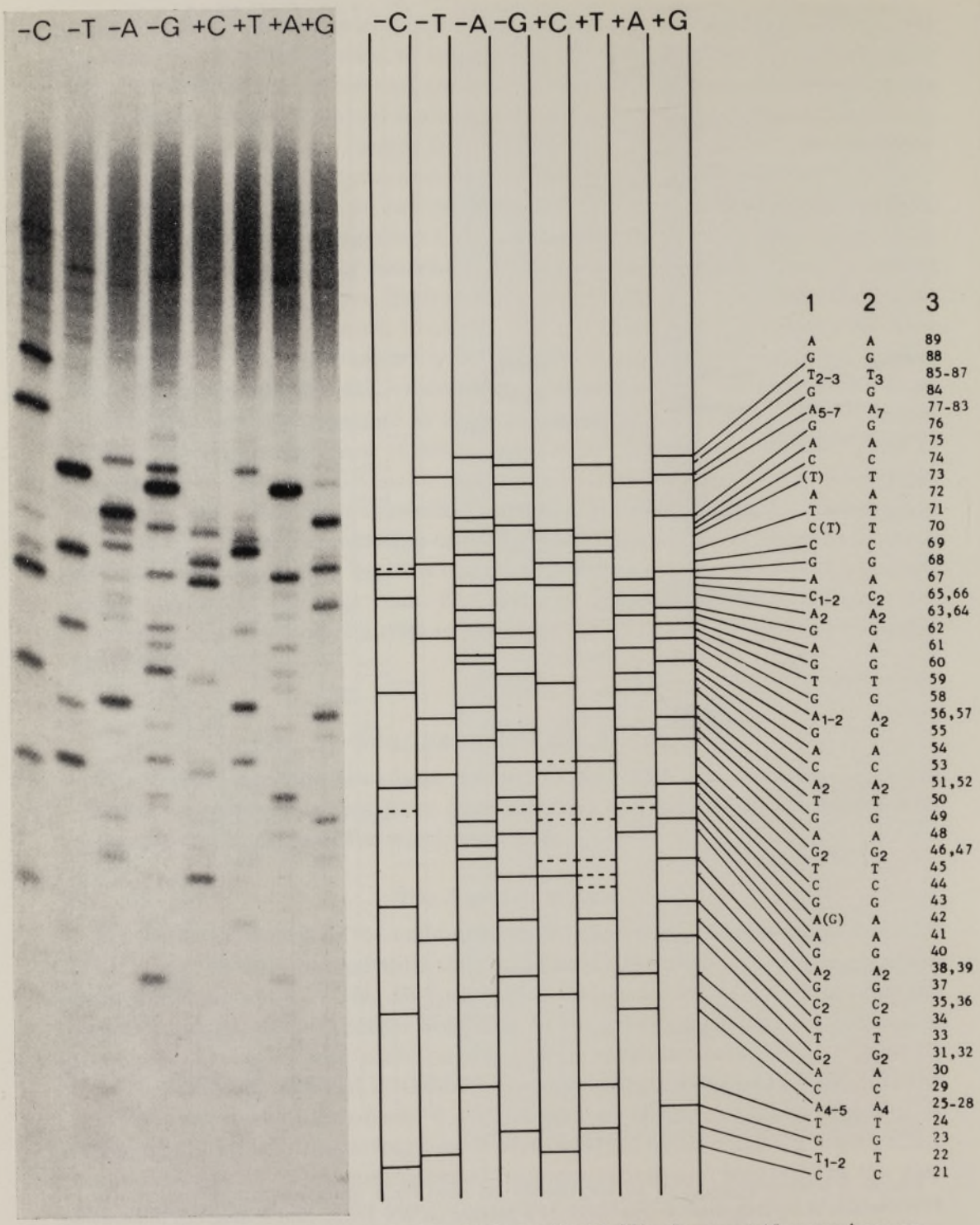

FIGURE 9. 'Plus and minus' experiment (as in figure 8) with Hin_d fragment 10 as a primer on ϕX DNA. Column 1 shows the sequence predicted from this single experiment and column 2 that finally deduced from other experiments and from the amino acid sequence (see figure 11).

figure 8. The combined results of these three techniques yielded a sequence of 283 nucleotide residues and this is shown in figure 10. The results agreed with the amino acid sequence, determined of course by an entirely different approach.

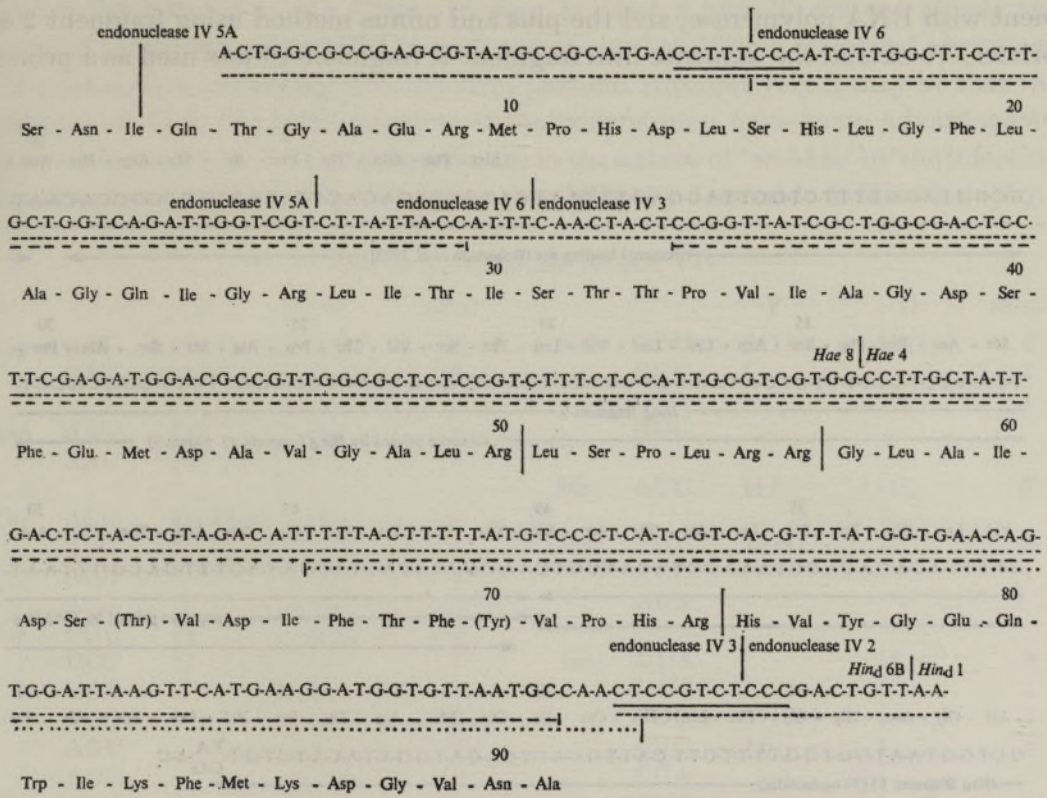

FIGURE 10. Nucleotide and amino acid sequences from gene F and its product. ----, sequences obtained by direct methods on the endonuclease IV fragments (Galibert *et al.* 1974; Sedat, Ziff & Galibert personal communication). – – – –, sequences determined by transcription with RNA polymerase (Blackburn 1975. ...., sequences obtained by priming with Hin_d fragment 1 (figure 8, plate 24) (Sanger & Coulson 1975). ———, sequences of the polypyrimidine tracts (Ling, 1972). The amino acid sequences were determined by G. M. Air (personal communication). Sites of cleavage by endonuclease IV and Hin_d II and *Hae* restriction enzymes are shown. Vertical bars indicate positions where either a nucleotide or amino acid sequence was overlapped only by reference to the other.

The G protein cistron

The synthetic decanucleotide primer A-G-A-A-A-T-A-A-A-A discussed above was complementary to a sequence near the *N*-terminus of the G protein and, when used as a primer on ϕX DNA, made the expected sequence that corresponded to the initiation site of this protein. It was found that the same sequence was made when the decanucleotide was used as a primer and the denatured form of Hin_d fragment 9 was used as template (J. C. Fiddes, personal communication). This established the position of the *N*-terminus of protein G in fragment 9. The relative order of Hin_d fragments was 9, 10, 2 (P. G. N. Jeppesen & L. Saunders, personal

communication; Lee & Sinsheimer, 1974), reading from 5′ to 3′ on the ϕX plus strand (see figure 1). Thus fragment 10 and the 5′ end of fragment 2 coded for amino acid sequences in protein G. Two methods were used to determine the 79-nucleotide long sequence of fragment 10; transcription of the denatured fragment with RNA polymerase, and the plus and minus method using fragment 2 as primer. To extend the sequence into fragment 9, fragment 10 was used as a primer

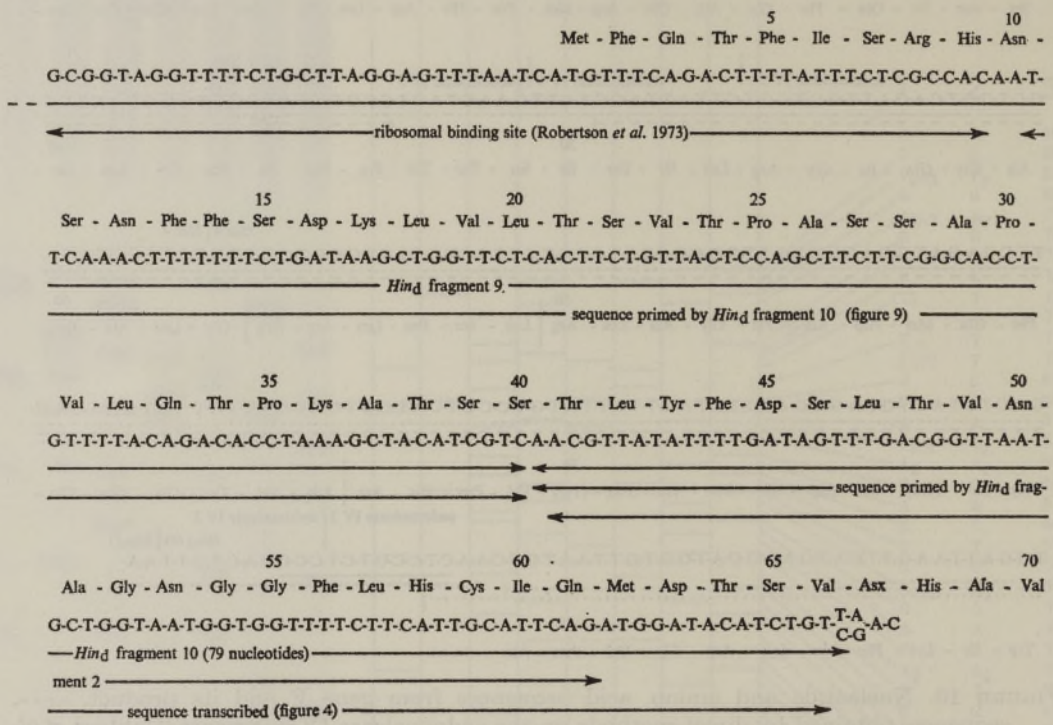

FIGURE 11. Amino acid sequence of the *N*-terminal region of the 'spike' protein of ϕX and the nucleotide sequence of the gene (G) that codes for it (Air *et al.* 1975).

and gave a 'predicted' sequence of 68 residues (figure 9). Although there were a number of uncertainties in this sequence, these could be eliminated by the knowledge of the amino acid sequence. Figure 11 shows the complete sequence of DNA and the corresponding amino acid sequence in gene G determined in this area (Air *et al.* 1975). It includes the DNA sequence in the initiation site obtained from the ribosomal binding site (Robertson *et al.* 1973) and by priming with the decanucleotide (J. E. Donelson & B. G. Barrell, personal communication).

The codons used in bacteriophage ϕX

In the above two sequences codons have been directly identified for 164 amino acid residues. These are enumerated in table 1. It will be seen that there is a strong preference for codons in which U is in the third position. 57 % of the total codons studied end in U, which is much more than would be expected on a random basis.

The biological significance of this finding is not clear at present. In the case of the RNA bacteriophages (R17, MS2 and Qβ) there was no such effect (Sanger 1971; Contreras, Ysebeart, Min Jou & Fiers, 1973): there appeared to be no significant preference for any particular codons. ϕX DNA has a high T content (32.7 %), which is obviously related to the effect. It may be that a high T content affects the physical properties of the DNA in such a way as to establish a biological advantage – for instance in packaging into the virus particle. Alternatively it may be that the presence of U in the third position of the codons may have some advantageous effect on translation; for instance a change in the nature of 'wobble' by the infecting bacteriophage has been suggested as a possibility by Denhardt & Marvin (1969).

TABLE 1. CODONS IN ϕX 174

		F	G	total			F	G	total
Asp	GAU	11	111	5	Val	GUU	1111	1111	8
	GAC	11111		5		GUC	1		1
						GUA	1		1
Asn	AAU	11	111	5		GUG			
	AAC	11	1	3					
					Met	AUG	111	11	5
Thr	ACU	111111	111	9					
	ACC	1		1	Ile	AUU	111111	11	8
	ACA		111	3		AUC	1		1
	ACG		11	2		AUA			
Ser	CUU	11	11111	7					
	UCC	11		2	Leu	UUA		11	2
	UCA	1	11	3		UUG	1	1	1
	UCG		11	2		CUU	111111	1	7
	AGU		1	1		CUC	11	1	3
	AGC					CUA			
Glu	GAA	1		1		GUG		1	1
	GAG	111		3					
					Tyr	UAU	111	1	4
Gln	CAA					UAC			
	CAG	11	111	5					
					Phe	UUU	11	111111	8
Pro	CCU	11	11	4		UUC	111		3
	CCC								
	CCA	1	1	2	Trp	UGG	1		1
	CCG	11		2					
					His	CAU	111	1	4
Gly	GGU	1111	111	7		CAC	1	1	2
	GGC	11111		5					
	CGA				Lys	AAA		1	1
	GGG					AAG	11	1	3
Ala	GCU	111111	111	9	Arg	CGU	1111111	1	7
	GCC	111		3		CGC		1	1
	GCA		1	2		CGA			
	GCG					CGG			
						AGA			
Cys	UGU					AGG			
	UGC		1	1					
					total		99	65	164

General conclusions

The lecture has been concerned mainly with the development of methods because it is on this that progress in sequencing depends, and it has been shown that techniques are now available for determining sequences, at least in some of the smaller DNAs. The simplest and most rapid method is probably the plus and minus technique described here which makes it possible to deduce sequences of about 50–80 residues in single-stranded DNA relatively quickly where suitable primers are available. At present the method is not completely reliable but the other techniques discussed above can be employed to provide confirmation relatively simply. One limitation of the plus and minus method is that it cannot be applied to double-stranded DNA, so that a strand separation either of the template or primer must first be carried out. For many double stranded DNAs this has been achieved experimentally, and in most cases it is theoretically possible.

Clearly further progress will be necessary before the larger DNAs can be studied. However the great variety of restriction enzymes that are becoming available make the outlook quite promising. They have already been used to obtain defined fragments from bacteriophage λ DNA (Maurer, Maniatis & Ptashne 1974) Allet, Roberts, Gesteland & Solem 1974) which is ten times as large as ϕX DNA.

Although up till now sequencing RNA has been easier than DNA, the availability of restriction enzymes, and the particular biological properties of DNA that make possible hybridization and polymerase copying procedures, are causing a rapid reversal of the situation, so that we may expect to see much progress in DNA sequencing in the near future; and it is hoped that this will provide a greater understanding of the basic chemistry of living matter.

References

Adams, J. M., Jeppesen, P. G. N., Sanger, F. & Barrell, B. G. 1969 *Nature, Lond.* **223**, 1009–1014.

Air, G. M., Blackburn, E. H., Sanger, F. & Coulson, A. R. 1975 *J. molec. Biol.* (In the Press).

Air, G. M. & Bridgen, J. 1973 *Nature New Biol.* **241**, 40–41.

Allet, B., Roberts, R. J., Gesteland, R. F. & Solem, R. 1974 *Nature, Lond.* **249**, 217–221.

Bambara, R., Jay, E. & Wu, R. 1974 *Nuc. Acids Res.* **1**, 1503–1520.

Benbow, R. M., Zuccarelli, A. J., Davis, G. C. & Sinsheimer, R. L. 1974 *J. Virol.* **13**, 898–907.

Berg, P., Fancher, H. & Chamberlin, M. 1963 *Symposium on informational macromolecules*, pp. 467–483. New York and London: Academic Press.

Billeter, M. A., Dahlberg, J. E., Goodman, H. M., Hindley, J. & Weissmann, C. 1969 *Nature, Lond.* **224**, 1083–1086.

Blackburn, E. H. 1975 *J. molec. Biol.* **93**, 367–374.

Brownlee, G. G. & Sanger, F. 1969 *Europ. J. Biochem.* **11**, 395–399.

Brownlee, G. G., Sanger, F. & Barrell, B. G. 1968 *J. molec. Biol.* **34**, 379–412.

Chen, C. Y., Hutchison, C. A. & Edgell, M. H. 1973 *Nature, New Biol.* **243**, 233–236.

Contreras, R., Ysebeart, M., Min Jou, W. & Fiers, W. 1973 *Nature, New Biol.* **241**, 99–101.

Denhardt, D. T. & Marvin, D. A. 1969 *Nature, Lond.* **221**, 769–770.

Edgell, M. G., Hutchison, C. A. & Sclair, M. 1972 *J. Virol.* **9**, 574–582.

Englund, P. T. 1972 *J. molec. Biol.* **66**, 209–224.

Galibert, F., Sedat, J. W. & Ziff, E. B. 1974 *J. molec. Biol.* **87**, 377–407.
Hindley, J. & Staples, D. H. 1969 *Nature, Lond.* **224**, 964–967.
Lee, A. S. & Sinsheimer, R. L. 1974 *Proc. Natn. Acad. Sci.* U.S.A. **71**, 2882–2886.
Ling, V. 1972 *J. molec. Biol.* **64**, 87–102.
Maniatis, T., Ptashne, M., Barrell, B. G. & Donelson, J. E. 1974 *Nature, Lond.* **250**, 394–397.
Maurer, R., Maniatis, T. & Ptashne, M. 1974 *Nature, Lond.* **249**, 221–223.
Min Jou, W., Haegeman, G., Ysebaert, M. & Fiers, W. 1972 *Nature, Lond.* **237**, 82–88.
Rensing, U. F. E. & Schoenmakers, J. G. G. 1973 *Europ. J. Biochem.* **33**, 8–18.
Robertson, H. D., Barrell, B. G., Weith, H. L. & Donelson, J. E. 1973 *Nature, New Biol.* **241**, 38–40.
Sadowski, P. & Hurwitz, J. 1969 *J. Biol. Chem.* **244**, 6192, 6198.
Salser, W., Fry, K., Brunk, C. & Poon, R. 1972 *Proc. natn. Acad. Sci. Wash.* **69**, 238–242.
Sanger, F. 1971 *Biochem J.* **124**, 833–843.
Sanger, F., Brownlee, G. G. & Barrell, B. G. 1965 *J. molec. Biol.* **13**, 373–398.
Sanger, F. & Coulson, A. R. 1975 *J. molec. Biol.* **94**, 441–448.
Sanger, F., Donelson, J. E., Coulson, A. R., Kössel, H. & Fischer, D. 1973 *Proc. natn. Acad. Sci.* U.S.A. 70, 1209–1213.
Schott, H. 1974 *Die Makromolekulare Chemie* **175**, 1683–1693.
Schott, H., Fischer, D. & Kössel, H. 1973 *Biochemistry*, **12**, 3447–3453.
Sinsheimer, R. L. 1968 In *Progress in Nucleic Acids Research and Molec. Biol.* (eds. W. E. Cohn & J. N. Davidson), **8**, 115–169. New York: Academic Press.
Steitz, J. A. 1969 *Nature, Lond.* **224**, 957–964.
Van de Sande, J. H., Loewen, P. C. & Khorana, H. G. 1972 *J. biol. Chem.* **247**, 6140–6148.
Wu, R. & Kaiser, A. D. 1968 *J. molec. Biol.* **35**, 523–527.
Wu, R. & Taylor, E. 1971 *J. molec. Biol.* **57**, 491–511.
Ziff, E. B., Sedat, J. W., & Galibert, F. 1973 *Nature, New Biol.* **241**, 34–37.