RAPID DNA SEQUENCE ANALYSIS

Author: Gillian M. Air
Department of Microbiology
John Curtin School of Medical Research
Australian National University
Canberra, Australia

Referee: Heinz Schaller
Microbiology Department
University of Heidelberg
Heidelberg, West Germany

INTRODUCTION

For some scientists, there are two standard responses to those who sequence nucleic acids: "Well, *I* don't believe in sequencing for the sake of sequencing" and "Well, what do you expect to get out of it?" Since the nucleotide sequence of DNA is the ultimate source of all the complex processes and variations scientists study, whether they are enzymologists, geneticists, immunologists, X-ray crystallographers, or developmental biologists, both remarks are extremely odd. They stem from the idea that DNA molecules are vast and unapproachable and that attempts to sequence them involve long years of hard labor for an insignificant amount of information. Eukaryotic DNA certainly provides a vast area of research, but the individual genes do not. The purpose of this article is to describe the new sequencing technologies by which complete sequences of the DNAs of organisms, such as bacteriophages ϕX 174[1] and fd[2] and animal virus SV40,[3] have been determined. Long sequences from more complex cells have also been obtained, not as a sequencing exercise, but to study processes such as interaction of the lambda operators with repressor,[4] transcriptional control in the prokaryotic *trp* operon[5,6] and eukaryotic genes,[7,8] and features of DNA replication and protein synthesis.[1] An important use of the rapid sequencing methods is to determine the exact genetic composition of an eukaryotic DNA fragment which is to be cloned in bacteria.[9,10]

The two techniques which made sequencing of whole genomes a relatively short-term project are those of Sanger and Coulson[11] and Maxam and Gilbert.[12] Both methods depend on the fact that single-stranded DNA fragments will migrate on polyacrylamide gel electrophoresis under denaturing conditions strictly according to size. Hence, if in a set of fragments of varying length from a fixed end point the terminal (or terminal plus one) nucleotide of each product can be identified, then the sequence of nucleotides can be read from the position of each product on the gel. The method

of sorting the various size classes of oligonucleotides according to the sequence is entirely different in the two techniques. Sanger and Coulson[11] used a staggered primed synthesis of DNA, incorporating an α-^{32}P-deoxyribonucleoside triphosphate, while Maxam and Gilbert[12] used partial chemical degradation after labeling one end of the oligonucleotide with ^{32}P. The enzymes used for primed synthesis of DNA, synthesis of primers, and end-labeling are shown in Table 1.

The original methods of sequencing nucleic acids[20,21] depended on the incorporation of ^{32}P in vivo. Such labeling is not very efficient, and the short half-life of the isotope means that by the end of a series of procedures to sequence the fragment of DNA, the results often can not be seen. In vitro labeling offers two major advantages over the in vivo experiments. First, the label is introduced with high efficiency and can be restricted to a particular region of interest. Second, since the precursors used are nucleoside triphosphates, the ^{32}P label can be restricted to each of the four nucleotides in turn; thus, the radioisotope is being used as part of the sequence determination as well as a means of detecting those amounts too small for chemical determination. The rapid methods all depend on in vitro labeling which is introduced either internally using α-^{32}P deoxyribonucleoside triphosphates or at the 5′ or 3′ ends (see Table 1). Originally, fragments of DNA of suitable length for sequencing (approximately 100 nucleotides) were obtained by experiments such as the isolation of regions of DNA protected from nuclease digestion by specific binding to proteins such as the lambda repressor[22] or RNA polymerase.[23] Ziff et al.,[21] Galibert et al.,[24] and Sedat et al.[25] were able to isolate and sequence fragments up to approximately 250 nucleotides long by partial digestion with endonuclease IV; however, this enzyme often does not give consistent results.[26,27] The primers used in copying reactions with DNA polymerase were oligonucleotides synthesized by somewhat long and laborious chemical methods.[28]

The new rapid methods of sequencing DNA would, therefore, have been very limited in scope. Sequencing long lengths of DNA was only made possible by the large variety of restriction endonucleases which became available about the same time as the new sequencing methods were developed. The restriction enzymes (those in Class II) recognize specific sequences in double-stranded DNA and cleave both strands, normally within the recognition sequence. For example, one of the enzymes from *Haemophilus aegyptius, HaeIII*,[30] (for the nomenclature of restriction endonucleases see Reference 29) cleaves the sequence:

$$\begin{array}{ll} 5' & \text{G–G}\overset{\downarrow}{\text{–C}}\text{–C} \\ 3' & \text{C–C}\underset{\uparrow}{\text{–G}}\text{–G} \end{array}$$

In this case, cleavage is at the same phosphodiester bond on both strands. Other enzymes may leave cohesive ends, for example HindIII:[31]

$$\begin{array}{l} \overset{\downarrow}{\text{A}}\text{–A–G–C–T–T} \\ \text{T–T–C–G–A}\underset{\uparrow}{\text{–A}} \end{array}$$

The recognition sequences are often symmetrical and may be tetra-, penta-, or hexanucleotides. In some cases, the cleavage site is away from the recognition site, for example MboII:[32]

$$\begin{array}{ll} 5' & \text{G–A–A–G–A–N–N–N–N–N–N–N–N}\overset{\downarrow}{\text{–N}} \\ 3' & \text{C–T–T–C–T–N–N–N–N–N–N–N}\underset{\uparrow}{\text{–N}}\text{–N} \end{array}$$

TABLE 1

Enzymes to Label or Synthesize DNA and Oligonucleotide Primers

Enzyme	Major reaction	Template/Primer requirements	Other reactions	Ref.
DNA polymerase I (*E. coli*)	$(DNA)_{OH}$ + n dNTP (or oligonucleotide) $\overset{Mg^{2+}}{\rightleftharpoons}$ DNA − $(pdN)_n$ + nPPi	ss template; primer with 3′-OH	With Mn^{2+} will polymerize rNTPs, or use RNA template. 5′ exonuclease; ds DNA → 5′ mono-, di-, tri-nucleotides. The 5′ activity can be specifically removed by proteolysis leaving the Klenow fragment. 3′ exonuclease; ds and ss DNA → 5′ mononucleotides; ds activity blocked by synthesis	13
T4 polymerase (Phage T4-infected *E. coli*)	$(DNA)_{OH}$ + n dNTP $\overset{Mg^{2+}}{\rightleftharpoons}$ DNA − $(pdN)_n$ + nPPi	ss template; free 3′-OH. In the absence of added primer can loop back to start synthesis	3′ exonuclease, ss>ds DNA → 5′ mononucleotides. ds activity blocked by synthesis	14
AMV reverse transcriptase (avian myeloblastosis virus)	$(DNA)_{OH}$ or $(RNA)_{OH}$ (or oligonucleotide) + n dNTP $\overset{Mg^{2+}}{\rightarrow}$ DNA−$(pdN)_n$ or RNA−$(pdN)_n$ + nPPi	RNA template; primer with 3′-OH	Also will use DNA template. Ribonuclease H. Specific for RNA-DNA hybrid. Processive 5′ and 3′ exonuclease, generating small oligoribonucleotides.	15
Terminal deoxynucleotidyl transferase (calf thymus)	$(DNA)_{OH}$ + n dNTP (or oligonucleotide) $\overset{Mg^{2+}}{\rightarrow}$ DNA−$(pdN)_n$ + nPPi	Deoxynucleotide primer with 3′-OH	Also adds 1—2 ribonucleotides. For 3′-end labeling: α-^{32}P-rNTP is used, followed by alkali to cleave off the ribonucleotides, leaving 3′-^{32}P	16
Polynucleotide phosphorylase (*E. coli*)	n NDP $\overset{Mg^{2+}}{\rightarrow}$ $(pN)_n$ + nPi Stimulated by primer with 3′-OH, at least a dinucleotide → primer − $(pN)_n$		Primer + n dNDP $\overset{Mn^{2+}}{\rightleftharpoons}$ primer − $(pdN)_{\geq 2}$ + nPi ($\geq$ deoxytrinucleotide)	17, 108
ATP:RNA adenyl transferase (*E. coli*)	$(RNA)_{OH}$ + ATP $\overset{Mg^{2+}Mn^{2+}}{\rightleftharpoons}$ RNA−$(pA)_n$ + PPi	RNA primer, > trinulceotide, with 3′-OH		18
Polynucleotide kinase (T4-infected (*E. coli*)	$_{OH}$(DNA) + ATP → pDNA + ADP For 5′-end labeling: γ-^{32}P-labeled ATP is used, giving 5′-^{32}P DNA	5′-OH	Also phosphorylates RNA and small oligonucleotides	19

The restriction endonucleases have recently been reviewed by Roberts[33] who is responsible for isolating and characterizing many of the 91 enzymes listed. Although several of these are isoschizomers (enzymes which recognize the same nucleotide sequence) so many recognition sequences have now been identified that, by using one or more restriction enzymes, practically any region of DNA can be broken down into fragments of 100 to 200 nucleotide pairs. This is the ideal length for either a primer in the Sanger and Coulson method[11] or direct sequencing by the Maxam and Gilbert procedure.[12]

The new rapid methods can potentially give sequences of several hundred nucleotides within a week, but as described in later sections of this article, some difficulties and ambiguities can arise. Thus, it is advisable to obtain data by other methods to confirm the sequence. The provisional sequence read from gels provides a framework in which those particular methods which will give the most useful information can be planned. If small oligonucleotides are sequenced, there is no requirement for several sets because the gel sequence provides the overlapping.

The older methods of sequencing DNA have been comprehensively reviewed.[34-36] Therefore, in this article, only those methods most useful in confirming sequences will be outlined. Enzymes and chemicals used to cleave DNA are shown in Table 2.

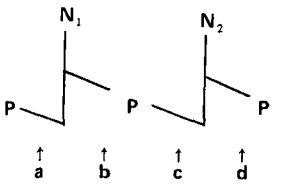
The depurination reaction[38] is one of the most useful procedures. Brown and Smith[39] confirmed a long sequence from ϕX 174 by depurinating both strands of small restriction fragments (preferably <100 nucleotides). This was done using in turn labeling with α-^{32}P- dATP or dGTP to identify the purine at the 3′ end of each pyrimidine tract, since the α-^{32}P becomes the 3′ phosphate (see Table 2). From the composition of each product, the 3′ purine, and the relative amounts of each isostich (compositional isomer), the sequence they obtained by the Sanger and Coulson method[11] was shown to be correct. Many sequences from the Maxam-Gilbert method[12] have also been confirmed by depurination analysis.

Small restriction-enzyme fragments, or products of digestion with an enzyme such as endonuclease IV, can be sequenced by partial digestion with spleen or venom phosphodiesterase. These exonucleases act at a 5′ hydroxyl and 3′-hydroxyl group, respectively (see Table 2). If necessary, a 5′ or 3′ phosphate can be removed with bacterial alkaline phosphatase. Pyrimidine tracts of up to 20 nucleotides long have been sequenced by partial exonuclease digestion, with separation of the products by electrophoresis at pH 3.5 on cellulose acetate, followed by homochromatography on DEAE-cellulose thin layers in the second dimension.[40] Since the T and C nucleotides have different mobilities on electrophoresis, and homochromatography separates according to size, a trail of spots is obtained extending up the chromatography plate and representing every possible product of the exonuclease. From the relationship of each spot to the one before, it is possible to identify whether the difference is due to loss of a C or a T and, hence, deduce the sequence. This method has been extended to identify all four nucleotides[21,24,25] and has been extensively used as a primary sequencing method, often with 5′-end labeling and partial venom phosphodiesterase digestion.[41,42] It is still useful as a supplementary method to check sequences obtained by the rapid methods.

In small DNA fragments, especially those which are single-stranded, nearest-neighbor analysis can give useful confirmatory information.[20] The α-^{32}P incorporated is at the 5′ site, whereas most enzyme digestions or chemical degradations cleave the 5′-phosphate bond (see Table 2). Hence, the input ^{32}P becomes the 3′ phosphate of the adjacent nucleotide.

One of the major difficulties in DNA sequence work is the lack of single-base-specific enzymes. However, the DNA can be transcribed into RNA and then the standard RNA sequencing techniques can be applied to digestion products of T1, U2, and pan-

TABLE 2



Methods of Cleaving DNA

Procedure	Substrate	Type of reaction	Specificity of cleavage	Predominant reaction products (5′→3′)
Snake venom phosphodiesterase	RNA, ssDNA	Exonuclease	At b where N_2 is the 3′ terminus and carries a 3′-OH group	pN
Spleen phosphodiesterase	RNA, ssDNA	Exonuclease	At c where N_1 is the 5′-terminus and carries a 5′-OH group	Np
Alkaline phosphatase (bacteria)	RNA, DNA	Phosphatase	At a where N_1 is the 5′ terminus; at d where N_2 is the 3′ terminus	P_i and oligonucleotide with 3′-OH and 5′-OH groups
Exonuclease III (E. coli)	dsDNA	Exonuclease and phosphatase	At b (and d) where N_2 is the 3′ terminus and carries 3′-OH (or -P)	pN
Lambda exonuclease	dsDNA	Exonuclease	At b where N_1 is the 5′ terminus and carries a 5′-P	pN
Micrococcal nuclease	DNA, RNA	Endonuclease	At c where N_1 and N_2 are any nucleotides	Np, NpNp
Endonuclease IV (T4-infected *E. coli*)	ssDNA	Endonuclease	At b where N_1N_2 is TC	pC———T
Deoxyribonuclease I (pancreas)	DNA	Endonuclease	At b where N_1 and N_2 are any nucleotides	Nucleotides and small oligonucleotides pN———N, pN
Deoxyribonuclease II (spleen)	DNA	Endonuclease	At c where N_1 and N_2 are any nucleotides	Nucleotides and small oligonucleotides N———Np, Np
Restriction enzymes	dsDNA[a]	Endonuclease	At b where N_1 and N_2 are part of the appropriate recognition sequence	pN———N
Formic acid plus diphenylamine	DNA	Depurination	At c where N_1N_2 is PyPu; at b where N_1N_2 is PuPy	Pyrimidine tracts pPy———Pyp
Hydrazine followed by alkaline hydrolysis	DNA	Depyrimidination	At c where N_1N_2 is PuPy; at b where N_1N_2 is PyPu	Purine tracts pPu———Pup
Dimethylsulphate followed by alkaline hydrolysis	DNA	Depurination	At c where N_1N_2 is PyPu; at b where N_1N_2 is PuPy	pPy———Pyp
Ribonuclease T_1	RNA[b]	Endonuclease	At c where N_1 is G	Np———Gp
Ribonuclease T_2	RNA[b]	Endonuclease	At c where N_1 is usually A, sometimes G	Np———Ap
Ribonuclease U_2	RNA[b]	Endonuclease	At c where N_1 is a purine	Np———Pup
Pancreatic ribonuclease	RNA[b]	Endonuclease	At c where N_1 is a pyrimidine	Np———Pyp
Alkali	RNA[b]	Esterolysis	At c where N_1 is any nucleotide	Mixture of 2′ and 3′ nucleotides

Note: N, any nucleoside; pN, nucleoside 5′-phosphate; Np, nucleoside 3′-phosphate; Py, pyrimidine nucleoside; Pu, purine nucleoside. For details of individual enzymes, see Reference 37; restriction enzymes are reviewed by Roberts.[33]

[a] Some restriction enzymes may also cleave ssDNA.

[b] Some restriction enzymes may also cleave ssDNA or ribo-substituted DNA.

creatic ribonucleases[20] (Table 2). Alternatively, ribosubstitution of DNA can be used. In the presence of manganese ion instead of magnesium, *Escherichia coli* DNA polymerase I will insert ribonucleotides into the synthesized sequence.[28,43,44] Therefore, if three deoxyribonucleoside triphosphates and one ribonucleoside triphosphate (e.g., rCTP) are present, then every C in the synthesized oligonucleotide is susceptible to alkali or pancreatic ribonuclease cleavage. The resulting oligonucleotides can be isolated, their composition determined, and, if necessary, sequenced to confirm a DNA sequence.

If the amino acid sequence of the protein coded by the DNA is known, it can be compared to the provisional DNA sequence.[9,10,45-48] Since each amino acid may be specified by up to six triplet codons, this comparison does not confirm every nucleotide. However, it does confirm whether the number of nucleotides is correct — a deletion or insertion is immediately obvious in the triplet codons. In reading gel sequences,[11,12] it is difficult to be certain that the sequence length is correct and that no bands have been missed or artifact bands included. Thus, the amino acid sequence can be very helpful. Unfortunately, protein sequencing remains a time-consuming project in spite of the development of automated machinery.[49]

The simple and rapid methods of sequencing DNA cannot be adapted for proteins for the same reason that nucleotide sequencing was once considered intrinsically more difficult than protein sequencing. The four nucleotides have a great deal of similarity and can be made to behave essentially identically in terms of solubility, electrophoretic mobility, susceptibility to both polymerizing and degradative enzymes, and so on. Amino acids, on the other hand, have vastly different side-chain structures. Peptides do not behave consistently as residues are added or removed, and rates of exo- or endo-protease reactions are unpredictable until the whole sequence is known. Insolubility remains a serious problem in protein sequence work, as there are still no general methods to purify small, insoluble peptides. These problems of solubility and differing susceptibility to enzyme attack are absent in nucleotide sequencing work — all fragments from a restriction enzyme digest can be seen on the appropriate polyacrylamide gel in equimolar proportions. Bands containing unresolved fragments are immediately recognizable as double the intensity of their neighbors. In a protein digest, peptides may be present in varying yields due to partial cleavage or insolubility. Since all detection methods are dependent to some extent on the amino acid composition, the major products cannot be readily distinguished from the many minor ones. Therefore, the easiest way to determine the amino acid sequence of a protein is to sequence the nucleic acid which contains the coding information.[48]

This article describes the two rapid methods by which a large amount of DNA sequence information has recently been obtained. Some newer methods are also described. Also included is a section on sequencing RNA, since at the present time the easiest way to sequence RNA is to transcribe it into a DNA copy and apply the DNA methods.

THE SANGER AND COULSON "PLUS AND MINUS" SEQUENCING METHOD

The "plus-and-minus" technique of Sanger and Coulson[11] involves two stages of DNA synthesis. Starting with a single-stranded DNA template, a complementary sequence is annealed as primer. The primer can be a synthetic oligonucleotide or, usually easier to obtain, a double-stranded restriction fragment of which only the strand complementary to the template will anneal. The four deoxynucleoside triphosphates, one of which is α-^{32}P labeled, and *E. coli* DNA polymerase I are added. DNA synthesis is random; in order to ensure a mixture of lengths of the synthesized products, aliquots

are taken at various times and the synthesis terminated. All of these aliquots are then recombined when, ideally, every possible length of product from 1 to 200 or more residues are present. If the primer is a long fragment, it is removed by the appropriate restriction enzyme, and the remaining products of synthesis are resolved by polyacrylamide gel electrophoresis under denaturing conditions. Each band on the gel is related to the next slowest by the addition of one nucleotide. Two methods are used to identify this nucleotide. In theory, either the "minus" or the "plus" system alone is sufficient; however, in practice, neither works perfectly all the time, and both are used to obtain the sequence.

The Minus System

The synthesized mixture of oligonucleotides is divided into eight aliquots: four for the minus reactions and four for the plus reactions. In the minus system, the oligonucleotides are elongated with DNA polymerase and three nucleoside triphosphates. The fourth triphosphate is missing; hence, in each of the four reaction mixtures (missing dATP, dGTP, dCTP, and dTTP, respectively), synthesis continues until it stops because the nucleoside triphosphate specified by the template is missing. The "−A" mix, for instance, which started with every possible length of product, now contains only those lengths in which the next nucleotide would be an A. The "−G" mix contains a different set of lengths of oligonucleotides, one in which the next residue would be a G, and so on for the −C and −T mixes. The result is that every possible length of product is now in only one of the four mixes, and when the four mixes are run side by side on a 12% polyacrylamide gel, after cleaving off the restriction-fragment primer and dissociating the template by heating in formamide, the sequence of nucleotides can be read from a radioautograph of the gel by noting in which mix each band (size) occurs, starting with the smallest fragments at the bottom of the gel.

The Plus System

To each of the second four aliquots of the original elongation mix, one deoxyribonucleoside phosphate (dATP, dGTP, dCTP, or dTTP) is added, with DNA polymerase from phage T4-infected *E. coli* (T4 polymerase, Table 1). This enzyme degrades double-stranded DNA from the 3′ end. However, if a single nucleoside triphosphate is present, the exonuclease action effectively stops, since the nucleotide is polymerized back faster than it is exonucleased.[50] Hence, the "+A" mix will contain only those products in which A is the 3′ nucleotide: the pattern of bands should be the same as in the −A reaction, but one nucleotide longer (i.e., slower on the gel). The sequence can therefore be read from a radioautograph of the polyacrylamide gel with all four plus reactions run side by side. In practice, the four minus reactions and the four plus reactions are run on the same gel so that use can be made of the one-nucleotide relationship between a "−N" and an "+N" product.

Interpretation of the Sequence

Ideally, either the minus or plus systems should determine a complete nucleotide sequence of 100 to 200 residues, but certain problems can occur. A primary problem is seen when there is a "run" of a particular nucleotide. In this case, only one product is usually seen; it is the shortest in the minus reaction and the longest in the plus reaction. Measuring the distance between these products indicates the number of bands not seen. However, when the run is four nucleotides or more and especially if it occurs near the top of the gel, it is difficult to be sure of its length. In some cases, bands are missing completely, and occasionally artifact bands appear. These missing and extra bands usually are not consistent in different incorporations, and repeating the experiment will often remove any ambiguities. Persistent problems can occur when there is

secondary structure in the DNA template.[51] Priming from another site or using the complementary single strand as template helps to confirm the sequence. The original procedure has been modified,[39,48] but it is still wise to check the sequence obtained by some other method, such as depurination analysis of both strands of a short restriction fragment,[39,51] or, in regions which code for proteins, the sequence can be compared to a known amino acid sequence.[46-48,52] (For detailed method protocol, see Barrell.)[133] It is important to run the polyacrylamide gels at a fairly high current, as the heating helps to keep the DNA denatured.[11,39] At lower currents, any secondary structure may cause anomalous migration resulting in a compression of bands and, therefore, an unreadable region of sequence. A more serious problem with secondary structure is that the polymerase may be halted (or even diverted) by a strongly base-paired hairpin structure. The result is a gap in the gel sequence; as there are no consistent ways of overcoming this difficulty, the region must be sequenced by other methods.[51,53]

The four gels shown in Figure 1, plus another shown in Reference 48, produced a continuous sequence of 348 nucleotides from gene G of bacteriophage ϕX 174.[48] Each experiment was repeated several times to ensure consistent results. Since the entire sequencing procedure is complete in 24 hr, the time and effort involved is remarkably small considering the quantity of information obtained.

These four gels do show some of the problems encountered in reading the plus-minus gels. Firstly, some experiments produce better results (Figures 1A and 1C) than others (Figure 1B). This seems to be at least partly due to the variations in quality of the restriction fragment preparations used as primers. The sequence in Figure 1A starts approximately 60 nucleotides from the priming site, and the nucleotide numbered 88 is therefore 146 residues long. Nevertheless, it is resolved sufficiently from the 145-residue-long nucleotide to obtain the correct sequence up to this point. However, there are a few artifacts in, for example, the +C system at positions 14 and 41. These were not present in other experiments with the same primer and, therefore, did not cause too much confusion in reading the sequence. Occasionally, there is a persistent artifact band, as at position 4 in the −T mix in Figure 1A. This band was found in most experiments with the same primer. When another primer was used from a close but different site, the artifact −T was entirely absent.

In the gel shown in Figure 1B the bands are fainter and somewhat broader, making the reading more difficult. There are also some artifact bands, e.g., in positions 94 (+C) and 96 (+T). The gel shown in Figure 1C was run for a long time — the C at position 229 is 108 nucleotides from the priming site. The purine-rich sequence which follows could not be confidently read from the shorter electrophoresis run, and in Figure 1C a longer run is shown. The order of the samples on the gel was such that the plus and minus systems of the purines in question were side by side. The sequence was then read without difficulty. The sequence obtained was 348 nucleotides long and coded for the known amino acid sequence of the gene G protein.[48]

The major application of the Sanger and Coulson method has been in sequencing the DNA of bacteriophage ϕX 174. This DNA is a single-stranded circular molecule of approximately 5375 nucleotides, and the sequence reported by Sanger et al.[1] was the first complete nucleotide sequence of a DNA genome. "Complete" is a slight exaggeration, for, as Sanger and co-workers have repeatedly emphasized, the gel plus and minus method is not absolutely reliable. Some nucleotides in the ϕX 174 sequence were not confirmed by independent methods and were, therefore, considered uncertain. The plus and minus method was developed using ϕX 174 DNA with either a synthetic decanucleotide or restriction fragments as primers. Some improvements have been made in the original protocol.[39,48,133] Much of the sequence was confirmed by the comparison with known amino acid sequences of the proteins coded by ϕX 174.[45-49,52] Where protein sequences were not available and in the earlier work using the plus and

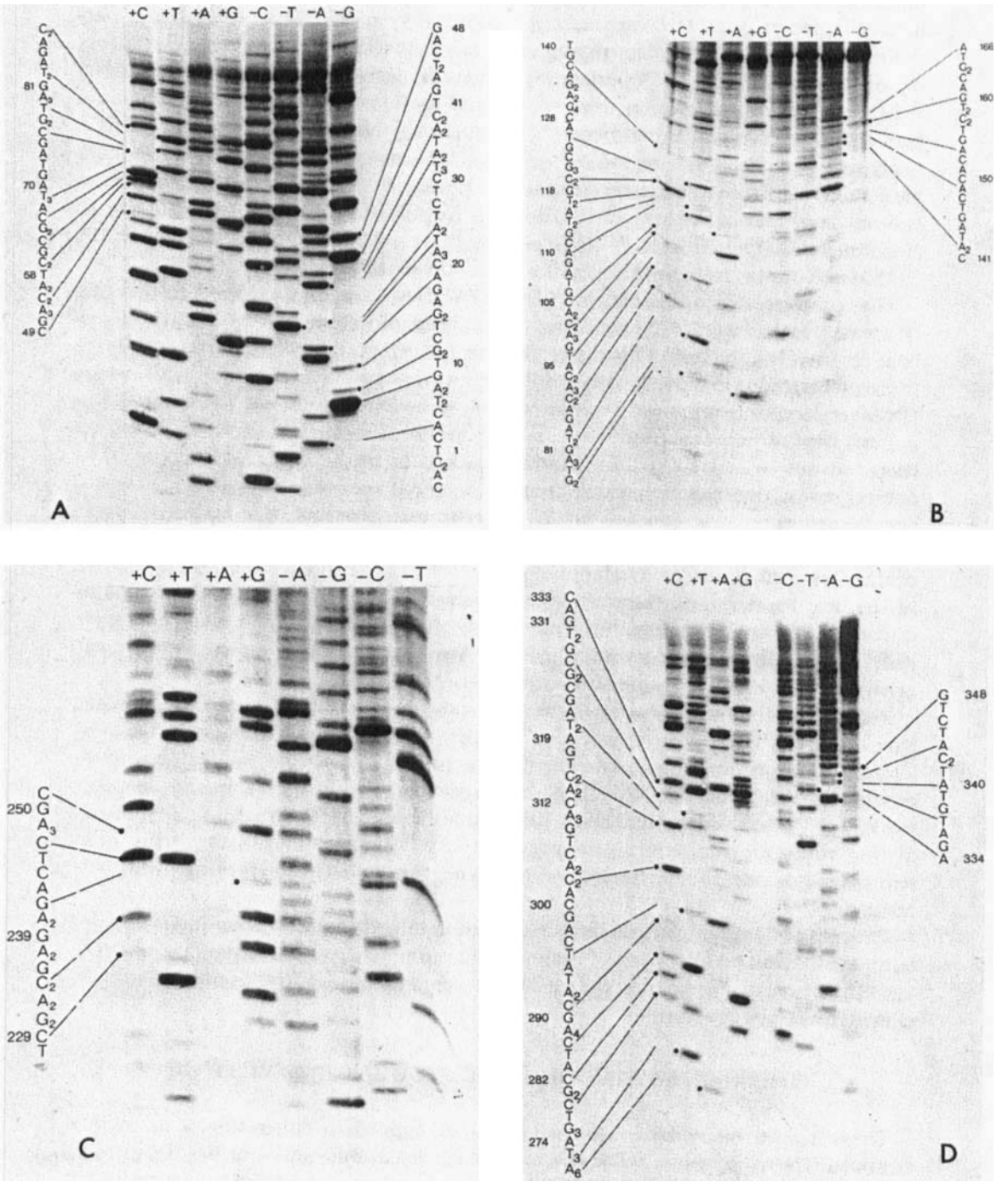


FIGURE 1. Four radioautographs of sequences from gene G of ϕX 174 determined by the Sanger and Coulson method.[1] A. The sequence obtained by priming with fragment *Hha* 2.[11] The nucleotide numbered "1" is 57 residues from the priming site and the sequence is read to 145 nucleotides from the site. B. The sequence primed by fragment *Hinf* 4. The first nucleotide shown is 31 from the priming site. Although the sequence extends for 91 nucleotides, there are several artifact bands while other bands were missing. This experiment was repeated several times to obtain the sequence shown, which was confirmed by the amino acid sequence. C. The sequence primed by fragment *Hap* 5. The gel shown is the long electrophoretic run, in which the first nucleotide indicated is 107 from the priming site. The short electrophoresis run (not shown) overlapped the data obtained from gel B. D. The sequence primed by fragment *Hinf* 8. The continuous sequence obtained from five polyacrylamide gel electrophoreses is 352 nucleotides long. (With permission from Air, G. M., Sanger, F., and Coulson, A. R. *J. Mol. Biol.*, 108, 519, 1976. Copyright by Academic Press, Inc. (London) Ltd.)

minus method, the sequences read from gels were confirmed by their comparison with data obtained by other DNA sequencing methods (by endonuclease IV digestion, partial exonuclease digestion, depurination analysis, ribosubstitution,[39,51] or the sequencing of RNA transcripts.[54,55] Overlapping sequences, derived by priming from a different (but close) restriction site, have been found useful if ambiguities or uncertainties persist through several experiments.[39,48] Sequencing the complementary strand is a fairly rigorous check of a region of sequence.[39,52] Sequences of up to 116 nucleotides have been read from a single gel (see Figure 1), and, if the reaction mix is divided into two and half loaded onto a gel which is run for a longer time, the sequence can be read beyond 200 nucleotides.[52] Routinely, the gels can be read with confidence to 100 to 150 nucleotides from the priming site,[39,48] as shown in Figure 1.

The complete nucleotide sequence of ϕX 174 DNA is not only a demonstration that it is easy to sequence 5375 nucleotides. The main objective was to identify the sequences involved in control processes, such as the promoters which initiate RNA synthesis, the ribosome-binding sites initiating protein synthesis, and the region where DNA replication is initiated. Such sequences were identified in ϕX 174 — three promoters, nine protein initiation sites, and the origin of replication; however, it has not been possible to single out the features required for these initiation processes.[1] The one region of the sequence which contains marked secondary structures, as well as peculiarities such as a sequence of 17 A·T base pairs, has not been assigned a function.[51] A remarkable feature of the ϕX 174 genome, not previously suspected, is the occurrence of two sets of overlapping genes.[52,56] The E protein, responsible for lysis of the host bacterium to release phage particles, is coded by the same nucleotide sequence as the gene D protein, but read in a different triplet reading frame.[52] Similarly, gene B is totally contained within gene A. Again, the reading frames are different, giving entirely different amino acid sequences of the proteins.[39,56,57]

Genetic studies in these regions produced anomalous results;[58-60] however, it would have been very hard to prove such gene overlaps solely by genetic techniques, the overlapping phenomenon being unknown. The nucleotide sequence in conjunction with the genetics of the phage established the overlaps unambiguously. Nucleotide sequence work in SV40 has shown that genes VP2 and VP1, coding for structural components of the virion, are similarly overlapped. The last 112 nucleotides of the VP2 coding sequence also code, in a different reading frame, for the N-terminal region of the VP1 protein.[61,62,82]

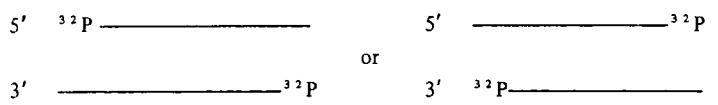
The plus and minus methods have been used in other systems when a single-stranded template is available. Beck et al.[53] sequenced a promoter from bacteriophage fd DNA, and Birnsteil et al.[63] have used the method on separated single strands of histone genes cloned in a lambda vector.

THE MAXAM AND GILBERT SEQUENCING METHOD

The chemical method of sequencing can be applied to either single- or double-stranded DNA fragments.[12] The fragment is labeled at one end — at the 3′ end using terminal transferase and α-^{32}P ATP[64] or at the 5′ end using polynucleotide kinase[65-67] and γ-^{32} ATP (see Table 1). In practice, the latter reaction has been used almost exclusively for sequence work since γ-^{32}P ATP could be obtained at over ten times the specific activity of α-^{32}P dATP. Schwartz et al.[68] have labeled at the 3′ end using terminal transferase. As an alternative method of 3′-end labeling, they used DNA polymerase I and α-^{32}P-labeled nucleoside triphosphates to fill in the cohesive ends generated by some restriction enzymes. Arrand and Roberts[69] had to use the 3′-end labeling to determine the nucleotide sequence of the ends of adenovirus-2 DNA, since the 5′ end is blocked by covalent attachment of a protein.[70] They used the repair activity of T4

DNA polymerase with α-^{32}P-labeled dGTP, the nucleotide complementary to the 5′ C which is attached to protein.

Either of the end-labeling procedures results in a fragment labeled at one end of both strands:



There are two ways of obtaining the single label required for sequence analysis. In some cases, the denatured complementary strands of the fragment can be separated by gel electrophoresis. Otherwise, the fragment is cleaved by another restriction enzyme into two unequal fragments, which are separated as the double-stranded structures, each having the ^{32}P-end label on only one strand.[12]

To obtain the sequence, the end-labeled DNA is partially degraded by reagents which damage, then excise, a base from its sugar. This sugar is then highly susceptible to alkali or amine cleavage of both 3′ and 5′ phosphates. Although the reagents used do not have absolute base specificity for cleavage, the purine and pyrimidine reagents will, under appropriate conditions, preferably attack adenine or guanine, and cytosine or thymine, respectively. Since the ^{32}P label is at only one end of the DNA fragment, partial cleavage will generate a series of products which can be seen on the radioautograph of an acrylamide gel only when the labeled end is intact. In other words, a specific set of lengths is obtained for cleavage at a specific nucleotide, and the sequence is read off the radioautograph as in the Sanger and Coulson method.[11]

Purine Cleavage

Dimethyl sulphate methylates guanine at the N7 position and adenine at the N3.[71] The methylation weakens the glycosidic bond, which breaks on heating: alkali (0.1*M* NaOH) then cleaves the sugar-phosphate bonds. Since guanine methylates five times faster than adenine, the reaction mixture, when run on a polyacrylamide gel, shows dark bands where guanine has reacted and light bands where there was an adenine in the sequence. There is not always as marked a difference in intensities as expected; hence, in another reaction mix, the methylated DNA is treated with dilute acid. The glycosidic bond of methylated guanosine, and the adenines are preferentially released, giving darker bands than those of guanines. Comparison of the two sets of conditions enables the sequence to be read. It is often advantageous to include alternative reaction mixes described by Maxam and Gilbert.[12] To cleave specifically at G residues, the methylated DNA is heated in 1*M* piperidine, when the 7-methylguanine ring is opened adjacent to the glycosidic bond and the base displaced from the sugar, which is then cleaved from the phosphates. The alternative adenine (and weak cytosine) cleavage involves heating end-labeled DNA in strong alkali. This opens the adenine and cytosine rings; piperidine then displaces the ring-opened bases and eliminates the phosphates.

Pyrimidine Cleavage

Hydrazine is used to cleave thymine and cytosine,[72] treatment with piperidine then displaces the products of hydrazinolysis from the sugars and catalyses cleavage of the phosphate bonds. The reaction is much the same for thymine and cytosine; however, in 2*M* NaCl the reaction of thymine is suppressed and bands are obtained only from cytosine. Therefore, the complete sequence can be read from four slots on the gel: G>A; A>G; C+T; and C, although the alternative reactions G and A>C are often useful.

The chemical method is less sensitive to any secondary structure in the DNA. Hence, the problems with the enzymic method of Sanger and Coulson[11] in interpreting sequences through base-paired regions do not occur.

Interpretation of the Sequence

One distinct advantage of this method over the plus-minus system is that every band in a run can usually be seen, and the bands in each reaction mix are more or less even in intensity on the radioautograph of the gel. Therefore, the sequences are easier to read than from plus-minus gels, and measuring distances between bands is not standard practice. This can lead to errors if the sequence is not confirmed by other methods. In experiments where bands are faint, one can be missed if the distances between the bands are not taken into account. Where a nucleotide is methylated (for example the 5-methylcytosine in the EcoRII site),[33] there is a gap in the gel bands which may not be noticed.[6,9] The other reason that nucleotides are omitted from a sequence is that the resolution between bands in a run, especially towards the top of the gel, may be poor and the sequence misread.[5,73] Other difficulties with the gel can be avoided. For example, in some gels, the pyrimidine bands are retarded with respect to the purines,[74-76] due to incomplete removal of the hydrazinolysis reagents and by-products. The gel system is highly sensitive to salt. Liu et al.[77] reported losing the first 21 nucleotides when the gel was not prerun — the salt effect caused a heavy pileup at the bottom of the gel.

Maxam and Gilbert[12] have described the sequencing protocol using 20% acrylamide gels in the presence of 8 *M* urea to separate the products; almost all users of this method follow suit. The description of the sequencing technique states that "the method is limited only by the resolving power of the polyacrylamide gel."

To increase this resolving power, the author recommends using 8 to 12% acrylamide gels on which sequences from primed synthesis can be read out to 300 nucleotides.[78] Sequences reported from the 20% gels, even those run for over 48 hr, rarely exceed 100 nucleotides. In fact, Humayun[74] and Porter[79] have used 15% polyacrylamide gels and Landy and Ross[80] 16% gels apparently without ill effects, although the sequences reported were no longer than those from 20% gels.

Figure 2 shows a sequence determination of 99 nucleotides from the *trp* operon of *E. coli*.[6] The 5′-^{32}P label was at a HpaII cleavage site at position 261; the sequence read from the three gels, run for 8.5, 26, and 50.5 hr, extends back from residue 262 to residue 161. It is clear that the sequence extends further, but the resolution between bands is becoming poorer. Lee et al.[6] wisely used another fragment for the region at the top of the 50-hr gel. This is a very good demonstration of the Maxam-Gilbert method, and the sequence obtained agreed with that obtained by sequencing an RNA transcript through the region.

The results from a single experiment are not always so clear-cut. Several groups have reported difficulty in distinguishing As and Gs,[10,77,81,82] and the alternative G and A>C cleavages are often necessary in reading the sequence. Scherer et al.[73] reported that there was consistent ambiguity in G and A identification in two positions. Sequence heterogeneity was ruled out because the corresponding pyrimidine in the complementary strand gave a clear sequence. At times the A cleavages are very weak, but reactivity may be increased by sequencing single strands or denaturing the double-stranded fragments.[83,84] The pyrimidines can also give ambiguous results, especially when Ts appear in the C reaction[10,85-87] and bands have appeared in both C + T and A + G mixes.[62] The resolution of tracts of purines or pyrimidines is occasionally not good enough to be certain of the sequence.[5,73,89]

Obviously, there are no universal problems in using the Maxam-Gilbert sequencing procedure. However, there are enough ambiguities in individual cases to warrant some checking of the sequences obtained. Almost all reported sequences have been confirmed to some extent. The sequence analysis of both complementary strands of a fragment is a valuable check; not only have all the purines become pyrimidines and vice versa, but the bands at the top of the sequence gel of one strand, where resolution

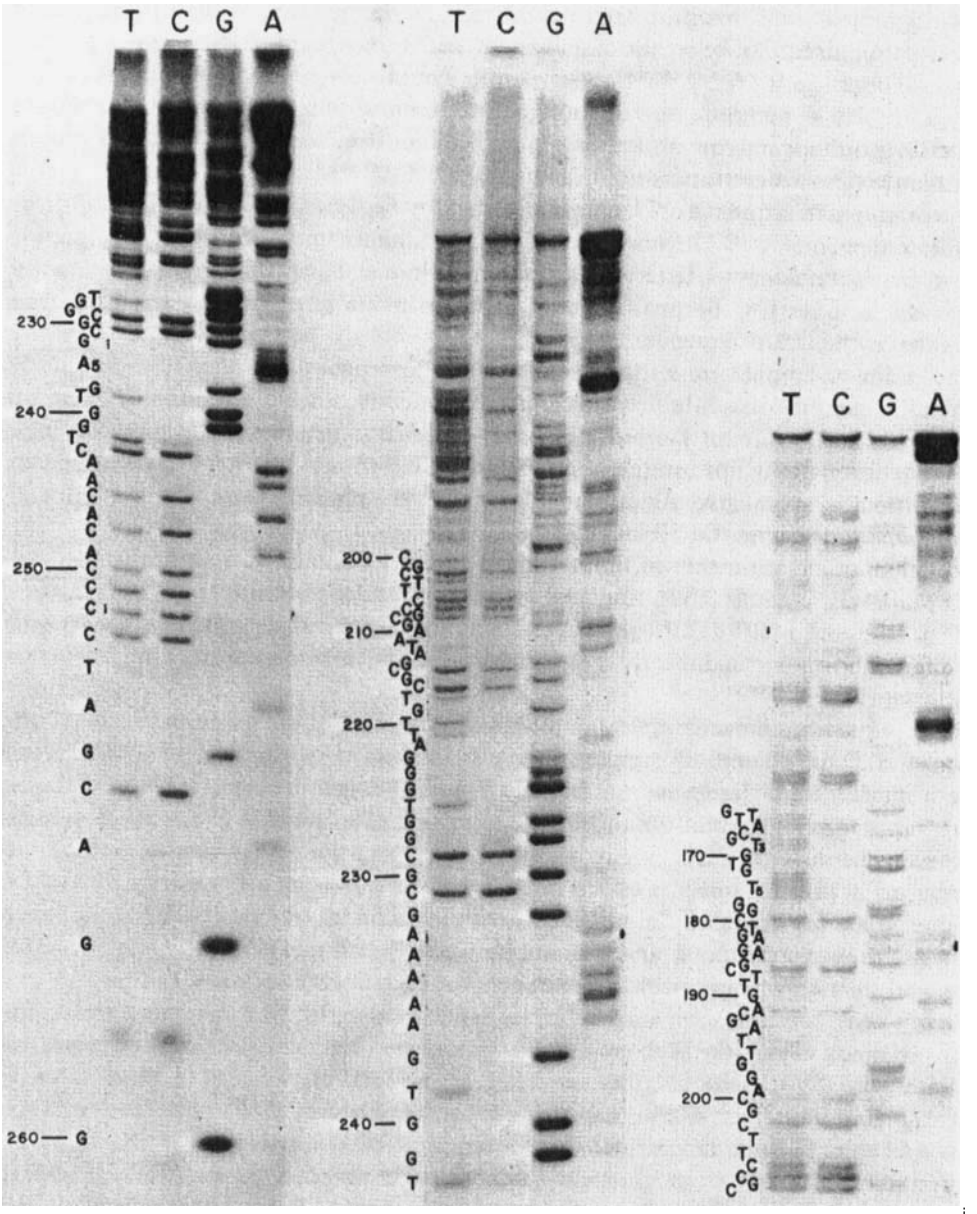


FIGURE 2. DNA sequencing by the Maxam-Gilbert method. A restriction fragment labeled at position +261 on the complementary strand (numbering from the initiation of transcription) of the *trp* operon of *E. coli*. The three gels were run for 8.5, 26, and 50.5 hr, and the sequence can be read for 99 nucleotides from +260 to +161. (With permission from Lee, F., Bertrand, K., Bennett, H. G., and Yanofsky, C., *J. Mol. Biol.*, in press. Copyright by Academic Press, Inc. (London) Ltd.)

may be doubtful, are at the bottom of the gel of the complementary strand. The analysis requires the presence of restriction sites approximately every 100 nucleotides through the region of DNA in question. Although the numbers of known restriction enzyme recognition sequences have accumulated rapidly,[33] there are still fragments too long to sequence both strands entirely. Overlapping, if not complete, sequences have been obtained in much of the sequence work to date.[10,68,69,73,76,80-83,89-91]

In many cases, RNA sequence data was available for comparison. Frequently, the

RNA oligonucleotides had not been completely overlapped due to the immense time and effort required; however, the data was sufficient to check the DNA sequence from Maxam-Gilbert gels.[5,76,77,85,88,87,92] Few people would now undertake a full sequence analysis by RNA methods, but a catalogue of compositions of, for example, ribonuclease T1 products may be all that is required to confirm a sequence.[82] Partial venom phosphodiesterase digestion is still often used.[74,76,82,85,87,88,92]

The amino acid sequence of the protein coded by the DNA is also useful in confirming the sequence.[9,63,74,77] However, for reasons explained in the introduction, there is now a strong tendency to determine a protein amino acid sequence by sequencing the DNA which codes for the protein, and mistakes in the protein sequence have been corrected by the DNA sequence.[9]

One point to emphasize is that the Maxam-Gilbert procedure does not give the 5′ terminal nucleotide of a 5′-labeled fragment. Frequently, this base is inferred from the known recognition site of the restriction enzyme used to generate the fragment. When the recognition site is not unique (e.g., HindII GTPy↓PuAC),[93] the 5′ nucleotide can be identified by exhaustive digestion with venom phosphodiesterase, when only the 5′ mononucleotide carries the ^{32}P label.[94]

When complete sequences of both strands cannot be obtained, quantitative depurination analysis[38] of the DNA fragment labeled by "nick-translation,"[95] in turn with α-^{32}P dATP and α-^{32}P dGTP, preferably using the separated strands, is a speedy and valuable addition to the data. If both strands are depurinated separately, the sequence is fully confirmed.[10,89,91]

Some sequencers have complete confidence in the data obtained from the gel. Contreras et al.[62] considered all supplementary RNA data to be "redundant." There are several nucleotide differences in a region of the SV40 genome reported by Contreras et al.[62] as compared to that obtained in Weissman's laboratory.[82] There are no amino acid sequence data for most of the proteins as yet, and the reader can be excused for wondering if the European SV40 is really so different from the American SV40 or whether there are errors in the sequence analyses. Modern methods of DNA sequence analysis are extremely rapid, and it is not particularly arduous to check every sequence by one of the many means available, however unnecessary it may seem.

In any case, an impressive amount of sequence data, most of it fully confirmed, has been obtained using the Maxam-Gilbert procedure, frequently after long years of struggle with RNA methods. The sequence of the 5200 nucleotides in SV40 DNA is virtually complete,[3] as is the genome of bacteriophage fd[2] (~6400 nucleotides). Schwartz et al.[68] have determined the sequence of 1000 nucleotides in the DNA of bacteriophage lambda from the rightward promoter through the *cro* and *cII* coding sequences and into the O gene protein-coding sequence. From another part of the lambda genome, Landy and Ross[80] have sequenced the regions involved in site-specific recombination (*att*), and the corresponding sites in the host (*E. coli*) DNA, and found that the two phage *att* sites and the two bacterial *att* sites have a common 15-base-pair sequence. Two eukaryotic mRNA species — rat insulin[9] and rabbit β-globin[10] — have been sequenced by making a double-stranded DNA copy with reverse transcriptase and inserting the DNA into a bacterial plasmid. After replication, the DNA is present in sufficient copies to be excised from the plasmid and sequenced.

DNA cloning techniques have also assisted the sequencing of the *trp* operon of *E. coli* and, for comparison, of *Salmonella typhimurium*.[5,6,96] Segments of the sequence were available for insertion into plasmids, amplification, excision, and sequencing. The sequences available from both organisms at the time of writing extend from about −115 to +260, the sequence being numbered from the nucleotide which initiates *trp* mRNA.[5,6] The sequence has given some insight into how the attenuator functions under the control of $tRNA_{trp}$, and promoter and other mutants have been sequenced to

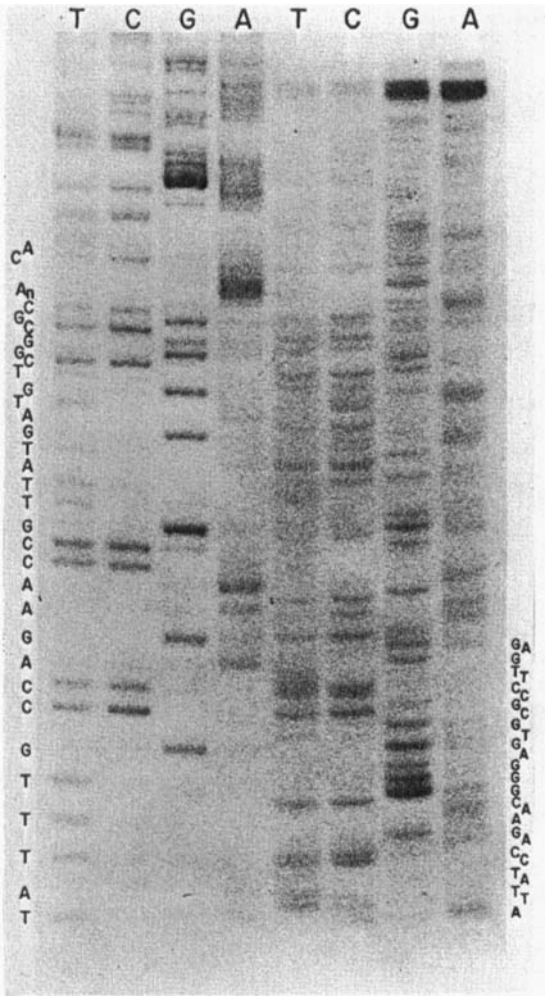


FIGURE 3. Maxam-Gilbert sequencing of the promoter-operator region of the *trp* operon of *E. coli*. The *Hinf*I-*Hinc*II fragment corresponding to positions −280 to −35 was ^{32}P-labeled at the *Hinc*II end. The electrophoresis was carried out at 600 V for 25 hr in the left series and in the right series for 50 hr. The sequence deduced in each case is noted along the side. The shorter run gives information on the region from −52 to −88 and the more extended run provides the sequence from −88 to −115. (With permission from Bennett, G. N., Schweingruber, M. E., Brown, K. D., Squires, C., and Yanofsky, C., *J. Mol. Biol.*, in press. Copyright by Academic Press, Inc. (London) Ltd.)

help identify which nucleotides are involved in the complex control systems operating in the *trp* operon. Figures 3 and 4 show some of the Maxam-Gilbert gel sequences. The sequences read from these five gels run continuously from nucleotide −115 to +24. This sequence is shown in Figure 5. Also indicated in Figure 5 are the ribonuclease products used to derive the same sequence before the rapid techniques were available. There were 79 oligonucleotides isolated and sequenced, since several overlapping sets have to be obtained before the sequence can be put together. The 79 oligonucleotides can be replaced by five polyacrylamide gels from three experiments, which nicely illustrates the power of the new methods.

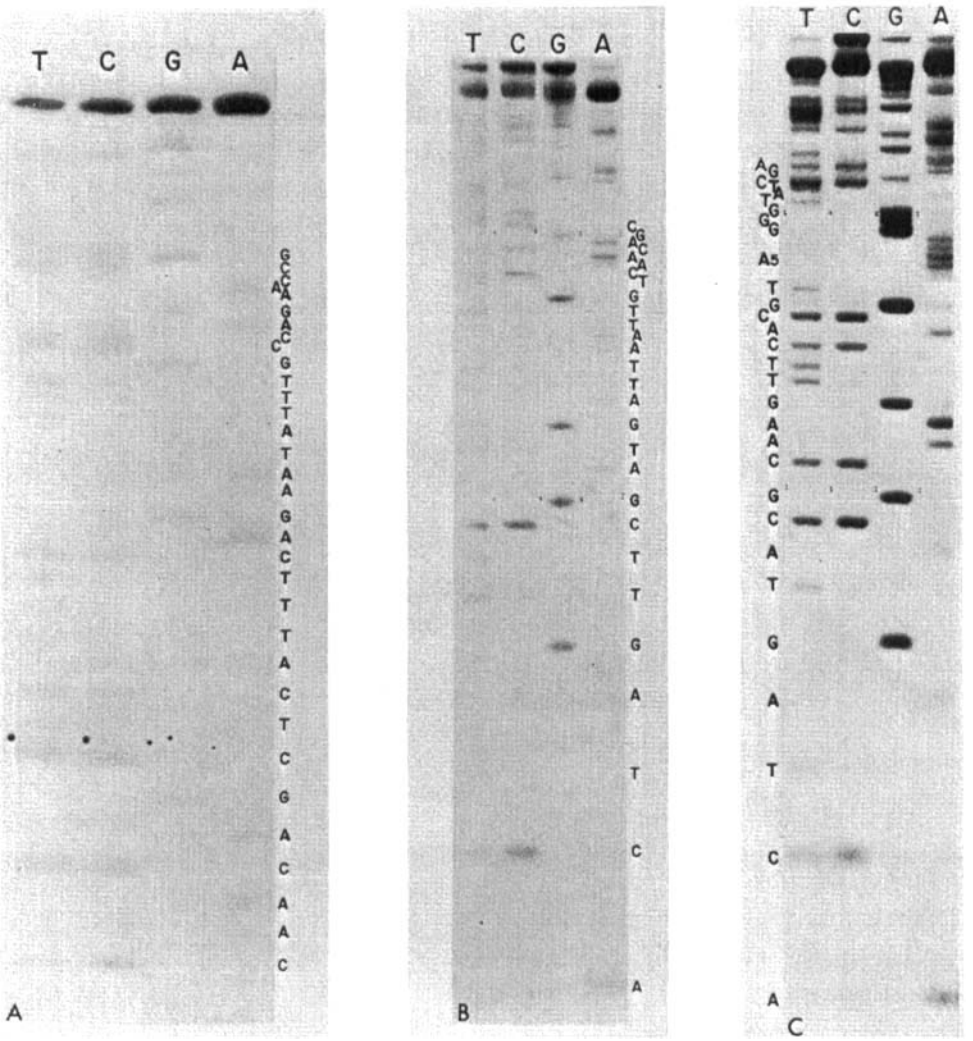


FIGURE 4. More DNA sequence of the *trp* operon of *E. coli*. DNA sequencing gel of the *Hha*I-*Hpa*I fragment corresponding to positions −78 to −12 (A and B) and the *Hpa*I-*Hha*I fragment corresponding to positions −11 to + 61 (C). The fragments were ^{32}P-labeled at the *Hpa*I end. The gel was run at 600 V for 9 hr in (B) and (C) and 24 hr in (A). The sequence deduced in each case is lettered along the side. Gel A corresponds to base pairs −34 to −66 and gel B shows the sequence of the region from −13 to −40. The position of the bromphenol marker dye is noted by the dots near the G in (B) and has a mobility equivalent to DNA of approximate length 10. The upper row of dots in (B) indicates the position of the xylene cyanol FF marker dye near the upper GC in (B) and near this same sequence in the longer electrophoresis in (A). Gel C shows the sequence from −10 to +24. (With permission from Bennett, G. N., Schweingruber, M. E., Brown, K. D., Squires, C., and Yanofsky, C., *J. Mol. Biol.*, in press. Copyright by Academic Press, Inc. (London) Ltd.)

SEQUENCING RNA BY PRIMED SYNTHESIS OF DNA

All nucleotide sequencing methods currently in use depend entirely on the introduction of a radioactive label, usually ^{32}P. Since it is difficult to obtain in vivo labeling of a high enough radioactive yield for sequencing, the label can be introduced in vitro by copying reactions. When sequencing RNA, copying is conveniently done by reverse transcriptase (usually from avian myeloblastosis virus (AMV), see Table 1). An added advantage is that the synthesized copy is more stable than the RNA template, especially

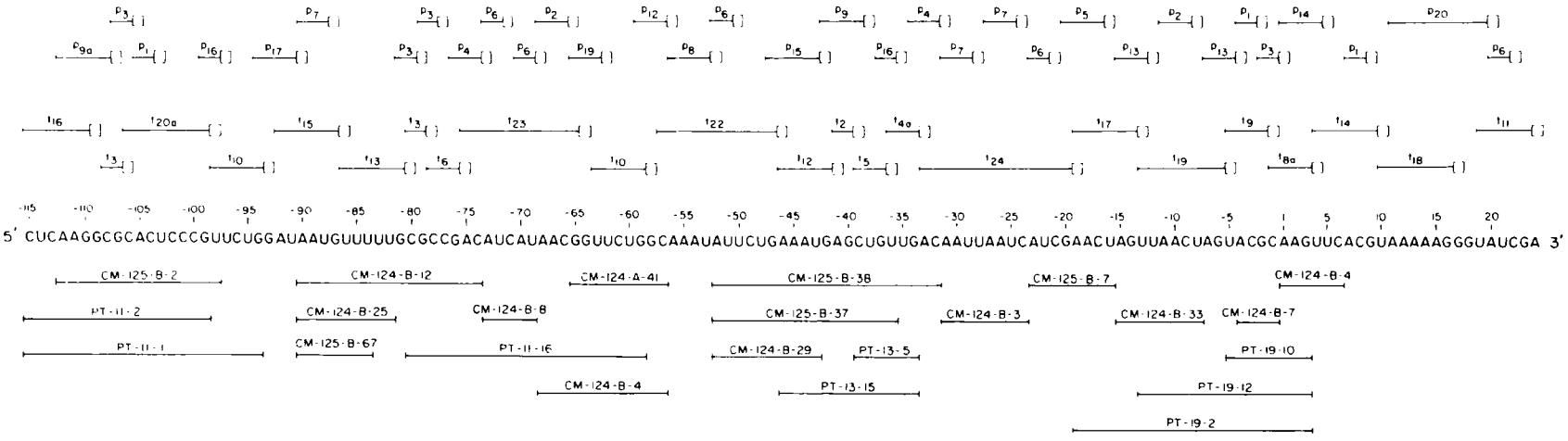


FIGURE 5. The RNA sequence corresponding to the DNA sequence shown in Figures 3 and 4, with the oligonucleotides used to derive the RNA sequence. Products of pancreatic and T1 ribonucleases are shwon above the sequence, while partial products are shown below, derived from CM ribonuclease A or T1 ribonuclease. These 79 oligonucleotides required for sequencing by RNA techniques were replaced by 5 gel runs of 3 experiments, as shown in Figures 3 and 4. (With permission from Bennett, G. N., Schweingruber, M. E., Brown, K. D., Squires, C., and Yanofsky, C., *J. Mol. Biol.*, in press. Copyright by Academic Press, Inc. (London) Ltd.)

since most deoxyribonucleases are inhibited by EDTA. Most ribonucleases are not dependent on divalent metal ions and are difficult to inhibit.

If the RNA template is polyadenylated at the 3′ end, synthesis can be initiated from an oligo-dT primer. The plus-minus and Maxam-Gilbert methods both require a fixed 5′ terminus. Attempts have been made to obtain this by limiting the concentration of dTTP in the reaction mix so that only those primers at the 5′ end of the poly-A tract will elongate. This procedure has been partially successful,[26,86,121] however, it is better to use oligo-dT which has been elongated by one or two nucleotides which are complementary to those immediately adjacent to the poly-A in the RNA. Primers which have been used in mRNA sequence work are $d(pT_{10}-G-C_{OH})$[26,97,98] and $d(pT_{10}-C-A_{OH})$.[27,99] Since the mRNA preparations are usually not 100% pure, the primer with two complementary nucleotides aids in excluding the transcription of unwanted sequences. When the RNA is highly purified, the primers $d(pT_8-C)$[79,100] and $d(pT_8-A)$[100,102] have been successfully used. The primer, $dT(pdT)_8prG$, was used in obtaining a sequence at the 3′ end of Rous sarcoma virus by spleen and venom phosphodiesterases.[101] The presence of the ribonucleotide enabled the primer to be easily cleaved from the synthesized DNA before exonuclease digestion.

If the RNA does not have a poly-A tail, it can be readily added by the *E. coli* enzyme adenosine triphosphate: ribonucleic acid adenyltransferase (see Table 1), and DNA synthesis initiated from the $d(pT_nN)$ type of primer. This has been used in sequencing the 3′ regions of RNA segments of influenza virus.[102] Since these segments are complementary to the mRNA, the cDNA sequence obtained is from the 5′ terminal (protein initiating) region of the mRNA. Ribosomal RNA sequences from the 3′ region have been obtained by polyadenylation and primed cDNA synthesis.[100]

Once the sequence from the 3′ region is established by priming with $d(pT_nN)$, there is a marked slowing down in the sequence work because of the problem in obtaining suitable specific primers. Where appropriate containment facilities are available, the best approach is to synthesize double-stranded DNA from the mRNA and insert it into a plasmid or phage vector. The DNA can then be amplified enough to isolate restriction fragments which can be sequenced directly by the Maxam- Gilbert method[9,10] or used as primers on either the mRNA or the cloned, separated DNA strands for sequencing by the Sanger methods.[63]

Where cloning is not permissible, or if the sequence of a particular limited region is required, oligonucleotide primers can be synthesized. The basic chemical technology has been improved and simplified by Khorana et al.,[103,104] using the phosphodiester approach, and Itakura et al.[105] and Bahl et al.,[106] using phosphotriester methods, and can be partially automated on a solid support.[107] However, the procedures are still complex enough to be the rate-limiting step in a sequence analysis. The final yield of primer is low, and the product must be carefully analysed to ensure it is correct. As an alternative to chemical synthesis, Gillam and Smith,[108] Gillam et al.,[109] and Gillam et al.[110] have used polynucleotide phosphorylase from *E. coli* (Table 1). Under certain conditions, the enzyme will add short tracts of ribo- or deoxyribonucleotides to a short primer. The products are then fractionated by size to isolate the one required. Certain sequences can be synthesized relatively easily by this method, although the wastage can be considerable. To prime efficiently and specifically, the oligonucleotide must generally be at the least approximately 8 nucleotides long. Shorter primers have been successfully used[111] but the sequence of a hepta- or hexanucleotide has a greater chance of not being unique in the template. It is not clear what constitutes a "good" primer, apart from its being specific. For instance, there is no clear relationship between length and base composition.[111,112] Zimmern[131] has had some success in finding internal primers by ribonuclease plus phosphatase digestion of any (but not too long) other RNA molecule. Most products are too short to prime, and with luck, efficient and specific primers will be present.

The Plus and Minus Method Using an RNA Template

The basic principles of the plus and minus method apply also to the use of an RNA template; however, there are several differences in the detailed protocol. Brownlee and Cartwright[97] have developed the method and obtained a sequence of 110 nucleotides preceding the poly-A tract of ovalbumin mRNA with a tentative sequence extending to 170 nucleotides. This protocol has been used in sequencing the 3′ terminal noncoding regions of rabbit and human β-globin mRNA,[113] where sequences could be read up to 140 nucleotides from the d($pT_{10}GC$) primer. Since there were several uncertain nucleotides, the sequence was confirmed by other techniques. Similarly, a 53 nucleotide sequence from the m^7GpppA^m cap at the 5′ end of rabbit β-globin mRNA to the initiating A-U-G codon was determined using a synthetic deoxyoctanucleotide primer complementary to the ribosome-protected sequence which includes the initiating A-U-G codon.[114] In the 3′ noncoding region of mouse immunoglobin light-chain mRNA, a sequence of 75 nucleotides was obtained[27] commencing 13 residues from the poly-A sequence; this was confirmed by depurination analysis and comparison with some T1 oligonucleotides which had been partly sequenced but not ordered.

Bernard et al.[99] slightly modified the Brownlee-Cartwright[97] protocol and, in conjunction with data from other methods, reported a sequence of 105 nucleotides of mouse immunoglobulin kappa chain mRNA. The authors considered the first 59 residues from the poly-A as established while the rest of the sequence was not fully confirmed. As far as nucleotide 60, the sequence is identical to that obtained by Hamlyn et al.[27] from a different mouse kappa chain mRNA. Beyond residue 60, the sequences are somewhat different, although both contain an exceptionally long run of pyrimidines — 24 in the confirmed sequences.[27] It appears, therefore, that there may be more than one kappa constant region in the mouse.

Figure 6 shows a radioautograph of a plus-minus gel from this kappa chain 3′ terminal sequence analysis.[99] The initial staggered synthesis of complementary DNA (cDNA) was produced by AMV reverse transcriptase incorporating one α-^{32}P-labeled dNTP. Alternatively, the extension can be carried out with *E. coli* DNA polymerase I in the presence of Mn^{2+} (see Table 1), although there is a danger that the 3′ exonuclease activity may destroy the phasing of a d(pT_nN)-type primer. This enzyme copies the RNA template correctly, but the products are somewhat shorter than the optimum range for plus-minus sequencing.[26,98,115] One critical step in the procedure is that, before the plus and minus reactions, the cDNA must be reannealed to fresh RNA template, presumably because the original template is degraded by the RNAse H activity intrinsic to AMV reverse transcriptase (see Table 1). The minus reactions involve a further extension of the cDNA products with AMV reverse transcriptase and three dNTPs. With a DNA template,[11] the plus reactions contained one dNTP and T4 DNA polymerase. On an RNA template, the 3′ exonuclease activity of T4 polymerase works, but the polymerization does not.[97] Therefore, the Klenow fragment of *E. coli* DNA polymerase is used (see Table 1). In this fragment, the 5′ exonuclease activity is lost, but at high pH there is 3′ exonuclease activity and polymerizing activity (Table 1), as required for the plus reactions. The conditions for the reaction are critical, and the plus results are variable.[97,99,100]

Radioautographs of plus-minus sequencing gels using RNAs which are not naturally polyadenylated are shown in Figures 7 and 8. The RNAs were purified and adenylated in vitro. Figure 7 shows a plus-minus sequence from an influenza viral RNA segment. The 3′ terminus was confirmed as U[116] by using the three primers d(T_8A), d(T_8G), and d(T_8C) in turn. Only the first gave specific priming. The 3′ ends of various ribosomal RNAs were similarly confirmed by checking the primer specificity. The primers d(T_8C) and d(T_8A), respectively, were used for synthesizing the cDNA copy of wheat 18S RNA and *E. coli* 16S RNA. The subsequent minus reactions are shown in Figure 8.

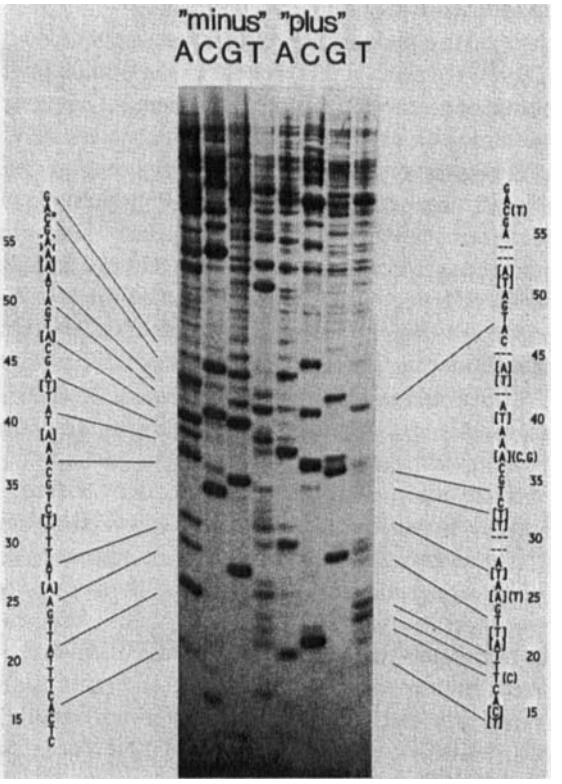


FIGURE 6. Fractionation of immunoglobulin kappa 41A cDNA products subjected to the minus and plus reactions. In this short run, a 15% polyacrylamide gel was run at 600 V for 8 hr; the sequence on the left is that read using the minus system, while the sequence on the right is that read from the plus system. Solid brackets enclose residues which cannot be seen clearly in this gel but were detected on others; broken brackets enclose residues not detected by the gel method but known from endonuclease IV products. Residues where the gel method leaves an ambiguity are shown within parentheses. The asterisk at position 57 indicates where two C residues are required from the analysis of an endonuclease IV product. Bernard, O. D., Jackson, J., Cory, S., and Adams, J. M., *Biochemistry*, 16, 4117, 1977. Copyright by the American Chemical Society.

A problem is immediately obvious in Figure 8: the AMV reverse transcriptase cannot read beyond about 23 *(E. coli)* or 20 (wheat) nucleotides. The 3′-terminal 50 bases of *E. coli* 16S rRNA have been sequenced[117] and can be illustrated as a hairpin structure with $N_6{}^2mAN_6{}^2mA$ at the top. It is most likely that cDNA synthesis stops at the methylated structures (base 23) in the *E. coli* 16S rRNA template. Wheat 18S rRNA also probably contains the $N_6{}^2mAN_6{}^2mA$ dinucleotide,[130] by analogy with *E. coli* at bases 20 and 21 from the 3′ end. Often, bands can be seen preceding the stop in the *E. coli* 16s RNA sequencing gels, but they do not correspond to the known sequence,[117] possibly due to the secondary structure causing compression of the bands.

Sequencing cDNA by the Maxam-Gilbert Method

With the same requirement for a fixed terminus, which can then be labeled with ^{32}P, a cDNA copy of an RNA template can be sequenced by the Maxam-Gilbert procedure.

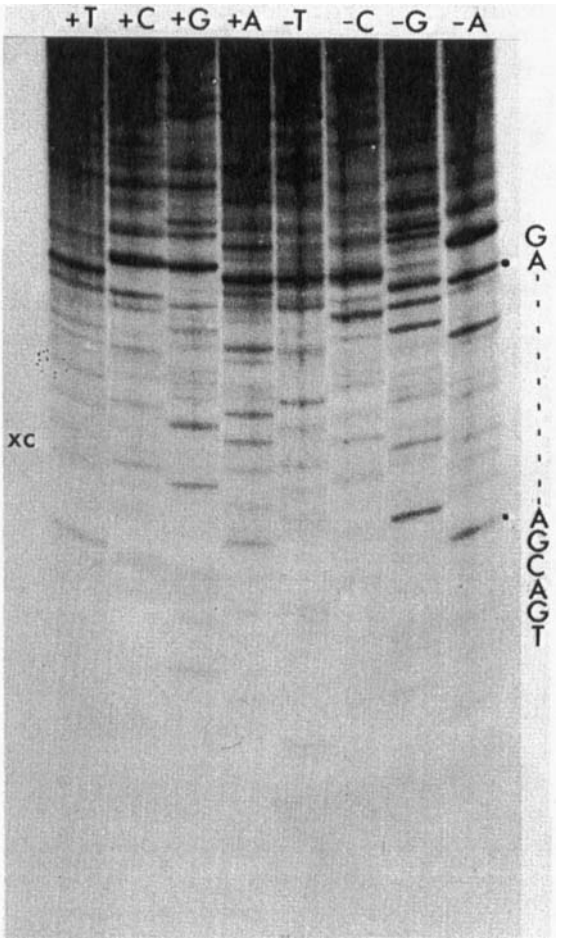


FIGURE 7. Plus and minus sequence of the 3′ terminal region of the matrix protein RNA of influenza virus (strain NWS/Ned, H_0N_2). Since the virion RNA segment is of the opposite sense to the mRNA, the sequence of the cRNA shown corresponds to that near the 5′ end of the mRNA. The sequence can be read from approximately 40 nucleotides from the A of the primer d(pT_6A) for 46 nucleotides in the minus slots; the plus reactions are patchy, with some obvious artifact bands. Amino acid sequence analysis of the matrix protein would confirm if the single reading frame which does not contain terminator codons is part of the coding region of this gene. (From Air, G. M., and Both, G. W., unpublished data, 1977. With permission.)

In several RNA tumor viruses, cDNA synthesis is initiated in vivo from a host tRNA molecule which primes from a region 100 or so nucleotides in from the 5′ terminus of the viral RNA.[86,94,118] In vitro, the major product of AMV reverse transcription of these RNAs extends from the primer to the 5′ end of the RNA and is an ideal length for sequencing after removing the tRNA primer by alkaline hydrolysis and labeling the 5′ end of the cDNA with γ-^{32}P-ATP and polynucleotide kinase (see Table 1). Haseltine et al.[118] obtained the sequence of 101 nucleotides from the $tRNA_{trp}$ primer to the 5′ end of Rous sarcoma virus cDNA. The corresponding sequence in an avian sarcoma virus RNA, also primed by $tRNA_{trp}$, is identical.[94] The sequence of the 5′

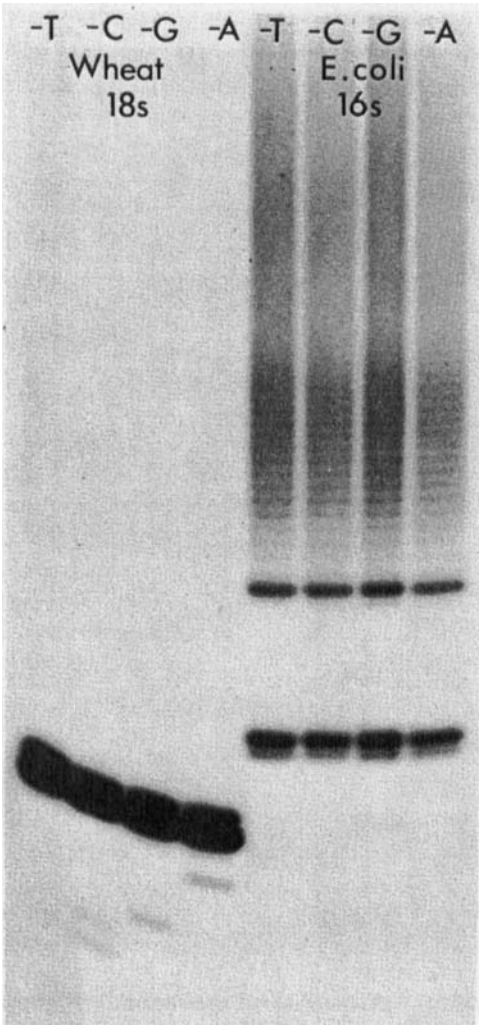


FIGURE 8. Plus and minus sequencing applied to 3′ ends of ribosomal RNAs. Polyadenylated wheat 18S and *E. coli* 16S ribosomal RNAs were trascribed by AMV reverse transcriptase using d(pT$_8$C) and d(pT$_8$A), respectively, as primers. As described in the text, synthesis has stopped at particular nucleotides in the sequence; presumably the m$_6^2$A residues in *E. coli* 16S[117] and probably present in wheat 18S[130] RNAs cannot be used as template by the enzyme. The bands which can be seen before the stop do not correspond to the known sequence of wheat 18S ribosomal RNA, perhaps because of the extended base-paired hairpin loop. (From Both, G. W., unpublished results. With permission.)

cDNA from avian myeloblastosis virus differs at a few nucleotides.[86] Preliminary experiments with the AMV cDNA indicated the recognition sequence for the restriction enzyme HaeIII (G–G–C–C) about 50 residues from the 5′ end. Endonuclease HaeIII cleaves single-stranded DNA,[119] thus, the cDNA was digested with HaeIII and the products separated and sequenced. Therefore, the sequence read near the top of the original sequencing gel, where the resolution is often not perfect, has been confirmed by sequencing a shorter product.[86]

In vivo, cDNA synthesis does not stop at the 5′ end of the RNA, and a mechanism for copying the rest of the template was proposed which required terminal redundancy, i.e., identical sequences at the 5′ and 3′ ends. Some evidence that the ends were redundant was obtained; for example, by demonstrating that the ribonuclease T1 product which was attached to the poly-A sequence at the 3′ end hybridized to the 5′ cDNA copy.[120] Stoll et al.[86] obtained the 3′ sequence of avian myeloblastosis virus RNA using $(pdT)_{13}$, labeled at the 5′ end with γ-^{32}P-ATP and polynucleotide kinase, as primer with a very low concentration of dTTP to encourage synthesis from the 5′ end of the poly-A tract. This strategy is partially successful, and the major products were purified by polyacrylamide gel electrophoresis and sequenced by the Maxam-Gilbert method. That the 3′ end was heterogeneous was an unexpected complication, since an A−C−C sequence preceding the poly-A was missing in some molecules. Hence, the bands of cDNA which were eluted from the preparative gel at particular sizes were not pure, and the sequencing gels had a background of other sequences. However, the major sequence read from each was the same as the sequence obtained from the T1 RNAse product which was attached to the poly-A. A sequence of 16 or 19 residues at the 3′ end was identical to the sequence at the 5′ end which followed the inverted 7MeG residue of the "cap."

The sequence adjacent to the short (about 38 nucleotides) poly-A sequence at the 3′ end of encephalomyocarditis virus RNA was established by limited exonuclease digestion.[121] This has been confirmed by the Maxam-Gilbert method, sequencing the cDNA synthesized from ^{32}P-labeled dT_8dC by *E. coli* DNA polymerase I in the presence of Mn^{2+}.[79] In this case, the products were separated on a 15% acrylamide gel, and the sequence on a single gel run was read to 27 nucleotides as shown in Figure 9.

A method of directly sequencing RNA by a gel method has also been described.[122] The RNA is 5′-end labeled with polynucleotide kinase and γ-^{32}P ATP. In urea, at 50°C, ribonuclease T1 produces partial cleavage at guanine residues, and ribonuclease U2 produces partial cleavage specifically at adenines. Limited alkaline hydrolysis[122] or heating in formamide[134] produces partial products from hydrolysis at any phosphodiester bond. Hence, four reaction mixes are run on a 20% polyacrylamide gel under denaturing conditions. In the ribonuclease T1 digest, the G residues are identified, and the ribonuclease U2 digest shows the A residues. Since all nucleotides in the sequence are displayed in the alkali-treated sample, those which are not G or A must be pyrimidines. Cytosine and thymine can be distinguished using RNAse I from *Psysarum polycephalum,* which cleaves at A, G, and U.[134] The sequencing protocol was developed using 5′-labeled yeast 5.8S ribosomal RNA[122] and tyrosine tRNA.[134] The method could also be applied to 3′-end-labeled RNA, although 3′ labeling of RNA is difficult to achieve. The technique could be very useful for comparing 5′ ends of RNA molecules, but the lack of enzymes (equivalent to the DNA-specific restriction endonucleases) to generate suitable fragments (100 to 200 nucleotides long) and the difficulties in working with RNA compared with the relative stability of DNA indicate that it is not the method of choice for extensive sequence analysis.

SEQUENCING DNA WITH CHAIN-TERMINATING INHIBITORS

Sanger et al.[78] have recently developed a DNA sequencing method which is considerably simpler than either the plus-minus or the Maxam-Gilbert techniques. This method relies on primed synthesis of DNA, incorporating a ^{32}P dNTP and also including an abnormal nucleoside triphosphate which is incorporated but which terminates DNA synthesis. The terminators which have been used are 2′3′-dideoxy derivatives of thymidine, cytidine, adenosine, and guanosine triphosphates. The four arabinoside triphosphates can also be used when *E. coli* DNA polymerase I is the synthesizing enzyme and low temperature is used. Mammalian DNA polymerases and the *E. coli*

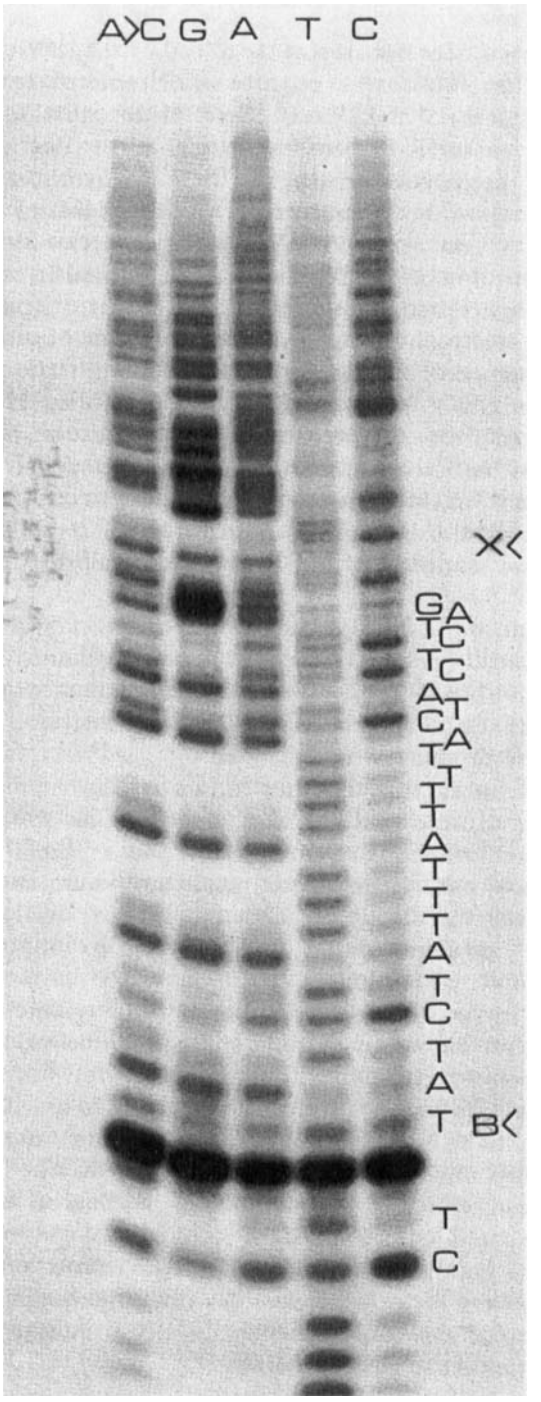


FIGURE 9. Maxam-Gilbert sequencing of cDNA. The sequence shown corresponds to the 3′ terminal sequence of the RNA of encephalomyocarditis virus which had been determined by other methods.[121] The primer used was 5′-end-labeled d(pT$_4$C) and extension was with *E. coli* DNA polymerase I. The products of partial chemical cleavage were separated on a 15% polyacrylamide gel. The sequence can be read for 26 nucleotides starting at the C of the primer, but there is a marked artifact around the third nucleotide, which is, in fact, a C.[121] (From Porter, A. G., unpublished results. With permission.)

enzyme at 37°C (such as during removal of the primer by a restriction enzyme) may not terminate at the arabinotide; thus, the dideoxy derivatives are probably more reliable. The proportion of dideoxy or arabinoside to normal triphosphate is such that only partial incorporation of the terminator occurs.

The sample of template single-stranded DNA with annealed primer is divided into four, and each aliquot is incubated with the four deoxyribonucleoside triphosphates (one of which is α-^{32}P labeled), one of the four chain-terminating inhibitors, and DNA polymerase I (Klenow fragment, Table 1). After the incubation, a "cold" chase is given to avoid any termination due to the limiting concentration of the ^{32}P-labeled nucleoside triphosphate. Therefore, elongation of individual DNA molecules is stopped whenever an inhibitor molecule is incorporated. This happens randomly, so when the mix containing dideoxy-TTP, for instance, is run on a polyacrylamide gel, a radioactive band is seen at every position where there is a T, since a certain proportion of molecules have terminated there. The pattern of bands from the four reaction mixes then provides the sequence as in the plus-minus method. A significant advantage over the plus-minus method is that every band in a run of the same nucleotide appears, as in the Maxam-Gilbert procedure. In general, sequences from 15 to 200 nucleotides from the priming site can be determined in one experiment, and sequences have been read to about 300 nucleotides.[78] There are occasional artifact bands, not as many as on plus-minus gels and more readily eliminated. The most serious problem is compression of bands where the DNA forms stable base-paired loops. Sanger et al.[78] consider that the sequence obtained by this method, as well as the other rapid techniques, should be checked by some other method or by priming on the complementary strand.

Figure 10 shows two radioautographs of gels with sequences obtained by the chain-terminating method. The sequences shown include regions of ϕX 174 DNA which were considered tentative. Apart from some compressions of bands due to secondary structure (position 4330 in Figure 10A, and 3545 in Figure 10B), the sequences obtained by the chain-terminating methods are read without difficulty and show that there were some errors (about 6 nucleotides) in the previously unconfirmed sequence.[1]

The chain-terminating method has also been used to sequence several hundred nucleotides around the origins of synthesis of viral and complementary DNA of bacteriophage G4,[132] which is an isometric phage related to ϕX 174. The viral strand origin shows considerable homology to that of ϕX 174,[1] while the complementary strand origin, which does not occur in ϕX 174, shows some homology to the λ origin region.

SEQUENCING DNA BY PARTIAL RIBO-SUBSTITUTION

A similar principle was used by Barnes[123] to develop a sequencing method based on random incorporation of a ribonucleoside triphosphate instead of the deoxy derivative during a primed synthesis of DNA. At appropriate concentrations of dNTP and rNTP, a certain proportion of sites become occupied by the ribo nucleotide and are subsequently susceptible to alkali cleavage. The labeling has to be at the 5′ end, either by the kinase reaction (Table 1) or else by using a short synthesis with only deoxynucleoside triphosphates (one of which is α-^{32}P labeled) and then ribo-substituting in a second extension with cold triphosphates. The sequence is then read from a radioautograph of a polyacrylamide gel, as in the other rapid methods. All bands in a run can be seen. Examples of radioautographs of gels using this technique are shown in Figure 11.

The method is not as straightforward as the chain-terminating technique, involving more steps in the protocol; however, it has the advantage that the substrates are readily available. In common with the Maxam-Gilbert method, it can be used to investigate topology of protein binding to DNA, since the protein can protect certain nucleotides from alkali cleavage[123] or methylation.[12]

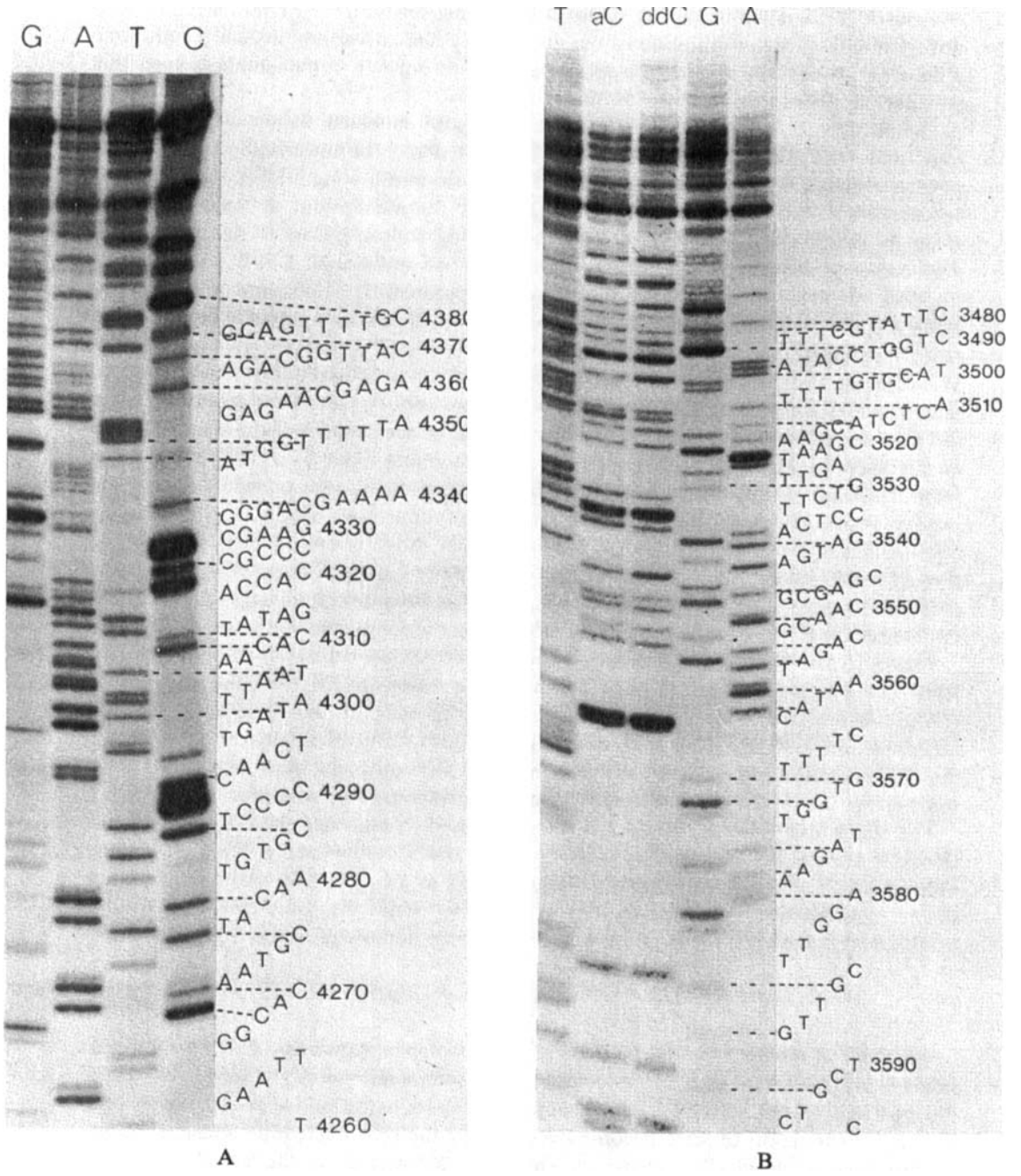


FIGURE 10. DNA sequencing with chain-terminating inhibitors. (A) Autoradiograph of an experiment using *Hind* II fragment 4 as primer on the complementary strand of ϕX 174. The numbering of the sequence is as given in Reference 1. The inhibitors used were dideoxy-(dd-)ATP, ddGTP, ddTTP, and ddCTP. Electrophoresis was on a 12% polyacrylamide gel for 14 hr at 40 mA. The sequence shown starts 66 nucleotides from the 3′ end of the primer. (B) Autoradiograph of an experiment with *Alu* I fragment 8 as primer on the viral strand of ϕX 174. The inhibitors used were ddATP, ddGTP, ddTTP, ddCTP, and the arabinoside of CTP (aC). The sequence shown starts 43 nucleotides from the 3′ end of the primer. (From Sanger, F., Nicklen, S., and Coulson, A. R., *Proc. Natl. Acad. Sci. U.S.A.*, 74, 5463, 1977.)

SEQUENCING DNA BY PHYSICAL METHODS

Beer and Moudrianakis[124] suggested that it might be possible to sequence DNA by electron microscopy. DNA is transparent to electrons: thus, if each of the four bases

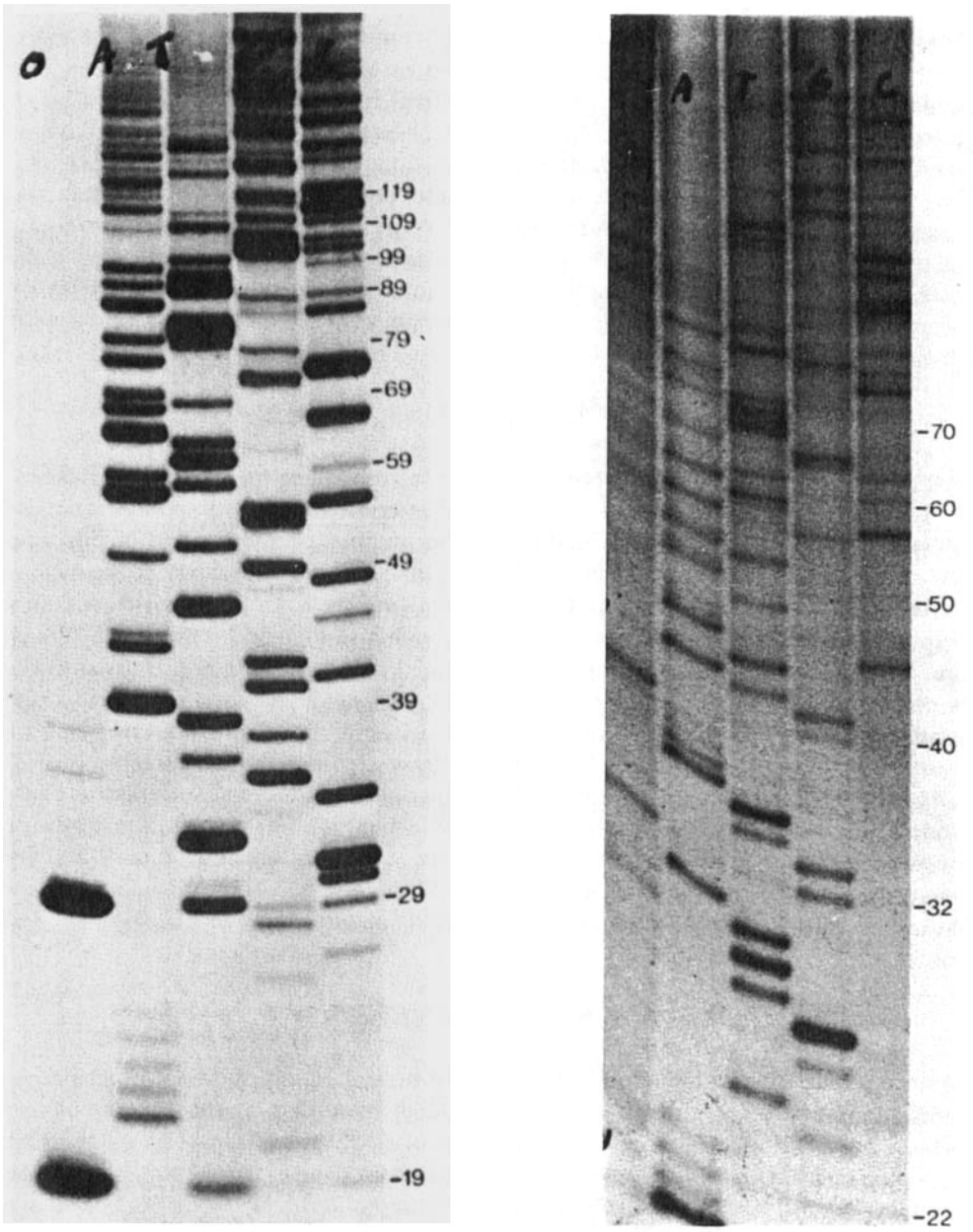


FIGURE 11A. Sequencing DNA by partial ribosubstitution. A. A sequence from ϕX 174 DNA with viral strand as template, *Hae*III fragment 3 as primer. The sequence can be read from nucleotide 19 to 112, the only difficulties being very faint G bands at positions 33, 47, 52, and 99, and a very strong C at 74, which obscures the C at position 73. The sequence read from this radioautograph agrees with that obtained by other methods.[1] B. A sequence from the *Salmonella typhimurium* histidine operon cloned in a mini-El plasmid.[123] The strands of the plasmid were separated and the H strand used as template with an *Hha*I fragment as primer. The sequence can be read from nucleotide 20 to 74. Photographs A and B courtesy of W. M. Barnes.

could be specifically coupled to electron-dense labels, such as heavy metals, the distribution of each label in the electron micrograph would give the sequence. Specific reactions have been found for two of the four bases,[125,126] however, to date, no real sequence analysis has been reported using electron microscopy.

Mass spectrometry can also be used to study oligonucleotides. Although the sensitivity is significantly lower than the radioactive tracer techniques, an advantage is that unusual or chemically modified bases are identified. Initial attempts to develop mass-spectrophotometric methods of sequencing nucleic acids were impeded by the requirement for volatile derivatives. Such derivatives resulted in mass values which, for anything larger than a tetraribonucleotide, were too high for analysis.[127] Wiebers and Shapiro[128] have developed a method in which the mass spectrometer itself cleaves intact underivatized oligonucleotides to fragments, which to some extent, are specific to the sequence. The relative ratios of particular ions can then be used to derive the sequence by computerized pattern-recognition techniques. While the technique is limited to short oligonucleotides, it should not be ignored if unusual bases are suspected to be present.

CONCLUSIONS

Although a compendium of DNA sequences has not been included in this review for obvious reasons of space as well as general interest, the areas in which extensive sequence analyses are currently available have been indicated and the sort of information on genetics or control processes which can be obtained from DNA sequencing is described. The ease of determining long sequences of DNA is still surprising to many nonspecialists, because the rapid methods were such a sudden and remarkable change from the older technologies (cf. References 34 to 36). If the need to confirm sequences has been emphasized to an extent that may be considered excessive, it is not to cast doubts on the new methods. Rather, they are so rapid that confirmation is not arduous. Many far-reaching conclusions can be drawn from a DNA sequence, such as overlapping genes and new mechanisms of replications, and it is clearly vital that there be no doubts as to the accuracy of the sequence when hypotheses resulting from the sequence are being considered. As long as DNA sequences which are not fully confirmed are indicated as such when they are published in scientific journals, the new methods will continue to be an amazingly powerful means of studying biological processes.

ACKNOWLEDGMENTS

I am indebted to the many colleagues who supplied unpublished data, gel photographs, and specific comments which have added immensely to the content of this review. I have included data available up to October 1977, and am very much aware that sequence information is accumulating faster than reveiws can be published. I am grateful to my colleagues here in Canberra, particularly Dr. G. W. Both, for comments on the manuscript. The work from this laboratory was with the expert technical assistance of Bronwyn Asquith.

REFERENCES

1. Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., Hutchison, C. A., III, Slocombe, P. M., and Smith, M., Nucleotide sequence of bacteriophage ϕX 174 DNA, *Nature* (London), 265, 687, 1977.
2. Beck, E., Auerswald, E.-A., Sommer, R., Kurz, C., Schaller, H., Sugimoto, K., Okamoto, T., and Takanami, M., unpublished data, 1977.
3. Weissman, S., unpublished data, 1977; Fiers, W., unpublished data, 1977.
4. Ptashne, M., Backman, K., Humayun, M. Z., Jeffrey, A., Maurer, R., Meyer, B., and Sauer, R. T., Autoregulation and function of a repressor in bacteriophage lambda, *Science*, 194, 156, 1976.

5. Bennett, G. N., Schweingruber, M. E., Brown, K. D., Squires, C., and Yanofsky, C., Nucleotide sequence of the promoter-operator region of the tryptophan operon of *Escherichia coli*, *J. Mol. Biol.*, 121, 113, 1978.
6. Lee, F., Bertrand, K., Bennett, G., and Yanofsky, C., Comparison of the nucleotide sequences of the initial transcribed regions of the tryptophan operons of *Escherichia coli* and *Salmonella typhimurium*, *J. Mol. Biol.*, 121, 193, 1978.
7. Valenzuela, P., Bell, G. I., Masiarz, F. R., De Gennaro, L. J., and Rutter, W. J., Nucleotide sequence of the yeast 5S ribosomal RNA gene and adjacent putative control regions, *Nature*, (London), 267, 641, 1977.
8. Maxam, A. M., Tizard, R., Skryabin, K. G., and Gilbert, W., Promoter region for yeast 5S ribosomal RNA, *Nature* (London), 267, 643, 1977.
9. Ullrich, A., Shine, J., Chirgwin, J., Pictet, R., Tischer, E., Rutter, W. J., and Goodman, H. M., Rat insulin genes: construction of plasmids containing the coding sequences, *Science*, 196, 1313, 1977.
10. Efstratiadis, A., Kafatos, F., and Maniatis, T., The primary structure of rabbit β-globin mRNA as determined from cloned DNA, *Cell*, 10, 571, 1977.
11. Sanger, F. and Coulson, A. R., A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase, *J. Mol. Biol.*, 94, 441, 1975.
12. Maxam, A. and Gilbert, W., A new method for sequencing DNA, *Proc. Natl. Acad. Sci. U.S.A.*, 74, 560, 1977.
13. Kornberg, T. and Kornberg, A., Bacterial DNA polymerases, in *The Enzymes*, Vol. 10, 3rd ed., Boyer, P. D., Ed., Academic Press, New York, 1974, 119.
14. Lehman, I. R., T4 DNA polymerase, in *Methods in Enzymology*, Vol, 29, Grossman, L. and Moldave, K., Eds., Academic Press, New York, 1974, 46.
15. Verma, I. M. and Baltimore, D., Purification of the RNA-directed DNA polymerase from avian myeloblastosis virus and its assay with polynucleotide templates, in *Methods in Enzymology*, Vol. 29, Grossman, L. and Moldave, K., Eds., Academic Press, New York, 1974, 125.
16. Bollum, F. J., Terminal deoxynucleotidyl transferase, in *The Enzymes*, Vol. 10, 3rd ed., Boyer, P. D., Ed., Academic Press, New York, 1974, 145.
17. Stadtman, E. R., Adenylate transfer reactions, in *The Enzymes*, Vol. 8, 3rd ed., Boyer, P. D., Ed., Academic Press, New York, 1973, 1.
18. Sippel, A. E., Purification and characterization of adenosine triphosphate: ribonucleic acid adenyltransferase from *Escherichia coli*, *Eur. J. Biochem.*, 37, 31, 1973.
19. Richardson, C. C., Phosphorylation of nucleic acid by an enzyme from T4 bacteriophage-infected *Escherichia coli*, *Proc. Natl. Acad. Sci. U.S.A.*, 54, 158, 1965.
20. Barrell, B. G., Fractionation and sequence analysis of radioactive nucleotides, in *Procedures in Nucleic Acids Research*, Vol. 2, Cantoni, G. L., and Davies, D. R., Eds., Harper and Row, New York, 1971, 751.
21. Ziff, E. B., Sedat, J. W., and Galibert, F., Determination of the nucleotide sequence of a fragment of bacteriophage ϕX 174 DNA, *Nature (London), New Biol.*, 241, 34, 1973.
22. Maniatis, T., Ptashne, M., Barrell, B. G., and Donelson, J., Sequence of a repressor-binding site in the DNA of bacteriophage λ, *Nature* (London), 250, 394, 1974.
23. Schaller, H., Gray, C., and Herrmann, K., Nucleotide sequence of an RNA polymerase binding site from the DNA of bacteriophage fd, *Proc. Natl. Acad. Sci. U.S.A.*, 72, 737, 1975.
24. Galibert, F., Sedat, J., and Ziff, E., Direct determination of DNA nucleotide sequences: structure of a fragment of bacteriophage ϕX 174 DNA, *J. Mol. Biol.*, 87, 377, 1974.
25. Sedat, J. W., Ziff, E. B., and Galibert, F., Direct determination of DNA nucleotide sequences. Structures of large specific fragments of bacteriophage ϕX 174 DNA, *J. Mol. Biol.*, 107, 391, 1976.
26. Proudfoot, N. J. and Longley, J. I., The 3′ terminal sequences of human α and β globin messenger RNAs: comparison with rabbit globin messenger RNA, *Cell*, 9, 733, 1976.
27. Hamlyn, P. H., Gillam, S., Smith, M., and Milstein, C., Sequence analysis of the 3′ non-coding region of mouse immunoglobulin light chain messenger RNA, *Nucleic Acids Res.*, 4, 1123, 1977.
28. Sanger, F., Donelson, J. E., Coulson, A. R., Kössel, H., and Fischer, D., Use of DNA polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in bacteriophage f1 DNA, *Proc. Natl. Acad. Sci. U.S.A.*, 70, 1209, 1973.
29. Smith, H. O. and Nathans, D., A suggested nomenclature for bacterial host modification and restriction systems and their enzymes, *J. Mol. Biol.*, 81, 419, 1973.
30. Middleton, J. H., Edgell, M. H., and Hutchison, C. A., III, Specific fragments of ϕX 174 deoxyribonucleic acid produced by a restriction enzyme from *Haemophilus aegyptius* endonuclease Z, *J. Virol.*, 10, 42, 1972.
31. Old, R., Murray, K., and Roizes, G., Recognition sequence of restriction endonuclease III from *Haemophilus influenzae*, *J. Mol. Biol.*, 92, 331, 1975.
32. Brown, N. L., Hutchison, C. A., III, and Smith, M., *J. Mol. Biol.*, in press.
33. Roberts, R. J., Restriction endonucleases, *Crit. Rev. Biochem.*, 4(2), 123, 1976.

34. Murray, K. and Old, R. W., The primary structure of DNA, *Prog. Nucleic Acid Res. Mol. Biol.*, 14, 117, 1974.
35. Salser, W. A., DNA sequencing techniques, *Annu. Rev. Biochem.*, 43, 923, 1974.
36. Wu, R., Bambara, R., and Jay, E., Recent advances in DNA sequence analysis, *Crit. Rev. Biochem.*, 2(4), 455, 1975.
37. Boyer, P. D., Ed., *The Enzymes*, Vol. 4, 3rd ed., Academic Press, New York, 1971.
38. Burton, K. and Petersen, G. B., The frequencies of certain sequences of nucleotides in deoxyribonucleic acid, *Biochem. J.*, 75, 17, 1960.
39. Brown, N. L. and Smith, M., The sequence of a region of bacteriophage ϕX 174 DNA coding for parts of genes A and B, *J. Mol. Biol.*, 116, 1, 1977.
40. Brownlee, G. G. and Sanger, F., Chromatography of ^{32}P-labelled oligonucleotides on thin layers of DEAE cellulose, *Eur. J. Biochem.*, 11, 395, 1969.
41. Ysebaert, M., Thys, F., Van de Voorde, A., and Fiers, W., Nucleotide sequence of the restriction fragments *Hin*d L and *Hin*d M of SV40 DNA, *Nucleic Acids Res.*, 3, 3409, 1976.
42. Subramanian, K. N., Dhar, R., and Weissman, S. M., Nucleotide sequence of a fragment of SV40 DNA that contains the origin of DNA replication and specifies the 5′ ends of "early" and "late" viral RNA I. Mapping of the restriction endonuclease sites within the *Eco* RII-G fragment and strategy employed for its sequence analysis, *J. Biol. Chem.*, 252, 333, 1977.
43. Berg, P., Fancher, H., and Chamberlin, M., The synthesis of mixed polynucleotides containing ribo- and deoxyribo-nucleotides by purified preparations of DNA polymerase from *Escherichia coli*, in *Symp. Informational Macromolecules*, Vogel, H., Bryson, V., and Lampen, J. O., Eds., Academic Press, New York, 1963, 467.
44. Salser, W., Fry, K., Brunk, C., and Poon, R., Nucleotide sequencing of DNA: preliminary characterization of the products of specific cleavages at guanine cytosine or adenine residues, *Proc. Natl. Acad. Sci. U.S.A.*, 69, 238, 1972.
45. Air, G. and Bridgen, J., Correlation between a coat protein amino-terminal sequence and a ribosome-binding DNA sequence from ϕX 174, *Nature (London) New Biol.*, 241, 40, 1973.
46. Air, G. M., Blackburn, E.H., Sanger, F., and Coulson, A. R., The nucleotide and amino acid sequences of the 5′ (N) terminal region of gene G of bacteriophage ϕX 174, *J. Mol. Biol.*, 96, 703, 1975.
47. Air, G. M., Blackburn, E. H., Coulson, A. R., Galibert, F., Sanger, F., Sedat, J. W., and Ziff, E. B., Gene F of bacteriophage ϕX 174. Correlation of nucleotide sequences from the DNA and amino acid sequences from the gene product, *J. Mol. Biol.*, 107, 445, 1976.
48. Air, G. M., Sanger, F., and Coulson, A. R., Nucleotide and amino acid sequences of gene G of ϕX 174, *J. Mol. Biol.*, 108, 519, 1976.
49. Air, G. M., Amino acid sequences from the gene F (capsid) protein of bacteriophage ϕX 174, *J. Mol. Biol.*, 107, 433, 1976.
50. Englund, P. T., The 3′ terminal nucleotide sequences of T7 DNA, *J. Mol. Biol.*, 66, 209, 1972.
51. Fiddes, J. C., Nucleotide sequence of the intercistronic region between genes G and F in bacteriophage ϕX 174 DNA, *J. Mol. Biol.*, 107, 1, 1976.
52. Barrell, B. G., Air, G. M., and Hutchison, C. A., III, Overlapping genes in bacteriophage ϕX 174, *Nature* (London), 264, 34, 1976.
53. Beck, E., Smith, M., and Schaller, H., *Nucleic Acids Res.*, in press.
54. Blackburn, E. H., Transcription by *Escherichia coli* RNA polymerase of a single-stranded fragment of bacteriophage ϕX 174 DNA 48 residues in length, *J. Mol. Biol.*, 93, 367, 1975.
55. Blackburn, E. H., Transcription and sequence analysis of a fragment of bacteriophage ϕX 174 DNA, *J. Mol. Biol.*, 107, 417, 1976.
56. Smith, M., Brown, N. L., Air, G. M., Barrell, B. G., Coulson, A. R., Hutchison, C. A., III, and Sanger, F., DNA sequence at the C-termini of the overlapping genes A and B in bacteriophage ϕX 174, *Nature* (London), 265, 702, 1977.
57. Brown, N. L. and Smith, M., DNA sequence of a region of the ϕX 174 genome coding for a ribosome binding site, *Nature* (London), 265, 695, 1977.
58. Hutchison, C. A ., III, Ph. D., thesis, California Institute of Technology, Pasadena, 1969.
59. Borrias, W. E., Weisbeek, P. J., and Van Arkel, G. A., An intracistronic region of gene A of bacteriophage ϕX 174 not involved in progeny RF DNA synthesis, *Nature* (London), 261, 245, 1976.
60. Weisbeek, P. J., Borrias, W. E., Langeveld, S. A., Baas, P. D., and Van Arkel, G. A., Bacteriophage ϕX 174: gene A overlaps gene B, *Proc. Natl. Acad. Sci. U.S.A.*, 74, 2504, 1977.
61. Celma, M. L., Dhar, R., Pan, J., and Weissman, S. M., Comparison of the nucleotide sequence of the messenger RNA for the structural protein of SV40 with the DNA sequence encoding the amino acids of the protein, *Nucleic Acids Res.*, 4, 2549, 1977.
62. Contreras, R., Rogiers, R., Van de Voorde, A., and Fiers, W., Overlapping of the VP2-VP3 gene and the VP1 gene in the SV40 genome, *Cell*, 12, 529, 1977.

63. Birnsteil, M. L., Schaffner, W., and Smith, H. O., DNA sequences coding for the H2B histone of *Psammechinus miliaris, Nature* (London), 266, 603, 1977.
64. Roychoudhoury, R., Jay, E., and Wu, R., Terminal labeling and addition of homopolymer tracts to duplex DNA fragments by terminal deoxynucleotidyl transferase, *Nucleic Acids Res.*, 3, 863, 1976.
65. Richardson, C. C., Polynucleotide kinase from *Escherichia coli* infected with bacteriophage T4, in *Procedures in Nucleic Acid Research*, Vol. 2, Cantoni, G. L. and Davies, D. R., Eds., Harper and Row, New York, 1971, 815.
66. Van de Sande, J. H., Kleppe, K., and Khorana, H. G., Reversal of bacteriophage T4 induced polynucleotide kinase action, *Biochemistry*, 12, 5050, 1973.
67. Lillehaug, J. R. and Kleppe, K., Effect of salts and polyamines of T4 polynucleotide kinase, *Biochemistry*, 14, 1225, 1975.
68. Schwartz, E., Scherer, G., Hobom, G., and Kössel, H., Nucleotide sequences from the DNA of bacteriophage λ coding for the *Cro* and CII regulatory proteins and for the N-terminal part of the O-protein, *Nature* (London), 292, 410, 1978.
69. Arrand, J. R. and Roberts, R. J., unpublished data, 1977.
70. Rekosh, D. M. K., Russell, W. C., Bellett, A. J. D., and Robinson, A. J., Identification of a protein linked to the ends of adenovirus DNA, *Cell*, 11, 283, 1977.
71. Lawley, P. D. and Brookes, P., Further studies on the alkylation of nucleic acids and their constituent nucleotides, *Biochem. J.*, 89, 127, 1963.
72. Cashmore, A. R. and Petersen, G. B., The degradation of DNA by hydrazine: a critical study of the suitability of the reaction for the quantitative determination of purine nucleotide sequences, *Biochin. Biophys. Acta*, 174, 591, 1969.
73. Scherer, G., Hobom, G., and Kössel, H., DNA base sequence of the Po promoter region of phage λ, *Nature* (London), 265, 117, 1977.
74. Humayun, Z., DNA sequence of the end of the CI gene in bacteriophage λ, *Nucleic Acids Res.*, 4, 2137, 1977.
75. Humayun, Z., Jeffrey, A., and Ptashne, M., Completed DNA sequences and organization of repressor-binding sites in the operators of phage lambda, *J. Mol. Biol.*, 112, 265, 1977.
76. Dhar, R., Reddy, V. B., and Weissman, S. M., Nucleotide sequence of the DNA encoding the 5′ terminal sequences of SV40 late RNA, *J. Biol. Chem.*, 253, 612, 1978.
77. Liu, A. Y., Paddock, G. V., Heindell, H. C., and Salser, W., Nucleotide sequences from a rabbit alpha globin gene inserted in a chimeric plasmid, *Science*, 196, 192, 1977.
78. Sanger, F., Nicklen, S., and Coulson, A. R., DNA sequencing with chain terminating inhibitors, *Proc. Natl. Acad. Sci. U.S.A.*, 74, 5463, 1977.
79. Porter, A., unpublished data, 1977.
80. Landy, A. and Ross, W., Viral integration and excision: structure of the lambda *att* sites, *Science*, 197, 1147, 1977.
81. Heyneker, H. L., Shine, J., Goodman, H. M., Boyer, H. W., Rosenberg, J., Dickerson, R. E., Narang, S. A., Itakura, K., Lin, S.-Y., and Riggs, A. D., Synthetic *lac* operator DNA is functional in vivo, *Nature* (London), 263, 748, 1976.
82. Reddy, V. B., Dhar, R., and Weissman, S. M., Nucleotide sequence of the genes for the SV40 proteins VP2 and VP3, *J. Biol. Chem.*, 253, 621, 1978.
83. Soeda, E., Miura, K., Nakaso, A., and Kimura, G., Nucleotide sequence around the replication origin of polyoma virus DNA, *FEBS Lett.*, 79, 383, 1977.
84. Brown, K. D., personal communication, 1977.
85. Bastia, D., The nucleotide sequence surrounding the origin of DNA replication of col El, *Nucleic Acids Res.*, 4, 3123, 1977.
86. Stoll, E., Billeter, M. A., Palmenberg, A., and Weissmann, C., Avian myeloblastosis virus RNA is terminally redundant; implications for the mechanism of oncorna virus replication, *Cell*, 12, 57, 1977.
87. Volckaert, G., Contreras, R., Soeda, E., Van de Voorde, A., and Fiers, W., Nucleotide sequence of simian virus 40 *Hind* H restriction fragment, *J. Mol. Biol.*, 110, 467, 1977.
88. Contreras, R., Volckaert, G., Thys, F., Van de Voorde, A., and Fiers, W., Nucleotide sequence of the restriction fragment Hind F-EcoR, 2 of SV40 DNA, *Nucleic Acids Res.*, 4, 1001, 1977.
89. Zain, B. S. and Roberts, R. J., Characterization and sequence analysis of a recombination site in the hybrid virus Ad2⁺ND$_1$, *J. Mol. Biol.*, 120, 13, 1978.
90. Pan, J., Reddy, V. B., Thimmappaya, B., and Weissman, S. M., Nucleotide sequence of the gene for the major structural protein of SV40 virus, *Nucleic Acids Res.*, 4, 2539, 1977.
91. Gray, C. P., Sommer, R., Polke, C., Beck, E., and Scaller, H., Structure of the origin of DNA replication of bacteriophage fd, *Proc. Natl. Acad. Sci. U.S.A.*, 75, 50, 1978.
92. Thimmappaya, B. and Weissman, S. M., The early region of SV40 DNA may have more than one gene, *Cell*, 11, 837, 1977.
93. Kelly, T. J. and Smith, H. O., A restriction enzyme from *Haemophilus influenzae*. II. Base sequence of the recognition site, *J. Mol. Biol.*, 51, 393, 1970.

94. Shine, J., Czernilofsky, A. P., Friedrich, R., Bishop, J. M., and Goodman, H. M., Nucleotide sequence at the 5′ terminus of the avian sarcoma virus genome, *Proc. Natl. Acad. Sci. U.S.A.*, 74, 1473, 1977.
95. Maniatis, T., Jeffrey, A., and Kleid, D., Nucleotide sequence of the rightward operator of phage λ, *Proc. Natl. Acad. Sci. U.S.A.*, 72, 1184, 1975.
96. Bennett, G. N., Brown, K. D., and Yanofsky, C., Nucleotide sequence of the promoter-operator region of the tryptophan operon of *Salmonella typhimurium*, *J. Mol. Biol.*, 121, 139, 1978.
97. Brownlee, G. G. and Cartwright, E. M., Rapid gel sequencing of RNA by primed synthesis with reverse transcriptase, *J. Mol. Biol.*, 114, 93, 1977.
98. Cheng, C. C., Brownlee, G. G., Carey, N. H., Doel, M. T., Gillam, S., and Smith, M., The 3′ terminal sequence of chicken ovalbumin messenger RNA and its comparison with other messenger RNA molecules, *J. Mol. Biol.*, 107, 527, 1976.
99. Bernard, O. D., Jackson, J., Cory, S., and Adams, J. M., Non-coding nucleotide sequence in the 3′ terminal region of a mouse immunoglobin kappa chain mRNA determined by analysis of complementary DNA, *Biochemistry*, 16, 4117, 1977.
100. Both, G. W. and Air, G. M., unpublished data, 1977.
101. Schwartz, D. E., Zamecnik, P. C., and Weith, H. L., Rous sarcoma virus genome is terminally redundant: the 3′ sequence, *Proc. Natl. Acad. Sci. U.S.A.*, 74, 994, 1977.
102. Air, G. M. and Both, G. W., unpublished data, 1977.
103. Khorana, H. G., Agarwal, K. L., Büchi, H., Caruthers, M. H., Gupta, N. K., Kleppe, K., Kumar, A., Ohtsuka, E., RajBhandary, U. L., van de Sande, J. H., Sgaramella, V., Terao, T., Weber, H., and Yamada, T., Studies on polynucleotides. CIII. Total synthesis of the structural gene for an alanine transfer ribonucleic acid from yeast, *J. Mol. Biol.*, 72, 209, 1972; 14 accompanying papers.
104. Khorana, H. G., Agarwal, K. L., Besmer, P., Büchi, H., Caruthers, M. H., Cashion, P. J., Fridkin, M., Jay, E., Kleppe, K., Kleppe, R., Kumar, A., Loewen, P. C., Miller, R. C., Minamoto, K., Panet, A., RajBhandary, U. L., Ramamoorthy, B., Sekiya, T., Takeya, T., and van de Sande, J. H., Total synthesis of the structural gene for the precursor of a tyrosine suppressor transfer RNA from *Escherichia coli*, I. General introduction, *J. Biol. Chem.*, 251, 565, 1976; 11 accompanying papers.
105. Itakura, K., Katagiri, N., Narang, S. A., Bahl, C. P., Marians, K. J., and Wu, R., Chemical synthesis and sequence studies of deoxyribo-oligonucleotides which constitute the duplex sequence of the lactose operator of *Escherichia coli*, *J. Biol. Chem.*, 250, 4592, 1975.
106. Bahl, C. P., Wu, R., Itakura, K., Katagiri, N., and Narang, S. A., Chemical and enzymatic synthesis of lactose operator of *Escherichia coli* and its binding to lactose repressor, *Proc. Natl. Acad. Sci. U.S.A.*, 73, 91, 1976.
107. Gait, M. J. and Sheppard, R. C., Rapid synthesis of oligodeoxyribonucleotides: a new solid-phase method, *Nucleic Acids Res.*, 4, 1135, 1977.
108. Gillam, S. and Smith, M., Enzymatic synthesis of deoxyriboligonucleotides of defined sequence. Properties of the enzyme, *Nucleic Acids Res.*, 1, 1631, 1974.
109. Gillam, S., Waterman, K., Doel, M., and Smith, M., Enzymatic synthesis of deoxyribo oligonucleotides of defined sequence. Deoxyribo oligonucleotide synthesis, *Nucleic Acids Res.*, 1, 1649, 1974.
110. Gillam, S., Waterman, K. and Smith, M., Enzymatic synthesis of oligonucleotides of defined sequence. Addition of short blocks of nucleotide residues to oligonucleotide primers, *Nucleic Acids Res.*, 2, 613, 1975.
111. Brownlee, G. G., unpublished data, 1977.
112. Smith, M., unpublished data, 1977.
113. Proudfoot, N. J., Complete 3′ noncoding region sequences of rabbit and human β globin messenger RNAs, *Cell*, 10, 559, 1977.
114. Baralle, F. E., Complete nucleotide sequence of the 5′ noncoding region of rabbit β globin mRNA, *Cell*, 10, 549, 1977.
115. Proudfoot, N. J., Sequence analysis of the 3′ non-coding regions of rabbit α- and β-globin messenger RNAs, *J. Mol. Biol.*, 107, 491, 1976.
116. Lewandowski, L. J., Content, J., and Leppla, S. H., Characterization of the subunit structure of the ribonucleic acid genome of influenza virus, *J. Virol.*, 8, 701, 1971.
117. Ehresmann, C., Stiegler, P., Mackie, G. A., Zimmerman, R. A., Ebel, J. P., and Fellner, P., Primary sequence of the 16s ribosomal RNA of *Escherichia coli*, *Nucleic Acids Res.*, 2, 265, 1975.
118. Haseltine, W. A., Maxam, A. M., Gilbert, W., Rous sarcoma virus genome is terminally redundant: the 5′ sequence, *Proc. Natl. Acad. Sci. U.S.A.*, 74, 989, 1977.
119. Horiuchi, K. and Zinder, N. D., Site-specific cleavage of single-stranded DNA by a Hemophilus restriction endonuclease, *Proc. Natl. Acad. Sci. U.S.A.*, 72, 2555, 1975.
120. Coffin, J. M. and Haseltine, W. A., Terminal redundancy and the origin of replication of Rous sarcoma virus RNA, *Proc. Natl. Acad. Sci. U.S.A.*, 74, 1908, 1977.

121. Merregaert, J., van Emmelo, J., Devos, R., Fiers, W., Porter, A., and Fellner, P., The 3′ terminal nucleotide sequence of encephalomyocarditis virus RNA, *Eur. J. Biochem.*, 82, 55, 1978.
122. Donis-Keller, H., Maxam, A. M., and Gilbert, W., Mapping adenines, guanines and pyrimidines in RNA, *Nucleic Acids Res.*, 4, 2527, 1977.
123. Barnes, W. M., DNA sequencing by partial ribosubstitution, *J. Mol. Biol.*, 119, 83, 1978.
124. Beer, M. and Moudrianakis, E. N., Determination of base sequence in nucleic acids with the electron microscope: visibility of a marker, *Proc. Natl. Acad. Sci. U.S.A.*, 48, 409, 1962.
125. Erickson, H. P. and Beer, M., Electron microscopic study of the base sequence in nucleic acids. VI. Preparation of ribonucleic acid with marked guanosine monophosphate nucleotides, *Biochemistry*, 6, 2694, 1967.
126. Langmore, J. P., Cozzarelli, N. R., and Crewe, A. V., A base-specific single heavy atom stain for electron microscopy, *Proc. Electron Microsc. Soc. Am.*, 30, 184, 1972.
127. Wiebers, J. L., Sequence analysis of oligodeoxyribonucleotides by mass spectrometry, *Anal. Biochem.*, 51, 542, 1973.
128. Wiebers, J. L. and Shapiro, J. A., Sequence analysis of oligodeoxyribonucleotides by mass spectrometry. I. Dinucleoside monophosphates, *Biochemistry*, 16, 1044, 1977.
129. Burgard, D. R., Perone, S. P., and Wiebers, J. L., Sequence analysis of oligodeoxyribonucleotides by mass spectrometry. II. Application of computerized pattern recognition to sequence determination of di-, tri- and tetranucleotides, *Biochemistry*, 16, 1051, 1977.
130. Lau, R. Y., Kennedy, T. D., and Lane, B. G., Wheat embryo ribonucleates. III. Modified nucleotide constituents in each of the 5.8s, 18s and 26s ribonucleates, *Can. J. Biochem.*, 52, 1110, 1974.
131. Zimmern, D., unpublished data, 1977.
132. Fiddes, J. C., Barrell, B. G., and Godson, G. N., Nucleotide sequences of the separate origins of synthesis of bacteriophage G4 viral and complementary DNA strands, *Proc. Natl. Acad. Sci. U.S.A.*, 75, 1081, 1978.
133. Barrell, B. G., in *International Review of Biochemistry*, Ser. II, Biochemistry of Nucleic Acids, Clark, B. F. C., Ed., University Park Press, Baltimore, in press.
134. Simoncsits, A., Brownlee, G. G., Brown, R. S., Rubin, J. R., and Guilley, H., *Nature* (London), 269, 833, 1977.