RECENT ADVANCES IN DNA SEQUENCE ANALYSIS

Author: **Ray Wu**
Robert Bambara
Ernest Jay
Section of Biochemistry, Molecular and Cell Biology
Cornell University
Ithaca, New York

Referee: Winston Salser
Department of Biology
U.C.L.A.
Los Angeles, California

INTRODUCTION

During most of the last 15 years, primary structure determination of nucleic acids has been focused on RNA, with relatively little work involving DNA. There were several very good reasons for this choice of emphasis: (1) Transfer RNA molecules are less than 100 nucleotides long. Ribosomal and certain messenger RNA molecules less than 500 nucleotides long can be isolated. In contrast, DNA molecules are very large. The smallest DNA from viruses is 5,000 nucleotides long, while cellular DNA molecules contain millions of nucleotides. (2) Base-specific nucleases are available which are able to hydrolyze RNA molecules to give specific fragments in the 2 to 20-nucleotide size range. T_1 RNase, which cleaves after G residues, and pancreatic RNase, which cleaves after pyrimidine nucleotides, are most commonly used. The first DNA endonucleases isolated were quite nonspecific. A digest of a fragment even 30 nucleotides long produces a complex mixture of many products. Nevertheless, many potentially interesting nucleotide sequences are available only in DNA. Operator, promoter, initiation sites of replication, and other control regions, all of which are of crucial importance in the understanding of cell and viral development, have awaited the advent of effective DNA sequencing methods.

Within the last 6 years, a new era of DNA sequencing has begun with the development of a variety of effective techniques which allow specific labeling, isolation, and sequencing of large segments of biologically important DNA. The recent discovery and utilization of specific endonucleases for DNA have allowed production of specific segments of DNA in the 30 to 500-nucleotide size range. Fractionation techniques have been worked out for purification of these large fragments. Methods have been devised for isolation of protein binding sites in the 20 to 100-nucleotide size range. Finally, a number of techniques have been developed for sequencing of oligonucleotides in the 2 to 50-nucleotides size range.

The following is a description of the development of the more recent DNA sequencing techniques, as well as a discussion of some of the new sequences which have been obtained.

In this review article, each section is prefaced by a short introduction which gives a general summary of the subject matter under consider-

ation. The remainder of the section provides details and critical evaluation. Although we have attempted to include most of the important developments in DNA sequence analysis in this article, the coverage is by no means exhaustive. The concepts, methods, and results covered in each section are presented chronologically, whenever possible. Because of the rapid progress of research work on DNA sequence analysis in recent years, different laboratories often report almost simultaneously the same or similar results which are obtained independently. This review is organized with respect to DNA sequencing techniques, rather than related results. Thus, the same results, obtained by workers using differing techniques, are likely to appear in different sections. However, results on related work are often summarized in the same table or figure.

Several other detailed review articles on the same subjects have recently appeared.[1-3] However, the coverage and emphasis of subject matter is usually different in each case. In general, the work from the authors' laboratory is treated in more detail, mainly because of their greater familiarity with the material, and sometimes at the request of the editor of the journal.[1] Furthermore, because of the rapid progress in this field, substantial amounts of new information have become available since the other articles were written.

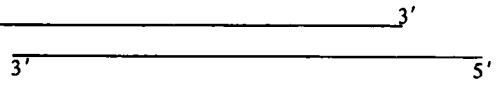
A. PREPARATION OR ISOLATION OF SPECIFIC SHORT DNA FRAGMENTS FOR SEQUENCE ANALYSIS

1. Repair Synthesis

A procedure has been developed for the sequence analysis of the single-stranded cohesive ends of a DNA molecule, such as λ DNA, by repair synthesis. The natural 3′ ends of λ DNA, which serve as primers, are extended by the incorporation of radioactive deoxynucleotides. Each labeled segment, which has a sequence complementary to its template strand, the cohesive end, can be more readily sequenced, since it is relatively short and highly radioactive. Furthermore, its sequence analysis is not complicated by the very long, unlabeled portion of the DNA molecule.

a. Sequence of the 5′ Protruding Cohesive Ends

The temperate bacteriophages are characterized by having DNA with cohesive ends. Cohesive ends are short, single-stranded regions extending the 5′ termini of the double-stranded phage DNA, as shown:



The single-stranded regions are called "cohesive ends" because they can hybridize to form circular DNA.

Temperate bacteriophages can be divided into two specificity families, distinguishable from each other by two properties. Within the same family, (1) mixed dimers may be formed by hydrogen bonding between two DNA molecules from different bacteriophages, and (2) in intact phage particle may help the free DNA from another phage to infect a bacterium. These properties are not displayed if two members of different families are tested.[4] The first specificity family, the lambdoids, includes the bacteriophages λ, φ80, 21, 424, 434, and 82. The second specificity family, the P2 types, includes P2, 186, P4, 299, φD, and N1. It is likely that both properties involve cohesion of the single-stranded ends of two DNA molecules. Therefore, members of each specificity family should have cohesive end sequences similar to those of other members of the same family, but quite different from those of members of the other family.

The sequence determination of the cohesive ends of bacteriophage λ DNA by Wu and collaborators[5-7] represented the beginning of DNA sequence analyses from known regions of the DNA molecule as well as the prelude to the new sequencing technique of primer extension. The strategy for sequencing these regions involved the use of the native 3′ ends of the DNA as the primers for the *E. coli* DNA polymerase-catalyzed repair reaction in order to copy the templates provided by the protruding 5′ end regions (see Figure 1). The repair reaction can be used for sequence analysis in two ways: (1) to determine short sequences by following the partial incorporation of nucleotides when one, two, or three radioactive nucleoside triphosphates are supplied; or (2) to provide long labeled DNA segments complementary to the template strand using all four radioactive nucleoside triphosphates. These radioactive segments can then be degraded and analyzed. Additional details of the repair reactions used and the results of the analyses can be found in Section B-1. The cohesive end sequences of λ DNA are given in Figure 2.

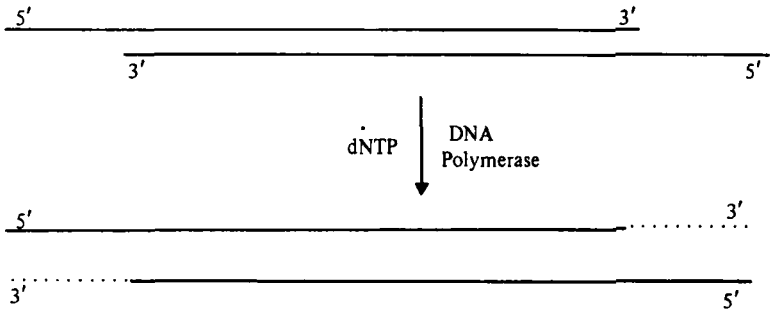


FIGURE 1. Repair synthesis of a bacteriophage DNA. The single-stranded cohesive ends of the DNA molecule, such as λ DNA, can be repaired in the presence of radioactive deoxynucleoside triphosphates (dNTP) and *E. coli* DNA polymerase I. The dots in the lower panel of this figure represent radioactive mononucleotides added to the 3′-ends of the DNA molecules.

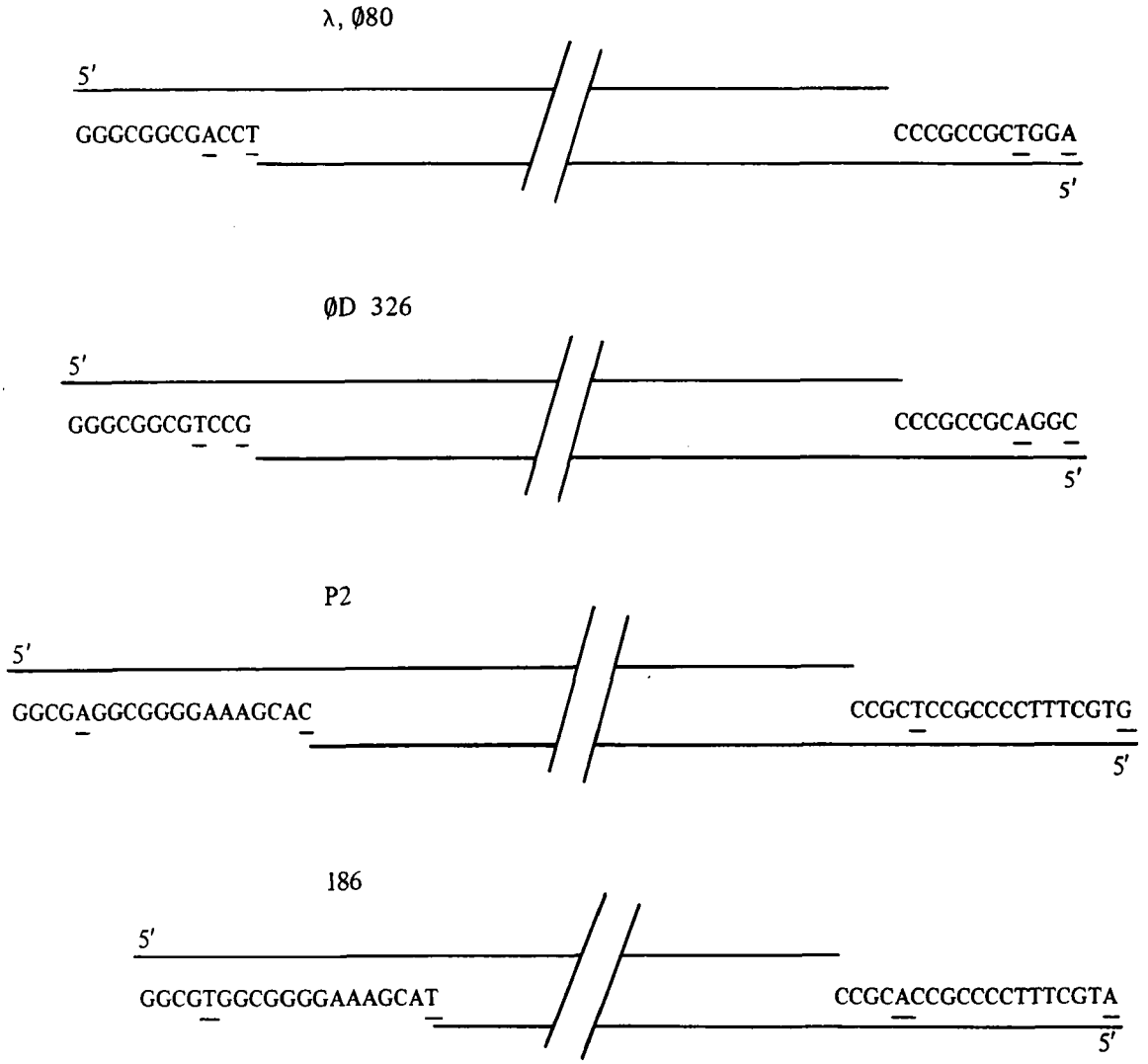


FIGURE 2. Cohesive end sequence information from several temperate bacteriophages.[6,14,22,24] Bacteriophages λ, φ80, and φD 326 are all members of the lambdoid specificity family. Bacteriophages P2 and 186 are both members of the P2 specificity family. Differing sequences between members of the same family are underlined.

It should be pointed out that for DNA sequence analysis using repair synthesis and labeled deoxynucleoside triphosphates, the relative as well as the absolute specific activity of the four nucleoside triphosphates can now be independently determined by partial or complete repair synthesis using λ DNA as the test system.[8]

The repair synthesis approach has not only proven very useful for determining the sequence of the single-stranded cohesive ends of phage DNA; it can also be extended for the sequencing of any double-stranded DNA after the removal of nucleotides from the 3′ ends with *E. coli* exonuclease III. Details of the production of DNA with 5′ terminated single-stranded ends which resemble the cohesive ends of λ DNA will be discussed in Section B-1-b. The partial repair synthesis has also been applied to sequence the nucleotides at the uneven ends of DNA after R_I restriction enzyme cleavage.[9] In this case, reverse transcriptase was used in place of *E. coli* DNA polymerase.

After sequence determination of the cohesive ends of λ DNA, those of 186 DNA were sequenced by the same procedure.[10] 186 DNA was examined because phage 186 is a member of the P2 specificity family. Although λ DNA and 186 DNA have the same terminal 5′ nucleotides, the remainder of the sequences is quite different, as was expected.

Phages λ and φ80 are both members of the lambdoid family, but φ80 DNA is 8% shorter than λ DNA, and heteroduplex formation experiments have indicated that only about 25% of the φ80 DNA molecule is highly homologous with the λ DNA molecule.[11] Therefore, it is of considerable interest to study the cohesive end sequences of φ80 DNA to determine the sequence relationship of DNA molecules within the same specificity family.

The comparison of the λ and φ80 terminal sequences in the cohesive end regions as well as the regions adjacent to the cohesive ends is especially important for another reason. The DNA molecules of all of the temperate phages replicate in a closed, double-stranded, circular form. Before packaging in the phage head, however, the closed circles are specifically cleaved by an endonuclease which produces linear, double-stranded DNA molecules with cohesive ends. The cleavage is unusually specific since it occurs only once on a DNA circle of 50,000 nucleotides. Although this specific endonuclease has not yet been successfully isolated, there is some genetic information on its specificity. It was found that φ80 can excise a λ genome from a tandem dilysogen of λ.[12] This means that the φ80 endonucleolytic function recognizes the λ cohesive end region. Thus, a comparison of the sequences of the λ, and φ80 terminal regions around the endonuclease cleavage sites would help in the understanding of the site of recognition and the mode of action of the specific endonuclease.

The sequence of the cohesive ends of φ80 DNA, determined by the repair synthesis technique, was found to be identical[13] to that of λ DNA. Both sequences have been confirmed by labeling the 5′ ends of the DNA, followed by digestion with pancreatic DNase and 2-D mapping techniques.[14] Details of these techniques, as well as a comparison of the two approaches for sequence analysis, are given in Section B-2.

b. Sequence at the 3′ Termini Adjacent to the Cohesive Ends

Sequence information at the 3′ termini of λ DNA was obtained by Weigel, Englund, Murray, and Old[15] after 3′ terminal labeling using the T_4 DNA polymerase nucleotide exchange reaction. Terminally labeled oligonucleotides obtained after partial pancreatic DNase digestion of the labeled DNA were separated on DEAE- and AE-cellulose electrophoresis, and sequencing was done by comparing the mobilities of adjacent oligonucleotides. The 3′ terminal sequences d(-G-T-T-A-C-G) for the right-hand 3′ end of λ DNA and d(-A-C-C-C-G-C-G) for the left-hand 3′ end were obtained.

Brezinski and Wang[16] specifically labeled the 3′ termini of λ DNA by partial repair synthesis with [^{3}H] nucleotides. Then they mixed a pancreatic DNase I digest of this material with a similar digest of uniformly ^{32}P-labeled DNA. After 2-D electrophoresis, oligonucleotides containing ^{3}H and ^{32}P were analyzed. Short sequence obtained in this fashion, d(pCpG) from the right-hand 3′ terminus and d(pCpGpCpG) from the left-hand 3′ terminus, agreed with those obtained by Weigel et al.[15]

Sequence information from the 3′ terminal regions of λ DNA and new 3′ terminal information from φ80 DNA were independently obtained in our laboratory using a different technique.[17-19] In order to obtain this sequence, each 3′ end was separately and specifically labeled using the DNA polymerase-catalyzed partial repair synthesis.[17] The terminally labeled DNA was partially digested,

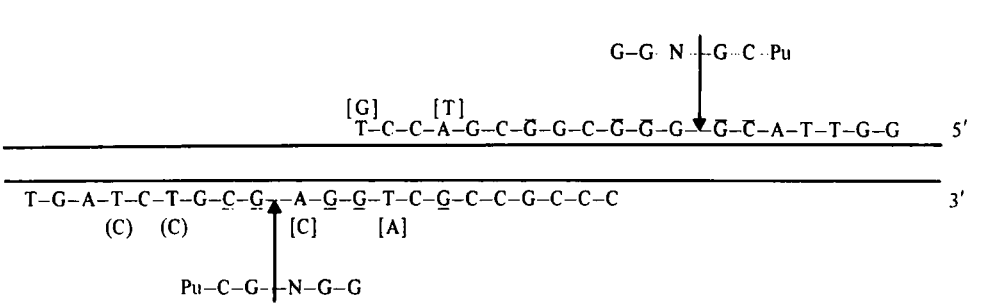


FIGURE 3. Terminal sequences of bacteriophage DNA's which may be involved in *ter* endonucleolytic recognition. All of the sequence information for the terminal regions of ϕ80 DNA[13,17] is shown. The illustration is drawn showing the arrangement of nucleotides before the endonucleolytic cleavage which produces the cohesive ends. The nucleotides shown in parentheses are those present in λ DNA.[6] The nucleotides shown in brackets are those present in ϕD 326 DNA.[22] The arrows point to the cleavage sites. The sequences underlined are the symmetrical recognition site sequences proposed by Weigel et al.[15] The sequences drawn across the arrows are the recognition sites proposed by Ghangas et al.[17] Pu = purines, N = any of the four deoxynucleotides.

and the 3′ end region was sequenced by interpretation of differential mobilities of the labeled oligonucleotides on 2-D homochromatography (details will be given in Section B-2). The 3′ terminal sequences, d(-G-G-T-T-A-C-G) for the right-hand 3′ end of λ DNA and d(-T-T-G-A-C-C-C-G-C-G) for the left-hand 3′ end, were obtained by this method.

All of the sequence information which has been determined for the terminal regions of λ DNA and ϕ80 DNA is shown in Figure 3. The illustration is drawn showing the arrangement of nucleotides before the endonucleolytic cleavage which produces the cohesive ends. The arrows point to the cleavage sites. The lower sequence between the arrows is the right-hand cohesive end. The sequence to the right of both arrows is the right-hand 3′ terminal sequence. The upper sequence between the arrows is the left-hand cohesive end. The sequence to the left of both arrows is the left-hand 3′ terminal sequence.

On the basis of the sequences of the λ DNA terminal regions, a number of *ter* endonuclease recognition sites have been proposed.

Weigel et al.[15] proposed that the enzyme is a dimer of identical subunits, arranged about an axis of rotational symmetry in such a way that it can bind to the DNA with its axis aligned with that of a symmetrical sequence found in the recognition site region (indicated in Figure 3). The subunits would be aligned so that the two cleavage reactions could occur simultaneously. In fact, in a space filling model the two cleavage sites are on the same side of the helix, so that the enzyme could easily exist as a dimer and contact both cleavage sites at the same time. This multiple subunit model resembles those of Bernardi[20] for spleen acid deoxyribonuclease and of Kelly and Smith[21] for *H. influenzae* restriction enzyme. It is also possible that the enzyme locally denatures the DNA and recognizes a single-strand sequence,[2] but there is no way for the primary sequence to suggest this.

On the basis of the sequence information available for the terminal regions of λ DNA and ϕ80 DNA, Wu and collaborators[17,18] proposed that the *ter* function endonuclease recognizes the two identical hexanucleotides purine-C-G-purine-G-G, which would be cleaved exactly in the center with one cleavage on each strand to produce a DNA molecule with two protruding 5′ terminated single strands.

Figure 3 also shows the sequence differences between the terminal regions on λ DNA and ϕ80 DNA. All of the sequences for λ DNA are the same as in ϕ80 DNA except that the two dC bases in the λ sequence (shown in parentheses) replace the two dT bases in the ϕ80 sequence just above them.

If the recognition sequences were wider than Weigel et al.[15] or Wu and collaborators[17,18] had proposed, they would overlap this region of sequence differences. As they are presently placed, however, they overlap no areas of sequence difference.

Preliminary sequence information from the cohesive ends of another lambdoid phage ϕD326 has just appeared.[22] The cohesive end sequences, shown in Figure 2, differ from those of λ and ϕ80 DNA in two positions. (The new bases to be substituted are given in brackets in Figure 3.) Nevertheless, the same *ter* function endonuclease is able to cleave at the λ and ϕD326 cohesive ends. In view of this new information, the proposed hexanucleotide recognition site for the lambdoid

ter endonuclease will have to be modified to two identical purine-C-G-N-G-G sequences, where N can be any one of the four nucleotides. This proposed recognition sequence, shown also in Figure 3, is now quite similar to that of Weigel et al.[15]

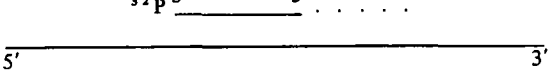
The complete, 19-nucleotide long cohesive end sequences for DNA from the bacteriophages 186 and P2 have been determined by Padmanabhan et al.[10,23,24] using repair synthesis and ribosubstitution techniques (see Section A-2). Both the cohesive end sequences of 186 DNA and P2 DNA contain an octanucleotide pyrimidine tract d(C-T-T-T-C-C-C-C) sequence at the right-hand end and a complementary purine tract d(G-G-G-G-A-A-A-G) sequence at the left-hand end. The 19-nucleotide sequence at the left-hand cohesive end of 186 DNA was determined by partial repair and complete repair synthesis using all deoxynucleoside triphosphates. For the right-hand cohesive end of 186 DNA and both ends of P2 DNA, ribosubstitution was found to accelerate the sequence analysis since specific cleavage products by pancreatic RNase and T_1 RNase can be obtained. Partial cohesive end sequences of these two phages, as well as those of phages 21, 82, 299, and 424, have been independently determined by Murray and Murray.[14] The sequences are given in Figure 2. Some di- and trinucleotide sequences have also been determined for the 3′ terminal regions of P2 and 186.[2,14] These bacteriophages belong to a specificity family different from λ and φ80, but they have cohesive ends which are recognized by a specific endonuclease. It is not yet known, however, if their endonuclease can cross react in a fashion similar to the cross reaction of the λ endonuclease.

Examination of the P2 type DNA sequences reveals essentially no sequence symmetry in the cohesive ends, as observed with the lambdoid phages. There is also no symmetry about the cleavage points. In fact, the cleavage point regions in the same molecule have dissimilar sequences. In view of these facts, neither of the recognition hypotheses proposed for the lambdoid phages appears to apply to the P2 family.

Because of the dissimilarity of sequences at cleavage sites in the same molecule, Wang and Brezinsky[25] have proposed a model in which two molecules of DNA are aligned for cleavage by an alignment protein. The duplex sequences can be made symmetrical by this process. The endonuclease could then cleave on either side of the alignment protein. The sequence d(G-G-C-G-G), which occurs in λ, φ80, P2, and 186, could be the recognition site of the alignment protein. Clearly, the information which is available right now is insufficient to allow definite conclusions about the production of the cohesive ends by the endonuclease. Resolution of the problem awaits more sequence information and purification of the endonuclease itself.

2. Primer Binding and Extension

A specific region along a long, single-stranded DNA molecule can be labeled by means of binding and extending a synthetic oligodeoxynucleotide primer.



The primer (short upper line) is to be bound at a specific site adjacent to the sequence of interest. The primer can then be extended from the 3′ end using DNA polymerase and radioactive deoxynucleoside triphosphates. The nucleotides added by the polymerase will have a sequence (dotted line) complementary to the sequence of interest. The extended primer can be isolated and sequenced by conventional methods or new methods to be described later.

a. Binding of the Primer

A considerable amount of information about the binding characteristics of short synthetic oligodeoxynucleotides was available.[26-28] It was found that the minimum length of short oligonucleotides required to form stable duplexes with the complementary oligo- or polynucleotides varied from 6 to 12 depending on the composition and concentration of the oligonucleotides used. The duplex of short oligonucleotides with 3′ hydroxyl ends can be repaired by DNA polymerase.[29-31] Goulian et al.[29] have shown that overall priming activity, using M13 DNA as template, is highest with purine rich oligonucleotides of lengths 8 to 12 for *E. coli* DNA polymerase I at 20°C. Pentamers gave less than 1% priming activity as compared to decamers. Pyrimidine octanucleotides gave less than 3% priming activity as compared to purine octanucleotides. Under certain conditions, especially with the use of much higher concentrations of oligonucleotides, an overlap of only four base pairs seemed to be sufficient for DNA ligase reaction.[30]

We have proposed a method for DNA sequence analysis using a short oligonucleotide primer which could be bound to a specific location on the

single-stranded region of a DNA molecule.[1,31] The nucleotide sequence beyond the 3′ end of the primer could be determined after extension of the primer with radioactive nucleotides using the DNA polymerase-catalyzed repair reaction. Wu[31] first used G-C rich oligonucleotides ranging in size from octanucleotide to dodecanucleotide for testing the stability of DNA-primer duplexes. The ^{32}P-labeled nonanucleotide d(T-C-G-C-C-G-C-C-C)-OH, derived from the left-hand end of the repaired λ DNA, was found to be the minimum size which can anneal to the native λ DNA efficiently. The binding of chemically synthesized dodecanucleotide to the λ DNA cohesive end region was carried out later by Heimer, Ahmad, and Nussbaum,[32] with similar results.

The following plan for sequence analysis by this technique was then proposed:[31] (1) An r-dodecanucleotide which would code for a suitable tetrapeptide sequence in a protein would be deduced with the use of the genetic code. (2) The corresponding d-dodecanucleotide would be chemically synthesized. (3) This would be used as the primer to anneal to the DNA template which contains the gene. (4) DNA polymerase would be used to extend the oligonucleotide under conditions of repair synthesis at 5°C. (5) The extended oligonucleotide would then be isolated for sequence analysis.

It was important to show that such processes of binding, extension, reisolation of the oligonucleotide, and sequence analysis were feasible in actual practice. This was tested by using a radioactive octanucleotide isolated from the repaired cohesive ends of bacteriophage 186 DNA.[33] The octanucleotide, d(C-T-T-T-C-C-C-C), is complementary to a portion of the 19-nucleotide long left-hand cohesive end of 186 DNA. This octamer should allow extension from its 3′ terminus because after binding, the template strand still has a protruding 5′ single strand to serve as template. In this experiment, no stable complex of the octanucleotide with 186 DNA was detected. However, addition of DNA polymerase and [^{3}H]dGTP allowed extension of the oligomer by one dG residue, which then gave stable binding. These observations substantiated the earlier finding[31,238] that a nonanucleotide is the minimum length for duplex formation with long DNA molecules which must be used at very low concentrations (e.g., 3 pmol of DNA/ml = 100 μg/ml) for practical reasons. Longer partial extensions and complete extension of the octamer to an undecamer and a hexadecamer allowed more efficient binding (60% efficiency). The extended oligonucleotides were then isolated and sequenced. The results indicated that the eight new nucleotides added to the primer were the exact complement of the cohesive end which served as template.

Oertel and Schaller[34] showed that an oligopyrimidine tract, C_9T_{11}, isolated from the fd DNA (+) strand could be bound to the fd DNA minus strand at a single site. The bound primer could be extended by DNA polymerase catalyzed repair synthesis at 20°C, and several specific pyrimidine tracts could be isolated from the products. Furthermore, pulse-labeled reaction products of various sizes were characterized by fingerprinting techniques, which allowed the deduction that the sequential arrangement of the pyrimidine tracts in the nucleotide sequence followed the 3′ terminus of the primer.

Several problems had to be resolved, however, before this method could be applied to sequencing regions in and around a specific gene in a long segment of DNA: (1) The oligonucleotide primer should bind only to the single specific site on the template DNA. As pointed out by Thomas,[35] when the size of the oligonucleotide increases above 8 nucleotides, it becomes very unlikely that the complementary sequence will be present more than once in a molecule of the size of λ DNA (50,000 base pairs). From this consideration, an octanucleotide primer is the minimum size for specific binding to DNA in the size range of λ DNA. (2) Theoretically, it may be possible to use the amino acid sequence of any protein to design the synthesis of deoxyoligonucleotide primer for binding to the DNA in that region. However, since the genetic code is degenerate, knowledge of the amino acid sequence of a protein does not provide complete information concerning the nucleotide sequences coding for that protein so that, in general, it would be impossible to choose a primer with a unique sequence coding for a tetrapeptide in the protein. Efforts should then be made to choose an amino acid sequence which leads to the least number of ambiguities in the triplet codes in the chemically synthesized primer.

In choosing the best primer sequence, the following guidelines were used by Wu:[31] (a) Whenever possible, choose Met and Trp which have unique codons. (b) Next, choose amino acids with the first two letters unique and with U or C

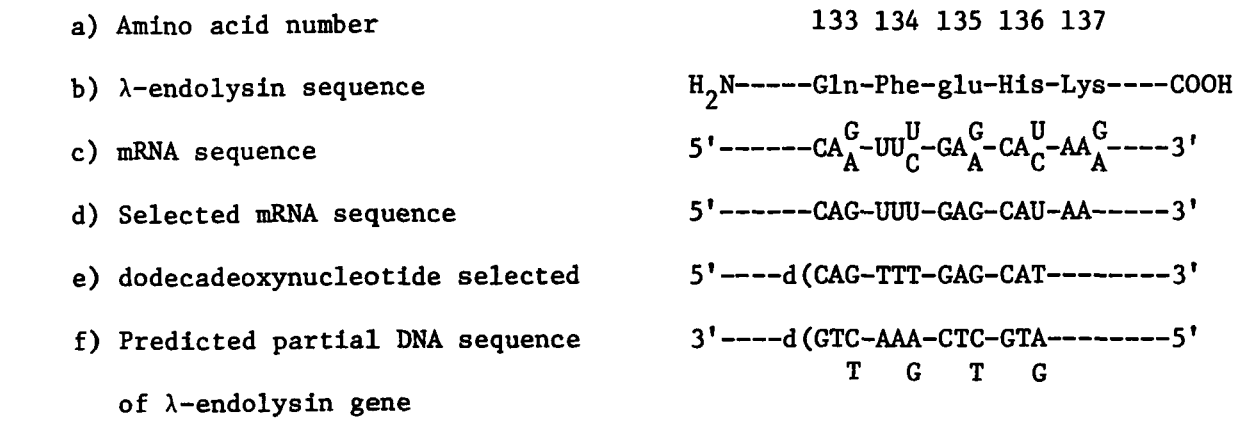


FIGURE 4. Design of the experiment and selection of the dodecadeoxynucleotide primer complementary to part of the γ-endolysin gene.[39]

in the third letter of the triplet (such as Asp, Asn, Cys, His, Phe, and Tyr) or with G or A in the third letter of the triplet (such as Glu, Gln, and Lys). (c) When the third letter ambiguity is G or A, G should be chosen because the G-T pair is able to form a somewhat stable base pair.[36-38] Similarly, when the third letter ambiguity is U or C, U should be chosen and T put into the primer because, again, the G-T pair could be formed. Using the above guidelines, Wu, Tu, and Padmanabhan[39] synthesized a primer which, in principle, would bind to a specific region of the endolysin gene of bacteriophage λ. Figure 4 (part e) shows the nucleotide sequence and the way it was selected (parts a to d). One of the reasons the λ endolysin gene was chosen for this study was that it is near one terminus of the λ DNA molecule. Treatment of λ DNA with *E. coli* exonuclease III can render the region of this gene single-stranded. Binding of the synthetic dodecamer to exonuclease III-treated λ DNA was low but measurable. In the presence of DNA polymerase and [^{3}H]dATP, binding was increased to 20% after 2 dpA residues were incorporated onto the end of the primer. The addition of these two dpA residues was expected on the basis of the λ endolysin protein sequence (Figure 4, part c) So, it appears that binding had occurred at the correct DNA site coding for the λ endolysin, and new sequence information can be obtained with the use of this primer.

One method for obtaining unambiguous primers is by analysis of frameshift mutants. In this case, a unique nucleotide sequence is deduced by comparing the amino acid sequence of the wild type and mutant proteins. A unique sequence of a tetradecamer coding for amino acids 36 to 40 of lysozyme from the T_4 phage was deduced in this way by Streisinger et al.[40] Studies were carried out using this sequence for the synthesis of a deoxytetradecamer, with an unambiguous sequence, to bind to the lysozyme gene of bacteriophage T_4.[41] Since the genes of the T_4 phage are circularly permuted, the lysozyme gene is distributed evenly along the T_4 DNA molecule. When exonuclease III is used to expose single-stranded regions of T_4 DNA, the percentage of nucleotides removed from the strand complementary to the primer used for binding is the percentage of lysozyme gene available for binding. It was possible to bind the tetradecamer, d(A-G-T-C-C-A-T-C-A-C-T-T-A-A), to an extent of approximately 30% of the exposed region, and to increase the binding to nearly 100% under conditions where extension of the primer could take place. Sequence determination of the oligonucleotide primer extended by two nucleotides suggested that binding was specific and in the correct location on the T_4 lysozyme gene.

b. Ribosubstitution and Primer Extension

A means should be found to obtain specific cleavage products of labeled DNA after repair synthesis. The technique of ribosubstitution appears to be one of the most useful means of accomplishing this. Berg et al.[42] showed that ribonucleotides could be incorporated along with deoxyribonucleotides in a DNA polymerase-catalyzed repair reaction carried out in the presence of manganese ion instead of magnesium ion. A repair reaction can then be carried out, for example, in the presence of [α-^{32}P]rCTP and

three ^{3}H-labeled dNTP compounds. A rC would then be incorporated in all of the positions where a dC would normally be placed. Furthermore, hydrolysis by means of alkali or pancreatic RNase produces oligonucleotides each with a terminal [^{32}P]Cp label. These could then be separated for sequencing. If [^{32}P]rG had been used, T_1 RNase digestion of the extended primer would produce [^{32}P]Gp terminally labeled oligonucleotides in the same fashion.

Conditions for the ribosubstitution reaction and fidelity of the incorporation during repair synthesis have been investigated in detail:

Salser et al.[43] reported that this method could be used for the extension of primers over long segments of DNA. Using M13 single-stranded DNA which had been broken by freezing and thawing so that it did not need added primer, they were able to show that large amounts of incorporation would occur if rCTP, rATP, or rGTP, were used with the other 3 dNTP. When rUTP was used, there was no incorporation. Their data suggest that the fidelity of the ribosubstitution system is good, since the fingerprint pattern obtained after cleavage of DNA at rG residues gave all the expected spots and the expected intensities as compared to that of a T_1 digest of labeled RNA.

Later experiments by Lillehaug and Kleppe[44] have shown that raising the pH from 7.4 (commonly used for ribosubstitution) to 9.1 approximately doubles the rate of incorporation with rCTP, rATP, and rGTP. Under these conditions, rUMP was observed to be incorporated at about 25% of the rate and extent of the other ribonucleotides.

Wu et al.[1] showed that ribosubstitution can be used to repair the cohesive ends of λ DNA. In order to restrict the repair to the single-stranded cohesive ends without introducing labeled nucleotides into single-strand breaks in the duplex DNA, low temperature (5°C) and high salt concentration were needed to prevent localized melting at the breaks. Under these conditions, rCMP can replace dCMP and rGMP can replace dGMP in repairing the cohesive ends. The fidelity of the ribosubstitution seems high, since the same sequences were obtained whether the cohesive ends of λ DNA,[170] P2 DNA,[24] or 186 DNA[23] were repaired with all four deoxynucleotides or with three deoxynucleotides and one ribonucleotide. However, it was found that when two adjacent rCMP or two adjacent rGMP molecules were to be incorporated, the rate of incorporation dropped appreciably and repair synthesis often stopped at these points. It was pointed out that one can actually take advantage of this observation in attempting to isolate rather short oligonucleotides (average length of 16) at rC-rC or rG-rG blocks in order to simplify sequence analysis.

Using highly repetitive satellite DNA from the kangeroo rat (*Dipodomys ordii*) as template for DNA synthesis with ribosubstitution at 37°C, Salser et al.[199] observed that oligonucleotides obtained from ribosubstituted DNA corresponded very well to oligonucleotides obtained by similar cleavages of RNA transcribed from the template. These results were obtained for either rC or rG substitution, and suggest that the substitution occurs with good fidelity. Additional experiments with long templates of complex sequence, produced clean, reproducible groups of spots on electrophoretic separations.[232,233] These data also suggest that good fidelity can be achieved at temperatures at which ribonucleotide incorporation is not inhibited.

Van de Sande, Loewen, and Khorana,[45] making use of short synthetic deoxyoligonucleotides of defined sequence as template and primer, found that at 37°C ribosubstitution occurred with rather poor fidelity. When the temperature was lowered to 10°C, good fidelity of rC incorporation was observed but rG was still incorporated into incorrect positions. Extensive misincorporation of nucleotides was usually found during partial repair synthesis where the correct nucleotide was absent. For example, when only ^{32}P-TTP was used, an extra TMP was incorporated in place of dCMP. With complete repair synthesis, the extent of misincorporation was much lower. During complete incorporation, it should be more unlikely that an incorrect nucleotide would successfully compete with the correct deoxynucleotide at a given position. Also, Salser[3] has suggested that at 37°C, fidelity may be lower near the terminus of a template, as in the synthetic template experiments, where the polymerase-DNA interaction may be weakened.

In general, it appears that the fidelity of the ribosubstitution is sufficiently high for sequence analysis of the oligonucleotide products obtained by repair synthesis, especially after complete repair synthesis. In fact, even a small amount of misincorporation, if it is random, should not affect the accuracy of sequence analysis of the product.

Later sequencing experiments using the ribosubstitution technique have affirmed its usefulness.

Sanger et al.[46] independently used the primer approach and combined it with the ribosubstitution technique to determine a sequence of 50 nucleotides in bacteriophage f1 DNA. The amino acid sequence of the major coat protein of phage fd DNA had been reported.[47] It contained the amino acid stretch Trp-Met-Val. For Trp and Met as well as the first two nucleotides for Val, the amino acid sequence unambiguously defines the sequence of an octanucleotide. Thus, the sequence d(A-C-C-A-T-C-C-A) selected for the primer was expected to bind to exactly complementary DNA. Since the phages fd and f1 are very closely related, Sanger and his collaborators assumed that a synthetic primer, having the complementary nucleotide sequence as the messenger coding for these three amino acids, would bind to the DNA from either bacteriophage to allow sequence analysis by extension with DNA polymerase. Accordingly, the primer, d(A-C-C-A-T-C-C-A), was synthesized by Roychoudhury, Fischer, and Kössel[48] and was shown to serve as a primer using f1 DNA as template.[46]

The primer was extended using repair incorporation with one, two, three, or four deoxynucleoside triphosphates in the presence of magnesium ion.[46] It was also extended using the ribosubstitution technique to facilitate the analysis of longer sequences. In these experiments, rC was used, and product fragments were obtained by digestion with pancreatic RNase A. Sequences were determined by a variety of analysis techniques, including sequence determination by mobility of homologous series of oligonucleotides on two-dimensional homochromatography, a technique discussed in detail in Section B-2. The order of the DNA segments between each ribosubstitution was determined by a clever application of homochromatography, which separates oligonucleotides on the basis of size. As shown in Figure 5, the primer molecules, after extension to different lengths, were first separated on 2-D homochromatography, producing spots A through H. Each of these extended primers was digested with pancreatic RNase to cleave next to each rC residue in order to produce the constituent oligonucleotides, which were separated on one-dimensional homochromatography. For example, digestion of spot H produced rC-terminated oligonucleotides 1 and 7, and digestion of spot E

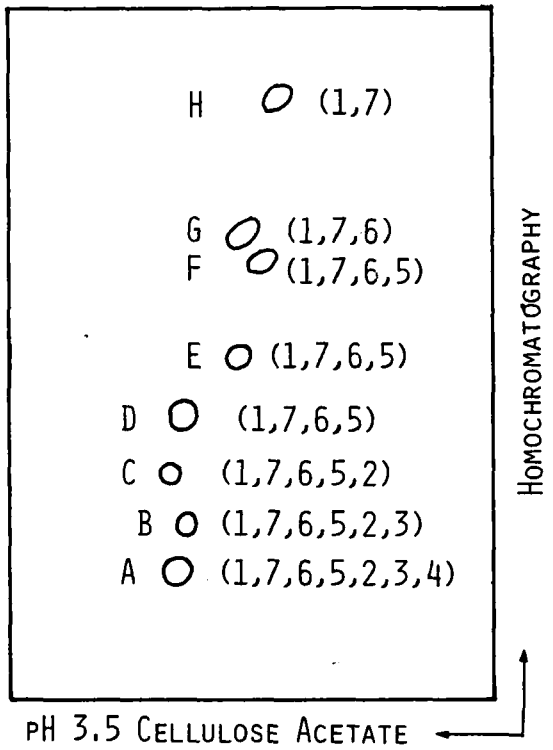


FIGURE 5. Fractionation of extended primer molecules synthesized by DNA polymerase I in the presence of the octanucleotide primer, Mn^{2+} and rCTP. This figure is a simplified version of that reported by Sanger et al.[46]

produced 1, 7, 6, 5, etc. Thus, the order of the 3′ rC-terminated segments was shown to be 1, 7, 6, 5, 2, 3, 4. The extended primers on the initial separation seemed to have discrete lengths, suggesting that the incorporation rate slowed down whenever a ribonucleotide had to be incorporated.

The 50-nucleotide-long sequence that was determined was not what would be expected from the published amino acid sequence of the coat protein. Later, amino/acid sequence determinations of proteins from the related bacteriophage ZJ-2[49] produced the sequence Trp-Ala-Met-Val for the corresponding region of the coat protein, and it appears that f1 may also have such an amino acid sequence in the coat protein. Thus, the synthetic octanucleotide primer, d(A-C-C-A-C-C-A), must be bound to some other region in the f1 DNA. An interesting feature of this 50-nucleotide-long sequence is that it is very rich in A residues, and there is considerable homology between different parts of the sequence.

Other investigators independently applied the primer approach to sequence specific regions of DNA. Loewen and Khorana[50] and Loewen,

Sekiya, and Khorana[51] made use of the tRNA sequence for the synthesis of d-oligonucleotide primers, and they have determined the sequences which follow the termination of transcription of the *E. coli* tyrosine tRNA (after the C-C-A end). The template used was the transducing phage ϕ80 psu III[52] which carries the suppressor tRNA gene. This gene is functional in vivo and should contain the normal control regions. Synthetic deoxyoligonucleotide primers were bound to the region adjacent to the 3′ end of the tRNA gene on the r-strand of ϕ80 psu III DNA. Extension from the 3′ end of the primers would thus include the transcriptional termination region. It was found that the sequence of the first 12 nucleotides could be determined on the basis of partial incorporation in the presence of combinations of dATP, dCTP, and dTTP. The sequence, primer-T-C-A-C-T-T-T-C-A-A-A-A, was obtained using these three labeled triphosphates. The extended primer was isolated and degraded by depurination, and the products were further purified. Standard enzymatic digestions and analyses were done to obtain the sequence. The sequence was confirmed by the behavior of the partial degradation products of the extended oligonucleotide on 2-D homochromatography. Isolation of the extended region for these experiments was facilitated by use of a primer which terminated in a 3′ rC. Selective ribonucleotide cleavage could then be used to separate the extended region from the primer. Sequence determination on 2-D homochromatography was easier when the shorter pieces were used.

Further elongation was performed in either of three ways: (a) The primer, elongated with A, C, and T, was repurified and elongated with G, T, and C. (b) The primer was elongated in a timed reaction using all four deoxyribonucleoside triphosphates. (c) The primer was elongated in a mixed ribo-deoxyribonucleotide incorporation, in which rC was used with d(T,G,A). The extended primer products were purified on gel electrophoresis, and then subjected to DNA sequencing procedures. Products which had been extended with rC were hydroloyzed with alkali. Products of the alkali digestion were purified on 2-D homochromatography, and then analyzed. The sequence, primer-T-C-A-C-T-T-T-C-A-A-A-A-G-T-C-C-C-T-G-A-A-C-T, was determined by the combination of all of these methods. The sequence possesses twofold symmetry, which may be involved in the process of termination recognition.

Similar procedures were used by Sekiya, Van Ormondt, and Khorana[53] to determine the sequence of the promoter region of the tyrosine tRNA gene. A synthetic oligodeoxynucleotide was annealed to the l-strand of ϕ80 psu III DNA and extended, using *E. coli* DNA polymerase I, into the promoter region. Analysis of labeled oligonuclotide digestion products of the extended primer led to the sequence primer-G-A-A-G-C-G-G-G-G-C-G-C-A-T-C-A-T-A-T-C-A-A-A-T-G-A-C-G-C-G-C-C-G-C. Several new innovations were applied for the analysis of this sequence. The primer was first extended by an rG at the 3′ end. Then, extension was allowed to go to completion in the presence of dATP, dGTP, and dCTP. It was possible to incorporate these 3 nucleotides into the first 13 positions beyond the 3′end of the primer before the requirement of dTTP, which was omitted in the repair reaction. The newly synthesized chain was isolated after alkaline hydrolysis, and sequenced by the mobility techniques and nearest neighbor analysis. Next, the primer was extended with dATP, dGTP, and rCTP. After the initial extension and removal of the free deoxynucleoside triphosphates, the oligonucleotide was elongated further by addition of dATP, dTTP, and dCTP. Eleven additional nucleotides were incorporated, which were isolated by alkaline cleavage and sequenced. Again, the primer, initially elongated with dATP, dGTP, and rCTP, was further extended using dATP, dCTP, dGTP, and a very low concentration (0.015 μM) of dTTP. Wu and Taylor[6] had shown that low concentrations of 1 of the 4 deoxynucleotides (0.05 to 0.001 μM) would allow repair synthesis at reduced rate. A 22-nucleotide-long segment was isolated after alkaline cleavage and sequenced.[53] As shown in Figure 6a, this sequence has large segments of twofold symmetry which allow formation of a prominent, single-stranded loop structure. The authors have pointed out that *E. coli* RNA polymerase binding sites in other DNA systems, bacteriophage λ,[54,55] bacteriophage fd RF,[56] and SV40,[57,58] all differ widely in primary sequence. They have concluded that it is likely that a single-stranded loop structure rather than the primary nucleotide sequence is recognized by the RNA polymerase. Such possible recognition site structures, first proposed by Gierer,[59] are seen more often as the sequences

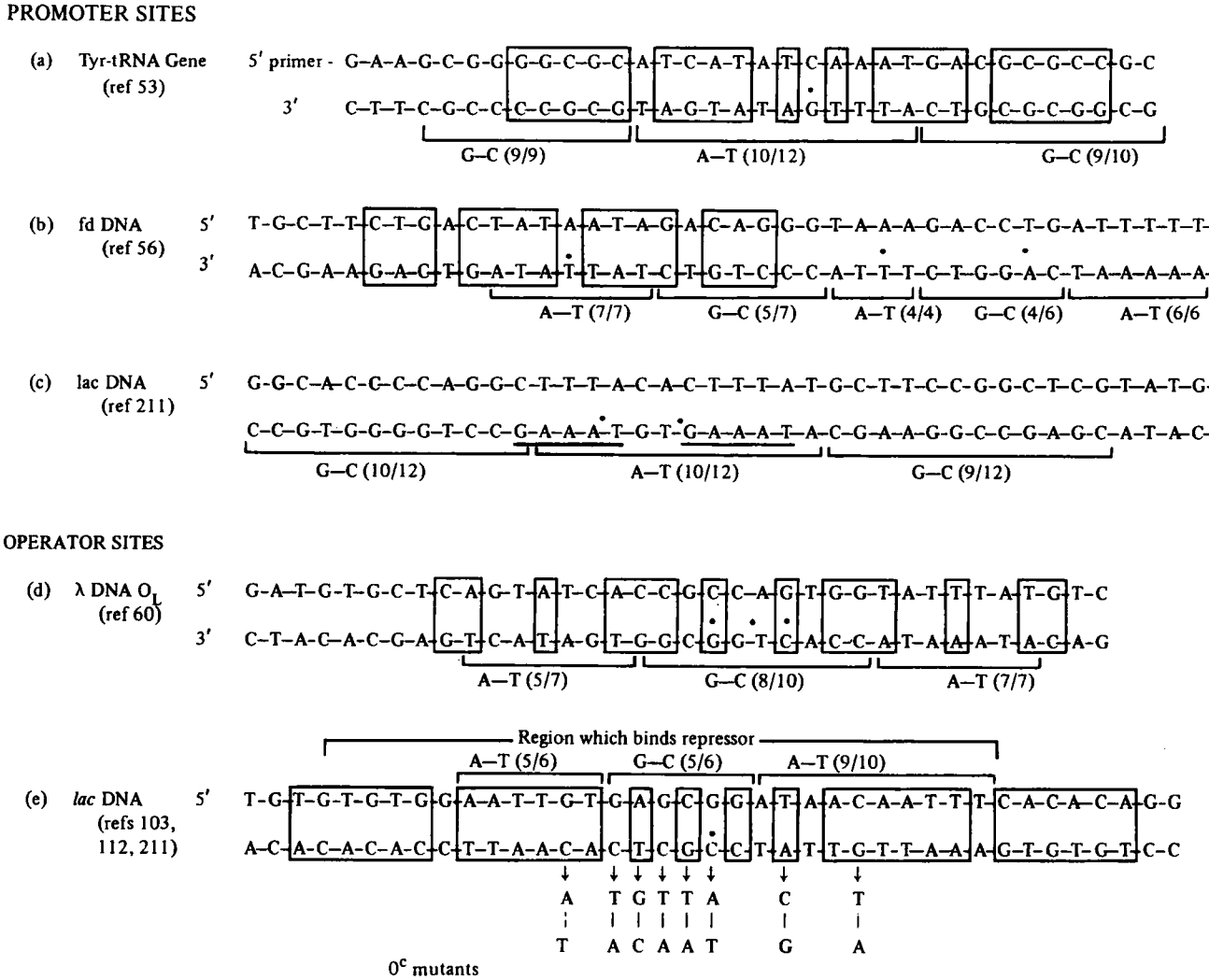


FIGURE 6. Promoter and operator sequences form various sources. The source of each sequence is given in the references. Regions o twofold symmetry are indicated by a central dot at the symmetry axis and boxes adjacent to symmetrical sequences. Regions of hig A-T or G-C content are indicated. The underlined sequences in the *lac* promoter represent repreated sequence G-A-A-A-T.

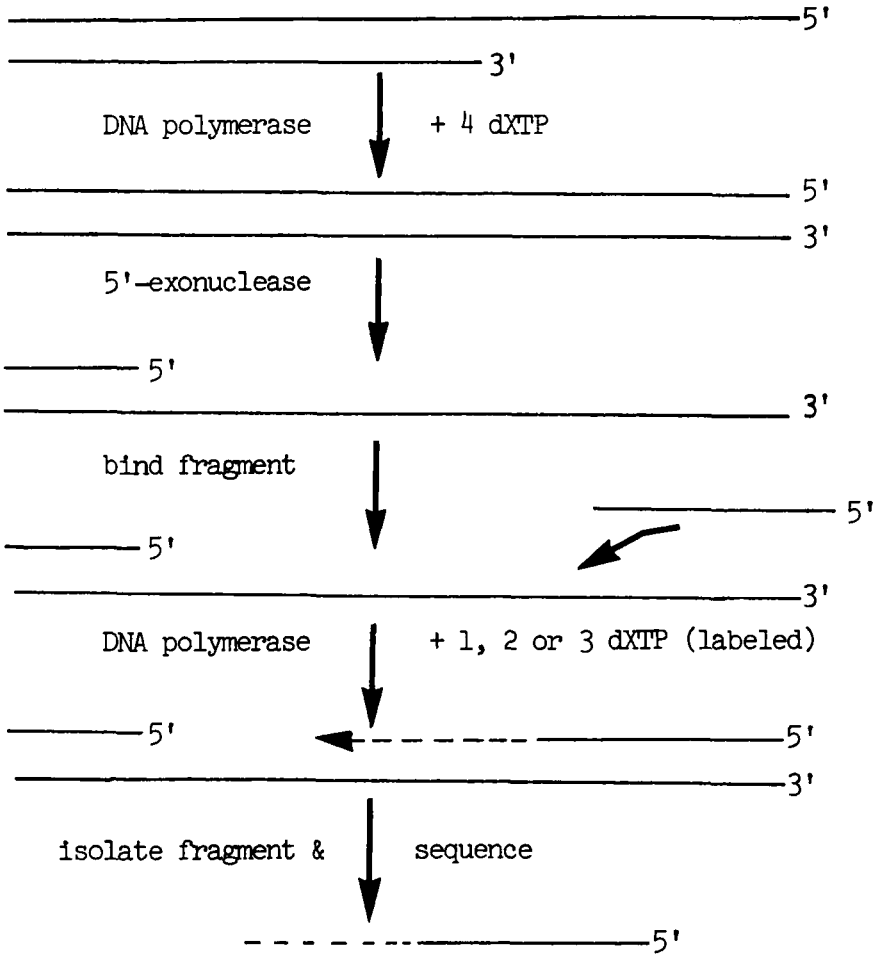


FIGURE 7. Summary of the procedure for sequence analysis by means of primer binding and extension at a terminus of duplex DNA. Bacteriophage termini may be made completely double-stranded by DNA polymerase-catalyzed repair. Double-stranded DNA is treated with 5' exonuclease to expose the DNA segment to be sequenced. A primer is bound to a terminal region of known sequence and extended with radioactive nucleotides over the region of interest. The labeled, extended primer is isolated and sequenced.

of control regions of DNA appear in the literature. These will be discussed in more detail in Section B-3-f.

An additional sequence was determined in the region preceding the initiation of transcription of the gene *N* in bacteriophage λ.[55] The same primer extension technique was used after specific binding of synthetic primer to the r-strand of λ DNA. Sequence information was obtained by mobility techniques, pyrimidine tract analysis, and nearest neighbor analysis. The sequence, primer-G-T-G-C-T-C-A-G-T-A-T-C-A-C-C-G-C-C, was found to lie in the center of the λ binding site sequence determined by Maniatis, Ptashne, Barrell, and Donelson[60] and is described in Section A-4.

Proudfoot and Brownlee[237] used rabbit globin mRNA as template, $(pT)_{12}OH$ as primer, and *E. coli* DNA polymerase I, in the presence of Mn^{2+} and four deoxynucleotides, to extend the primer in copying the mRNA sequence. With this method, a unique sequence, -A-U-U-G-C-poly A, was deduced for the globin mRNA.

It is also possible to use a variation of the primer binding and extension technique to sequence from the termini of double-stranded DNA. Bambara and Wu[61] used this technique to obtain some terminal sequence information from the right-hand 3' end of φ80 DNA. As shown in Figure 7, the cohesive ends were first completely repaired using DNA polymerase and unlabeled nucleoside triphosphates. Then, between 50 and 100 nucleotides were sequentially removed from the 5' ends of the DNA with λ exonuclease.[62] After the exonuclease treatment, a synthetic

dodecanucleotide[63] of the same sequence as the native right-hand cohesive end was bound to the complementary right-hand incorporated region, exposed by the λ exonuclease. The bound dodecanucleotide was extended by partial incorporation, isolated from the DNA on gel electrophoresis, and subjected to sequence analysis by enzymatic digestion. The sequence primer-CpGpTpA was determined by this method. This sequence is in agreement with information obtained by other methods.[15,17] A considerably longer sequence should be obtainable using this method.

Since terminal sequence information is relatively easy to obtain, the sequence required for the synthesis of a primer at the terminus of any double-stranded DNA could be determined. After λ exonuclease treatment, the primer could be bound for extension and further sequence determination.[61]

The use of various restriction endonucleases should allow the production of primer molecules from digestions of native DNA. Maniatis, Ptashne, Barrell, and Donelson[60] have used the primer technique to investigate the sequence of bacteriophage λ operators. The general features of this approach are outlined in Figure 8. When λ DNA was digested with Hind II + III, about 50 different double-stranded fragments were produced. One segment, 1,125 bases long, was used as a primer for extension (details will be discussed in Section 4-b-ii). This segment was first denatured, and the r-strand of this segment was annealed in the presence of equimolar quantities of purified 1-strand of λ DNA. The primer was extended under ribo-incorporation conditions worked out by Sanger et. al.[46] The extended segments (shown by dotted lines) were isolated by redigestion with Hind II + III or digestion with RNase, followed by purification. Standard DNA sequencing techniques were used to analyze the labeled, extended regions.

Sequences which have been determined are shown in Figure 6d. The region shows three possible axes of twofold symmetry, any of which may be involved in the recognition process by one or more protein molecules (e.g., repressor or RNA polymerase). There seem to be about 33 residues between the binding site for RNA polymerase, which occurs at the Hind site, and the beginning of transcription.

3. Use of Restriction Enzymes for DNA Sequence Analysis

A number of nucleases have recently been isolated which display an unusually high specificity toward DNA. These are the restriction endonucleases which recognize specific nucleotide sequences of double-stranded DNA of the length ranging from a tetra- to a hexanucleotide. The cleavage site appears so rarely that in most cases, after complete digestion, the products are still hundreds or thousands of nucleotides long. However, the use of a combination of these enzymes has yielded specific oligonucleotides 30 to 300 nucleotides in length, a useful size range for modern sequencing techniques. The use of restriction

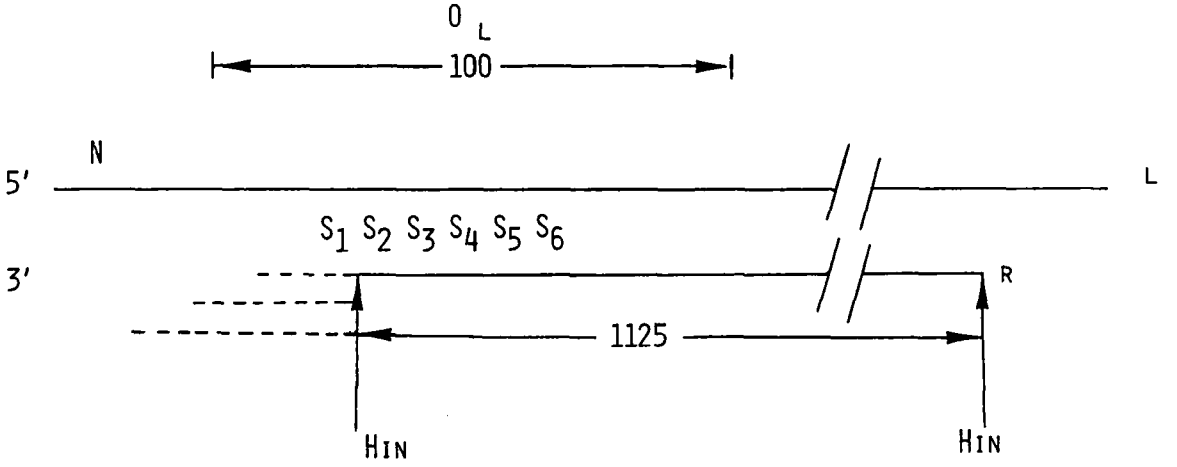


FIGURE 8. The use of a DNA fragment produced by restriction enzyme cleavage of λ DNA as primer for sequence analysis of the operator region of λ DNA.[54]

enzymes now appears to be one of the most promising means of specifically degrading large segments of DNA to pieces which can be effectively sequenced. Distinctive and reproducible patterns of DNA bands have been seen on acrylamide gels for the digestion products of SV40 DNA, of the replicative form of φX174 DNA, of the DNA from lambdoid phages, of adenovirus DNA, of polyoma DNA, and of DNA molecules from other organisms.

Bacterial restriction enzymes capable of degrading foreign DNA at a number of limited sites were first purified by Meselson and Yuan.[64] Later, Smith and Wilcox purified an endonuclease from *Haemophilus influenzae*[65] which is capable of degrading foreign DNA at specific sites[21] recognizing a specific hexanucleotide sequence. Since these two exciting findings, the search for other restriction enzymes and their use to assist in studying various problems in molecular biology have progressed at an immense pace. Because of the different sequence specificity of these restriction enzymes, various DNA molecules can be degraded into relatively short specific fragments with the use of a combination of these enzymes. For this reason, restriction enzymes are becoming increasingly important in the dissection and analysis of the structure and function of DNA molecules. Although a considerable amount of interesting information about the genetics and location of various functions on both bacterial and viral DNA genomes has recently been unraveled with the assistance of restriction enzymes, this material will not be dealt with in this article. A recent review article on restriction enzymes dealing specifically with these aspects should be consulted.[66] Instead, only the use of restriction enzymes immediately related to DNA nucleotide sequence analysis will be discussed here.

Restriction-modification systems can be detected by either genetic assays[67,68] or, more efficiently and accurately, by purely biochemical methods. Among the different biochemical methods for detecting restriction endonuclease activity by analysis of the digested DNA are (1) velocity centrifugation, (2) viscosity measurement, (3) gel electrophoresis on acrylamide, and (4) gel electrophoresis on agarose containing ethidium bromide. The first two methods have been used by earlier investigators. Both methods, velocity centrifugation in particular, suffer from the drawback that only a few assays can be performed at one time. In addition, both methods, viscosity measurement in particular, cannot distinguish degradation of DNA by specific restriction endonucleases from degradation by nonspecific endonucleases. Electrophoresis on either acrylamide or agarose is simple and fast, and numerous samples can be examined simultaneously. Usually, degradation of the DNA by nonspecific endonucleases can be distinguished from specific restriction endonuclease degradation because restriction products show discrete sharp bands on the gel. The sharpness of the bands can often provide information regarding other exo- or endonuclease contaminating activities. Electrophoresis on agarose gel containing ethidium bromide[69] has the further advantage that it is relatively insensitive to salt concentration, so that desalting after enzyme digestion of the DNA is not necessary prior to application onto the gel. Also, staining after electrophoresis in order to visualize the DNA bands is not necessary. Furthermore, very large DNA molecules or their fragments, which cannot enter acrylamide or acrylamide–agarose gels, are resolved in agarose gels.

Table 1 summarizes most of the present information on base-sequence specific restriction endonucleases which have been detected and isolated. The names of these enzymes are as suggested by Smith and Nathans.[70] The recognition sequences and the sites of cleavage which have been sequenced are also shown. It may be interesting to note that all of these recognition sites contain a twofold axis of symmetry. A few have been shown to make staggered breaks, 2 to 4 base pairs apart, leaving protruding 5′ ends.

One needs to solve two problems when using restriction enzymes for genetic or structural studies of DNA, as well as for DNA sequence analysis of a particular DNA molecule. The numbers of specific DNA fragments which are produced and the order of these fragments must be determined to give a physical map of the genome. Fragmentation patterns of many DNAs, degraded with a number of different enzymes, have been examined on gel electrophoresis. Sizes of DNAs for these studies have varied from the smaller SV40, polyoma, φX174, and fd DNA, to the larger λ, T_7 and adenovirus DNA, as well as some very large eucaryotic DNAs from human and calf thymus. However, only a few physical maps, which show the order of the DNA fragments of the smaller class DNAs produced by some of the

TABLE 1

Restriction Enzymes

Strain	Enzyme	Recognition sequence and cleavage site	Number of cleavage sites in DNA: λ	Ad2	SV40	References
Arthrobacter luteus	Alu I	–	>50	>50	23–27	82
Anabaena variabilis	Ava I	5′ –CGR↓YCG– –GCY↑RGC– 5′				83
Bacillus subtilis	Bsu X5	5′ –GG↓CC– –CC↑GG– 5′				83
Escherichia coli R_I (endo I^-, R^+)	Eco RI	5′ –G↓AATT–C– –C–TTAA↑G– 5′	5	5	1	9
	Eco RI′	5′ –AGA↓TCT– –TCT↑AGA– 5′ 5′ –GAA↓TTC– –CTT↑AAG– 5′				83
Escherichia coli R_{II} (fi^- R^-)	Eco RII	5′ –C↓CTG–G– G–GAC↑C– 5′	>35	>35	16	85, 89
Haemophilus aegyptius	Hae II	–	>30	>30	1	82
	Hae III	5′ –GG↓CC– CC↑GG– 5′	>50	>50	17	83, 86, 90
Haemophilus aphirophilus	Hap I	–	>30			83, 87, 88
	Hap II	5′ –C↓CG–G– –G–GC↑C– 5′	>50	>50	1	
Haemophilus gallinarum	Hga I	–	>50	>50	0	87, 88
Haemophilus haemoglobinophilus	Hhg I	5′ –GG↓CC– –CC↑GG– 5′	>50	>50	17	82
Haemophilus haemolyticus	Hha I	–	>50	>50	2	82
Haemophilus influenzae serotype b	Hinb	Same as Hind II + Hind III				91
	Hinb III	Same as Hind III	6	11	6	82
Haemophilus influenzae serotype c	Hinc II	Same as Hind II	34	>20	5	92

TABLE 1 (Continued)

Restriction Enzymes

Strain	Enzyme	Recognition sequence and cleavage site	Number of cleavage sites in DNA: λ	Ad2	SV40	References
Haemophilus influenzae serotype d	Hind II	5′ –GTY↓RAC– · · · · · · –CAR↑YTG– 5′	34	>20	5	21, 65
	Hind III	5′ –R↓AGCT–Y– · · · · · · –Y–TCGA↑R– 5′	6	11	6	65, 98, 2
Haemophilus influenzae serotype e	Hine	–				91
Haemophilus influenzae H-1	Hin H-1	–	>50	>50	1	94
Haemophilus parahaemolyticus	Hph I	–	>50	>50	4	91
Haemophilus parainfluenzae	Hpa I	5′ –GTT↓AAC– · · · · · · –CAA↑TTG– 5′	11	6	3	95, 69, 83, 72, 99
	Hpa II	5′ –C↓CG–G– · · · · · –G–GC↑C– 5′	>50	>50	1	95, 69, 83, 77
Haemophilus suis	Hsu I	Same as Hin II + Hin III				91
Moraxella nonliquefaciens	Mno I	5′ –C↓CG–G · · · · · G–GC↑C– 5′	>50	>50	1	82
Serratia marcescens SB	Sma I	–	3	12	0	96
Streptococcus faecalis var *zymogenes*	Sfa I	–	>30		>8	97
Streptomyces albus G	Sal I	–		2	0	82

The table is constructed from information compiled by R. Roberts and K. Murray. Y represents pyrimidine nucleotide T or C; R represents purine nucleotide A or G. Several restriction enzymes, Ava I, Eco $R_{I'}$, Hin II, and Hin III, recognize two nucleotide sequences, as shown in Column 3. It has not been shown whether each of these enzymes possesses two activities or, in fact, each contains two distinct enzymes which have not been separated. Hind in the text refers to Hind II plus Hind III.

restriction enzymes, have been completed. Some of these are shown in Figure 9.

The first physical map was reported by Danna and Nathans.[71] They observed that the mixture of enzymes Hind II and III from *Haemophilus influenzae* d cleaves SV40 DNA into 11 fragments (designated A to K). The SV40 DNA was isolated from African green monkey cells infected with small plaque SV40. To order these 11 fragments, they partially digested Form I SV40 DNA, which was uniformly labeled with ^{32}P in vivo, with Hind (mixture of Hind II and Hind III). The partial degradation products were separated on acrylamide gel electrophoresis, and subsequently

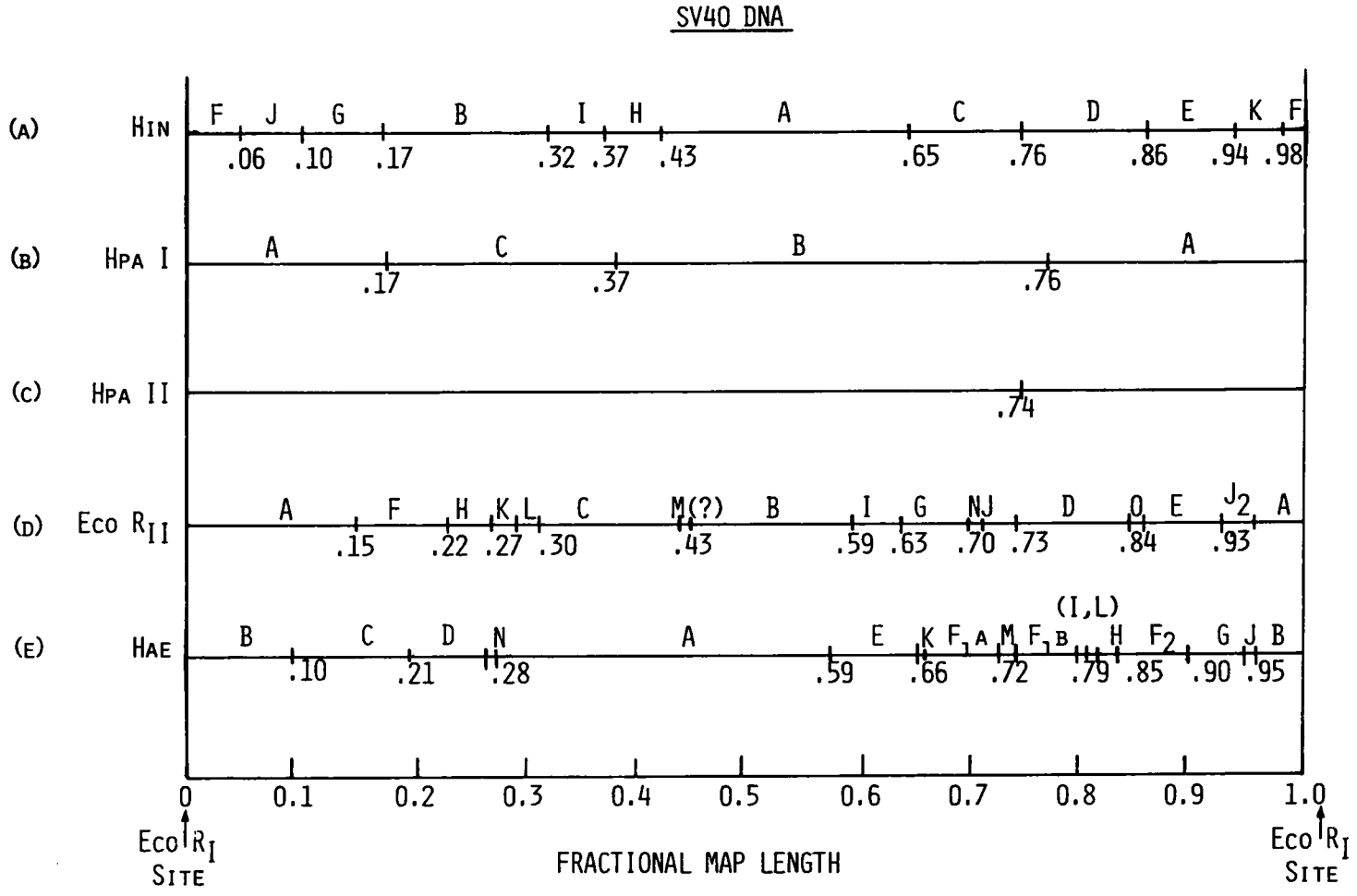


FIGURE 9. Physical maps of DNA molecules obtained after cleavage of DNA with restriction enzymes.

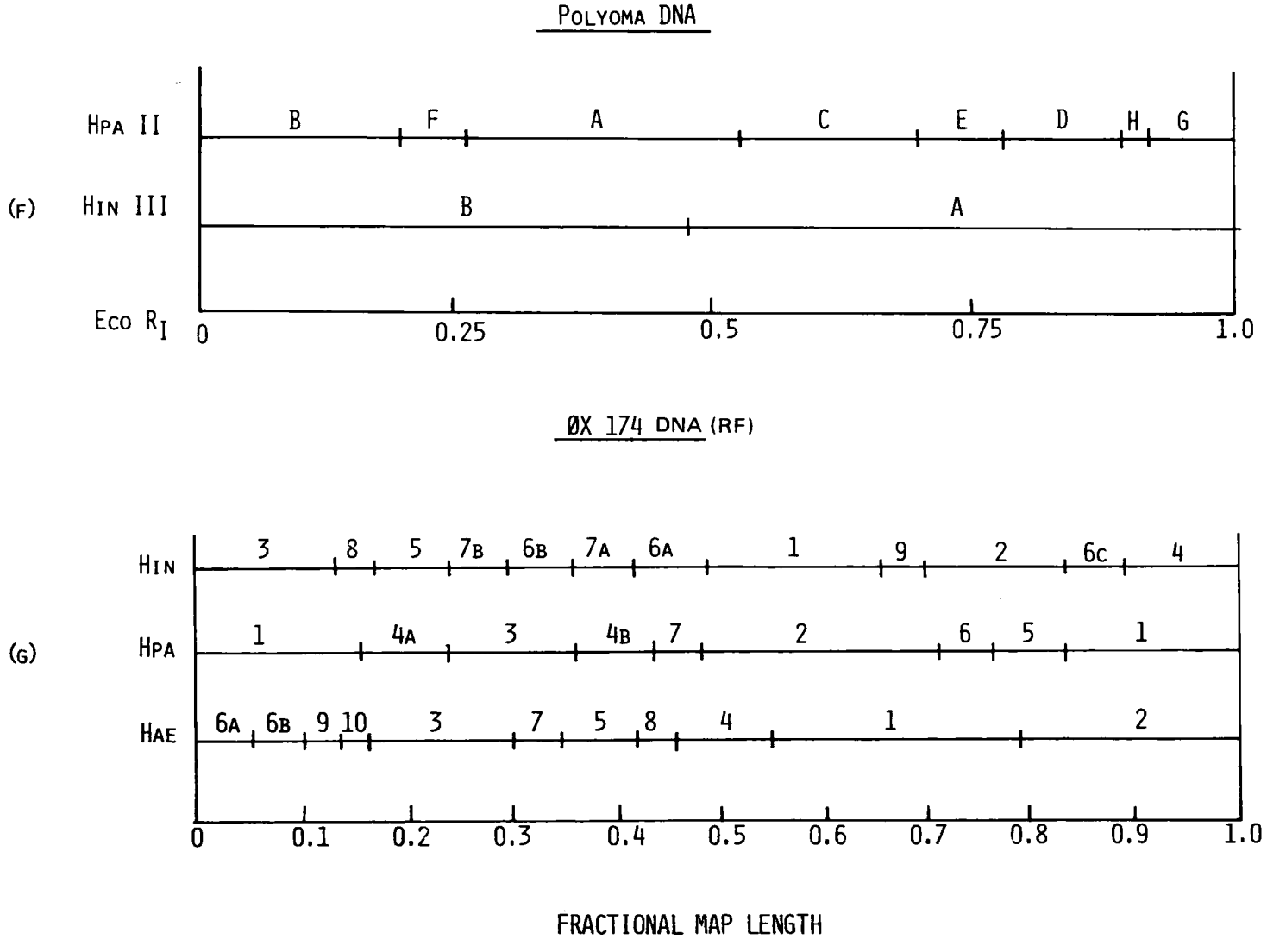


FIGURE 9 (continued)

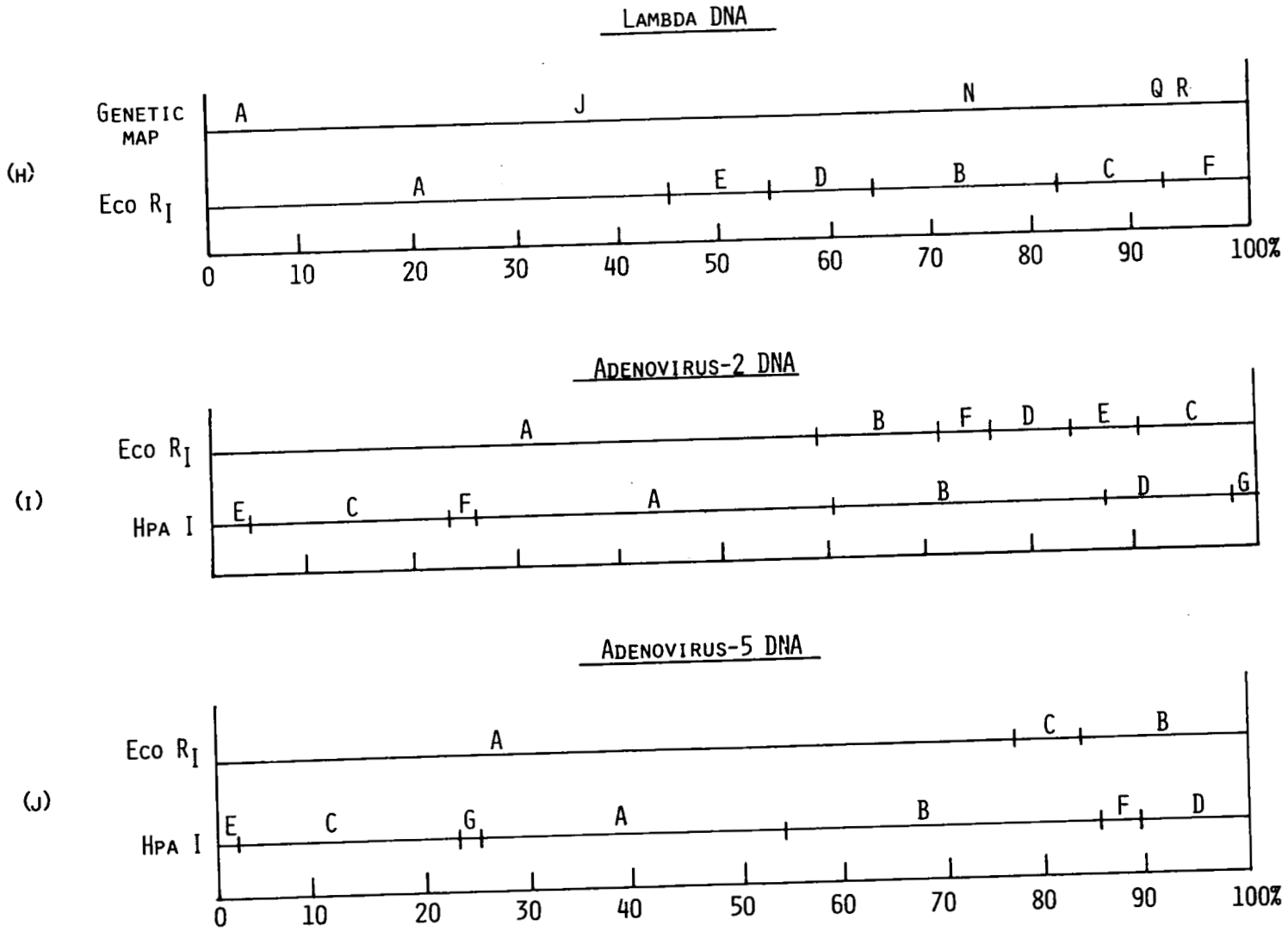


FIGURE 9 (continued)

isolated. By complete digestion of these partial degradation products to produce the final fragments, and by overlapping these fragments, the order of the 11 fragments was deduced (Figure 9A).

For circular DNA, such as SV40 and polyoma, a reference point is required around which the location of the various biological functions may be arranged in the genome. Such a reference point was provided when several investigators independently showed that EcoRI cleaves superhelical Form I SV40 DNA[71-73] and polyoma DNA[74] at one unique site. Thus, the location of any biological function or specific cleavage produced by another restriction enzyme can be defined as the distance from the EcoRI cleavage site, divided by the entire length of the genome. Danna et al.[75] have further shown that EcoRI cleaves SV40 within the Hind F fragment. The points of scission by Hind relative to the EcoRI site are also shown (Figure 9A). Following a similar approach, Griffin et al.[74] ordered the fragments produced from polyoma DNA by Hpa II. Eight fragments were produced, and the order of these fragments is shown in Figure 9F.

Determination of the order of the fragments produced by one enzyme can greatly assist in ordering the fragments produced by a second enzyme. This can be done by analyses of overlapping sets of fragments produced by the two different restriction enzymes. Subramanian et al.[76] have ordered SV40 fragments produced by EcoRII and Hae III. By redigesting the fragments produced by Hind[71] with EcoRII and vice versa, and by examining the products, they were able to deduce the order of the fragments produced by EcoRII, as shown in Figure 9D. Similarly, by redigesting the EcoRII fragments with Hae III and vice versa, the order of the Hae III fragments was also deduced (Figure 9E). Likewise, the order of fragments produced by Hpa I and Hpa II[75,77] was resolved (Figures 9B and 9C).

Using partial digestion and analyzing overlapping sets of fragments produced by different restriction enzymes, Lee and Sinsheimer[78] have also ordered the fragments from the replicative form ϕX174 DNA produced by Hind, Hpa, and Hae (Figure 9G).

A different method for ordering large restriction fragments was developed by Mulder et al.[79] Endonuclease R EcoRI digests DNA from human adenovirus serotype 2 and 5 (Ad2 and Ad5) into 6 and 3 fragments, respectively. Inman and Schnös[80] observed that denaturation maps of double-stranded DNA molecules after partial melting are highly characteristic of the DNA molecule, depending upon the A+T content throughout different regions of the molecule. Mulder et al.[79] made use of this observation to order the six Ad2 and three Ad5 EcoRI fragments by matching their individual partial denaturation maps against the denaturation map of intact Ad2 and Ad5 DNA, respectively. This method is applicable only if the restriction fragments are large enough to exhibit characteristic denaturation maps. The results thus obtained were confirmed by the method, mentioned above, of partial digestion and redigestion with another endonuclease. The composite cleavage map of Ad2 and Ad5 produced by EcoRI and HpaI is shown in Figures 9I and 9J.

Another method which makes use of well-characterized mutants to assist in ordering fragments produced by restriction enzymes has been employed by Allet et al.[81] One of the most studied and characterized DNA molecules is that from λ phage. Wild type λ DNA, upon digestion with EcoRI, produces six fragments. To order these six fragments, they made use of several well-characterized deletion, substitution, and insertion mutants of λ phage. By comparing the bands present, absent, or changed in position in digests of DNA from these mutants with those of wild type λ DNA, and knowing the point or region of mutation in the various mutants, the order shown in Figure 9H was established for λ DNA.

Thus far, the ordering of fragments produced by these restriction endonucleases has only been done on those endonucleases that generate relatively few fragments on a given DNA molecule. Most of these endonucleases give less than 18 fragments on the DNA tested. If the number of fragments produced by a given restriction enzyme is greater than 30, for mapping purpose, a second restriction enzyme would be required to first digest the DNA to give a few large fragments (e.g., 3 to 10) and then redigest each large fragment by the first restriction enzyme in order to provide complete ordering information.

One immediate use of restriction enzymes to assist DNA sequence analysis is to digest the large DNA segments into manageable, smaller sized fragments. Because of the very large size of DNA molecules, a major problem in sequence analysis

has been to obtain a homogeneous population of intact DNA. Owing to the length and rigidity of native, double-stranded DNA molecules, some breaks and nicks are usually introduced into the molecules during their purification. This breakage, together with their large size, usually interferes with most of the methods for sequence analysis. Thus, in many cases, it would be most helpful if a restriction endonuclease were used first to digest the DNA into specific, shorter fragments which could then be extensively purified. The fragment containing the region of interest may be further cleaved by a second restriction enzyme before analysis by any of the appropriate methods. Furthermore, a greater variety of methods can be applied for sequencing such relatively small, homogeneous DNA fragments.

An example of such an application of restriction enzymes for sequence analysis is the elegant work of Weissman and collaborators[76] using short fragments of SV40 DNA produced by Hind. They were able to sequence parts of some of the fragments by transcription and subsequent sequencing of the RNA transcripts. A more detailed account of this work is covered in Section B-3. Without the use of restriction enzymes to provide short, unique fragments of DNA, sequence analysis of the intact DNA would generally be very difficult.

Restriction enzymes can also be used to help in differentiating the two ends of a DNA molecule. The methods used for sequencing the terminal nucleotides of double-stranded DNA usually involve labeling the termini and partially digesting the DNA to provide short, labeled oligonucleotides for sequence analysis after 2-D fractionation. After labeling, the two strands of the DNA can be separated before partial digestion, or one of the two termini can be selectively labeled. These methods are rather tedious, and we believe that restriction enzymes would be most useful in providing an alternative. For example, both ends of a DNA molecule can be simultaneously labeled, and the labeled DNA can then be digested with a restriction endonuclease to produce several fragments. After fractionation by gel electrophoresis, the two labeled terminal fragments can be isolated and separately sequenced by further digestion and mapping.[93]

Incorporation of the label into the internal regions of the DNA because of breaks or nicks in the preparation of the DNA has been a problem with terminal labeling experiments. This internally labeled material, although present in relatively small amounts, if not separated from the terminal fragments, will appear as extraneous ghost spots on the 2-D maps for sequence analysis by mapping, and may confuse sequence determination of the termini. Digestion of the intact DNA with a restriction enzyme to generate two relatively short, terminally labeled fragments, purification by gel electrophoresis, and then partial digestion for mapping would effectively reduce such undesirable labeled impurities.

Jay, Roychoudhury, and Wu[93] have recently used this method for sequencing the termini of the SV40 DNA fragments produced by Hind. After terminal labeling, each fragment was digested with either Alu or Hae III, and the two labeled terminal fragments, after purification on acrylamide gel electrophoresis, were partially digested with pancreatic DNase for mapping. This technique has proven to be a considerable improvement over other terminal labeling techniques.

From Figure 9, it can be observed that by using a combination of different restriction enzymes, many parts of the SV40 genome can be degraded into very short pieces, 20 to 50 nucleotides long. Recently, efficient methods have been developed for sequence analysis of such small oligonucleotides by mapping. With the continuing search for and discovery of more of these restriction enzymes, any part of a DNA molecule will hopefully become available as short oligonucleotides, 20 to 50 bases long, by sequential digestion with these specific restriction endonucleases. Specific regions of the DNA molecule can then be efficiently sequenced.

Another method which has been studied and recently used extensively for sequencing internal regions of DNA molecules is the "primer extension" method with DNA polymerase (Section A-2). Using this method, DNA segments 50 to 100 nucleotides long have been sequenced by several investigators. A rather time-consuming step of this approach is the chemical synthesis of the primer for extension. Furthermore, prior knowledge or deduction of the sequence for the primer is required. Because of this problem, the primer extension approach has been restricted to specific regions of the DNA where the sequence of a primer can be deducted for synthesis. With the availability of restriction enzymes, natural primers for sequence analysis by extension with DNA

polymerase have become available for a number of regions of many DNA molecules. The DNA can be digested with an appropriate restriction enzyme to specific, short fragments. In principle, the DNA fragment adjacent to any region of interest can be isolated and used as a primer for extension. Thus, Thus, prior knowledge of the nucleotide sequence of the primer is not required.

Maniatis et al.[60] have already made use of this technique for sequence analysis of the region immediately preceding the start point of transcription of an operon in bacteriophage λ. Details of this experiment have already been described in Section A-2.

4. Isolation of DNA Fragments Corresponding to Protein Binding Sites

Some proteins bind specific segments of DNA, forming complexes so stable that the bound region is not susceptible to nuclease digestion. This fact has been exploited to specifically isolate protein binding sites from bulk DNA.

a. Ribosome Binding Sites

It was reported by Steitz[100] that ribosome binding to phage RNA could be used to protect specific segments of RNA from nuclease digestion. This procedure has been used to isolate a number of specific binding sites on bacteriophage RNAs.[101-104] All of these protected sites include the initiation codon A-U-G, as well as either or both of the sequences A-G-G-A and Pu-Pu-U-U-U-G-A. The sequences of some of these binding site regions are able to form hairpin configurations, which may be involved in the recognition of ribosomes during the initiation of the translational process.

It has also been possible to apply the same technique to obtain the sequence of a ribosome protected fragment from φX174 DNA.[105] Bretscher[106] found that the specific binding of *E. coli* ribosomes on phage DNA would occur if the correct complement of protein synthesis initiation factors, GTP and *N*-formylmethionyl-tRNA, were present. This work suggested that the ribosomes were recognizing and binding to the DNA at the region corresponding to the start signals for translation. Robertson et al.[105] used Bretscher's conditions to protect the single-stranded (+) strand of φX174 DNA, which contains the same sequence as the φX174 mRNA. After pancreatic DNase digestion, about 0.5% of the DNA sedimented with the ribosomes on sucrose gradient centrifugation.

The protected DNA segment gave a band on 12% acrylamide gel electrophoresis which suggested the size range of 40 to 50 nucleotides. Two-dimensional homochromatography separated the products into one major and several minor spots. The major fragment was sequenced by two methods: First, the major fragment was depurinated. The products were isolated and sequenced by the relative mobilities of their partial degradation products on two-diminsional homochromatography, as described by Ling.[107] Second, the major fragment was digested with T_4 endonuclease IV. This enzyme (described in detail in Section A-5) cleaved at most, but not all, of the -TpC- sequences present in this DNA fragment to produce oligonucleotide products with pC at the 5′ end and pT at the 3′ end. Two of the oligonucleotides were partially digested with venom phosphodiesterase, the products were separated, and each product was analyzed for 5′ end and depurination products. This procedure provided complete sequence information. The third oligonucleotide was irregular in size and could not be analyzed by this method. Instead, it was partially digested with streptodornase, a streptococcal nuclease, to four major products, which were then sequenced by techniques involving partial digestion with venom phosphodiesterase.

The final sequence shown in Figure 10a contains the trinucleotide which corresponds to the initiation codon A-T-G and the nucleotides coding for the first 9 amino acids of the N terminus of the φX174 spike protein. There is a sequence d(A-G-T-T-T-A-A) which parallels the Pu-Pu-U-U-U-G-A sequence found in the RNA bacteriophages. The binding site region is able to form a loop of secondary structure with the sequence A-T-G near the tip of the loop. The configuration is similar to those found in some of the RNA bacteriophages.

Segments of ribosome protected DNA were also isolated from bacteriophage f1 DNA by Robertson.[108] It was pointed out that ribosomes can be bound to heat-denatured double-stranded replicative form φX174 DNA. Binding in this case occurs at only 30% of the efficiency of that of the single-stranded DNA. The minus strands show no binding affinity.

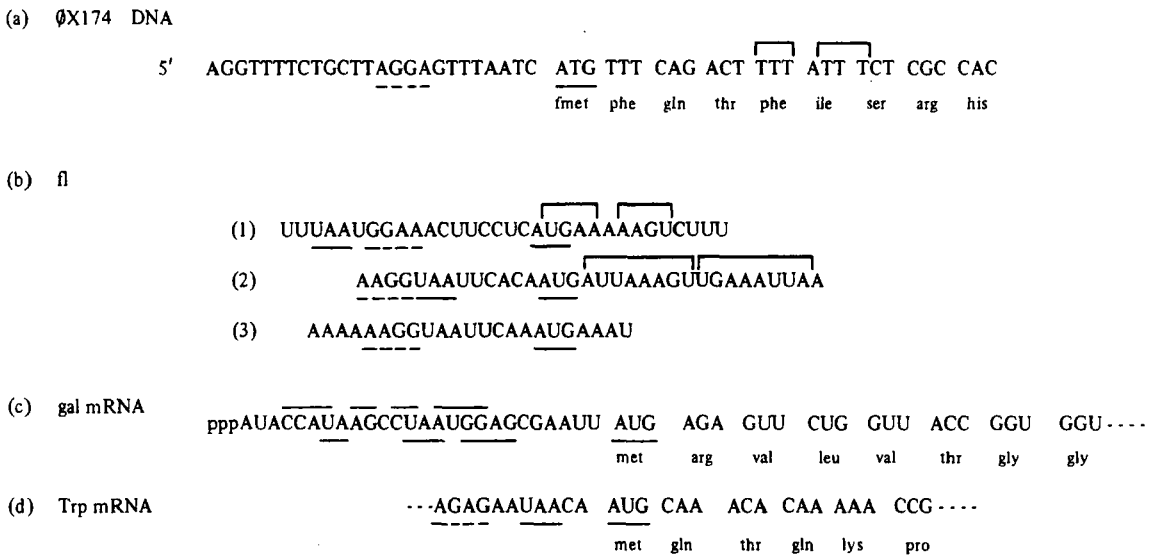


FIGURE 10. Ribosome binding site sequences, (a) and (b), and sequences which precede mRNA sequences, (c) and (d). The source of each sequence is listed. Underlined triplets are the termination signal (UAA) or the initiation signal (AUG). Sequences underlined by dotted lines are purine rich sequences which are located 7 to 9 nucleotides at the 5' side of AUG. Brackets on top of the sequence indicate true palindromes. Overlining indicates complementary palindromes.

b. Operator Sequences

Repressor proteins are known to bind very tightly to specific operator DNA segments. The binding is extremely specific, and it does not require other added factors, such as those required for binding of ribosomes to specific DNA segments.

i. The lac Operator

Transcription from the *E. coli lac* operon is controlled by the *lac* repressor–operator system. In this system, transcription is prevented by the binding of the repressor to the operator. Binding of lactose to the repressor, however, reduces its affinity for the operator and allows production of the proteins for lactose metabolism. Gilbert and Muller-Hill[109,110] made use of this affinity of the *lac* repressor to the lactose analogue, isopropylthiogalactoside (IPTG), to isolate the pure repressor protein. Later experiments showed that the *lac* repressor does not bind to denatured DNA or DNA from *lac* operator mutants.

The *lac* control region was expected to contain interesting sequence information. The region should contain the negative control site to which the repressor binds. It should also contain the positive control site to which the catabolic activator protein (CAP) binds. This protein binds to the *lac* control region to stimulate *lac* enzyme production when cyclic AMP levels are high. It appears that the position of CAP binding is also near the site for RNA polymerase initiation.[111]

Gilbert and Maxam[112] used the following procedure for the isolation of the *E. coli lac* operator region. DNA from bacteriophage strains containing the *E. coli lac* system (λp*lac*5 or λ*h*80d*lac*) was sonicated to segments about 1,000 base pairs long. The *lac* repressor was added, and the solution was passed through a nitrocellulose filter. The affinity of the protein to the filter allowed filter binding of only those fragments which contained the *lac* operator. The filters were then washed with a solution of IPTG, which caused the repressor molecules to release the *lac* operon containing DNA.[113]

The repressor was found to bind to the operator with a dissociation half-time of about 15 min. Pancreatic DNase was used to digest away all DNA except the repressor protected site. Less than 1% of the DNA could be retained by the repressor on a nitrocellulose filter after digestion. Release of this DNA fragment from the repressor could still be effected with IPTG. The apparent size of the segments was 27 base pairs.[112]

Although the operator was uniformly labeled in vivo with ^{32}P during the isolation and purification, the final amount of operator contained too little radioactivity for direct sequence analysis, except for pyrimidine tract analysis. Therefore, direct DNA sequence analysis was not done.

Instead, microgram quantities of the unlabeled operator region were isolated and then transcribed with RNA polymerase, and the radioactive RNA transcript was sequenced (details of this procedure are discussed in Section B-3).

ii. The λ Operators

The λ operators control synthesis of early mRNA from bacteriophage λ. Isolation of the λ operator sequences was achieved by the binding of the λ repressor to λDNA, followed by digestion of unbound DNA.

The λ repressor was shown to cosediment with λDNA in sucrose gradients,[114,115] but not with λ DNA lacking the λ operators or possessing mutated λ operators.[116] The repressor binds to two operators, O_L and O_R, blocking normal transcription from the "leftward" and "rightward" promoters. The binding is independent, so each operator could be studied separately using appropriate operator mutants.

After repressor binding, the highly labeled λ DNA was digested with DNase. The repressor operator complexes were purified by nitrocellulose binding and gel electrophoresis. Maniatis and Ptashne[117] found that, as the ratio of repressor to operator was increased, six successively larger DNA segments were recovered from each operator (see Figure 8). The smallest segment was about 35 nucleotides long. Segment sizes increased by increments of about 15 nucleotides to about 105 nucleotides. Pyrimidine tract sequences of the O_L fragments were not identical to those of the O_R fragments, as expected on the basis of the observed difference in repressor affinity for the two regions. The pyrimidine tract fingerprints became more complex for DNA segments isolated from the O_L region as the amount of bound repressor was increased. This suggested that the binding site sequences were different for the later-binding repressor molecules.

The O_L system is fortuitously cleaved by the restriction enzyme from *Hemophilus influenzae* (Hin) to form 2 DNA segments, Hin 320 and Hin 1125 (numbers refer to length in nucleotides), each with binding affinity for λ repressor.[118] After repressor binding and DNase digestion, a fragment 30 bases long was recovered from the Hin 320, apparently the primary binding site (S_1). Fragments varying in size from 30 to 75 bases long were recovered from Hin 1125, apparently containing segments of the secondary binding sites (S_2 to S_6) (see Figure 8). Correct arrangement of the Hin fragments placed S_1 adjacent to the *N* gene, with the S_2 to S_6 sites distributed further away from the gene. A similar arrangement was found for the multiple repressor sites in the O_R region.

It also appears that the RNA polymerase binding site sequence is superimposed on the operator sequence.[118] RNA polymerase was able to protect the Hin sites in O_L and O_R from cleavage, although expected cleavages occurred at other Hin sites in the λ DNA molecule.

Sequence information in the O_L region was obtained by annealing with denatured Hin 1125 fragment to the *l*-strand of λ DNA, extending the Hin primer with DNA polymerase in the presence of labeled nucleoside triphosphates, and analyzing the labeled oligonucleotide products.[60] Details of the sequence analysis techniques and the results are given in Section A-2.

c. RNA Polymerase Binding Sites

In order to study promoter region sequences, one of the approaches used is the isolation of the DNA segment protected by specific binding of *E coli* RNA polymerase. A problem with this technique has been that any large DNA segment is likely to contain a number of RNA polymerase binding sites of variable affinities, so that isolation of a single protected site has not been a simple operation (see Reference 119 for earlier reports).

The most successful attempt has been the isolation of a single RNA polymerase binding site from the double-stranded bacteriophage fd replicative form DNA by Heyden, Nusslein, and Schaller.[119] The binding of RNA polymerase at 37°C in the presence of sigma factor is very stable, with a half-life of about 10 hr. Lowering the temperature or the amount of sigma factor reduces binding. When the RNA polymerase bound fd DNA is digested with pancreatic DNase, segments of DNA protected by RNA polymerase can be isolated by gel filtration. The number of nucleotides protected per fd DNA depends on the ratio of moles of RNA polymerase per moles of DNA, up to a ratio of 5.

For sequence analysis, the DNA protected against DNase digestion (pDNA) was isolated after complexing either 2 RNA polymerase molecules ($pDNA_2$) or 4.8 RNA polymerase molecules ($pDNA_5$). The DNA for these experiments was labeled with ^{32}P in the (+) strand and ^{3}H in the (-) strand. When $pDNA_2$ was depurinated, it could

be separated out into 6 products from the (+) strand and 7 from the (-) strand. Separation of pyrimidine tracts was done on a polyethyleneimine plate according to the method of Southern and Mitchell.[120] A 2 *M* pyridinium formate pH 3.6 solvent is used in the first dimension, and a 1 *M* LiCl solvent is used in the second dimension. Fingerprints of $pDNA_5$ show additional strong spots where there were weak spots in the $pDNA_2$ fingerprint. This suggests that the additional site bound in $pDNA_5$ is a distinct sequence with a lower affinity for RNA polymerase. More recently, the sequence of the primary binding site has been determined on the basis of sequence analysis and alignment of pyrimidine tracts from both strands.[56] For some experiments, the (+) strand was ^{32}P labeled in vivo; for others, the (-) strand was ^{32}P labeled in an in vitro synthesis from the (+) strand. [α-^{32}P]dCTP and [α-^{32}P]dTTP were used for labeling so that nearest neighbors and the sequence of the pyrimidine tracts could be determined. [α-^{32}P]dGTP and [α-^{32}P]dATP were used to detect purines neighboring the 3′ ends of pyrimidine tracts. The compositions of the pyrimidine tracts were determined from their positions on the two-dimensional fractionation. The sequences were determined by 5′, 3′ termini and nearest neighbor analyses (see Section B-1). Long oligopyrimidines were partially degraded by micrococcal nuclease, reseparated, and then analyzed. Further alignment of oligopyrimidines on the two strands was done by the Watson-Crick base pairing rules. Additional alignment was achieved by analysis of labeled products after partial exonuclease III digestion, followed by incorporation of radioactive nucleotides using DNA polymerase I. Finally, sequence information was obtained from analysis of RNA, initiated and synthesized from the binding site, and from analysis of λ exonuclease digestion products.

The total sequence of the strong RNA polymerase binding site, shown in Figure 6b, has several regions of hyphenated symmetry which may be involved in the recognition process.

5. T_4 Endonuclease IV

Of all the enzymes which are able to degrade DNA to small oligonucleotides, only one, endoculease IV, displays considerable cleavage specificity. Although restriction enzymes can cleave DNA at very specific sites, the products are usually too large for direct sequence analysis. This enzyme, however, can produce a wide size range of specific products.

The existence of RNA base-specific endonucleases, such as pancreatic RNase and T_1 RNase, has been the prime reason for the success of the presently well-established methods for RNA sequence analysis. Pancreatic RNase splits RNA specifically after pyrimidines, while T_1 RNase splits RNA specifically after G residues. Up to the last 2 years, however, no DNA-specific endonuclease with a sufficient base specificity for use in DNA sequence analysis had been isolated.

Because of the lack of such specificities among the DNA-specific endonucleases, nucleases such as pancreatic DNase[123] and micrococcal nuclease[5,6] had been used to degrade large oligonucleotides or DNA segments to smaller oligonucleotides for analysis. Although micrococcal nuclease has been observed to possess some specificity, cleaving preferentially at NpA and NpT bonds,[124] degradations of even small, labeled DNA segments by these two endonucleases yielded rather complicated mixtures of oligonucleotide products. Purification of these products was often difficult simply because of their large numbers.

Sadowski and Hurwitz[125] have discovered a DNA endonuclease which does show considerable promise for use as a base-specific nuclease for DNA sequence studies. This enzyme, T_4 endonuclease IV, was found to cleave single-stranded DNA, leaving a cytidine at the 5′ end for 95% of the product oligonucleotides.

The 3′ terminal specificity of the enzyme still appears to be somewhat unclear. Ziff, Sedat, and Galibert[130] found a strong preference for cleavage of T-C bonds in sequence analysis of relatively few, long, specific oligonucleotides. Other evidence suggests that the 3′ termini may be produced in only a partially specific fashion. Sadowski[127] found T to be the 3′ end in about 50% of the fragments obtained after digesting bulk DNA with endonuclease IV. Roychoudhury and Wu[128] used different levels of endonuclease IV (a gift of Dr. Sadowski) for the digestion of denatured ϕ80 DNA to give products in the average size range of 20 to 300 nucleotides long. These products were then labeled at the 3′ ends with a single ^{32}P-rGMP residue, using deoxynucleotidyl terminal transferase (see Section B-4-a). Nearest neighbor analysis of the products showed the presence of all 4 nucleotides at the 3′ termini, with

the following distribution: T (50 to 60%), G (25 to 30%), A (15 to 20%) and C (5 to 10%).

The enzyme may, however, be more specific than the data suggest. The standard preparation of the enzyme showed some 3′ exonuclease activity. When oligonucleotide fragments terminally labeled with rGMP by transferase were incubated with endonuclease IV for a short period (10 to 20 min), 3′ terminal rGMP was liberated as a mononucleotide (no oligonucleotides were found). Attempts to inhibit this exonuclease activity with high levels of dNMP or with spermidine[128] were not successful. Thus, because of this 3′ exonuclease activity, the actual extent of specificity of the T_4 endonuclease IV could not be determined accurately. The lower degree of specificity in the digestion of bulk DNA may also be related to the presence of many different types of susceptible sequences, some of which are cleaved less specifically.

In spite of such difficulties with the absolute specificity of this enzyme, it has already proven useful for DNA sequence analysis. In an initial study, Ling[129] showed that a partial digestion of fd phage DNA yielded products which gave discrete bands on polyacrylamide gel electrophoresis. Such encouraging results stimulated further use of this endonuclease. The level of digestion used by Ling, however, still produced too many products for complete separation.

Ziff, Sedat, and Galibert[130] and Galibert, Sedat, and Ziff[131] used milder conditions for digestion of ϕX174 DNA. Under such conditions, the ϕX174 DNA was expected to assume a secondary structure in which large sections of the molecule would be insusceptible to cleavage by endonuclease IV. After digestion, separation of the labeled products on polyacrylamide slab gel produced only a small number of sharp bands. A number of these bands were isolated, purified further on homochromatography, and further digested with endonuclease IV. The original oligonucleotides, in the 50 to 80-nucleotide size range, were each degraded to the 2 to 15 size range for separation on 2-D homochromatography. Small oligonucleotide products were subjected to sequence analysis either by mobility on 2-D homochromatography followed by additional confirmatory analyses, or by the exonuclease I technique, which will be described in Section B-4-b. Since cleavage was primarily between T-C bonds, analysis of pyrimidine tracts provided the overlap sequences needed to order these short fragments. In the second digestion of a given large oligonucleotide by endonuclease IV, there were relatively few susceptible bonds, and not all the T-C bonds were cleaved. It is possible that the enzyme recognizes a sequence longer than the T-C dinucleotide. The insusceptibility of some T-C bonds does not appear to be caused by hydrogen bond formation producing double-strand character.

B. SEQUENCE ANALYSIS OF OLIGONUCLEOTIDES OR SHORT SEGMENTS OF DNA

1. Analysis after Further Degradation of the Oligonucleotides or DNA Segments

a. Chemical Degradation

The study of the primary structure of polydeoxyribonucleotides was facilitated by the observation that the glycosidic linkage of purine deoxyribonucleotides is labile under conditions where it is stable in pyrimidine deoxyribonucleotides, and vice versa under a different set of conditions. These facts have been used extensively to degrade DNA molecules into oligonucleotides containing either pyrimidines or purines only. To date, this is one of the few technical advantages to DNA sequencing not available to RNA sequencing. However, this type of degradative approach is useful mainly to provide supplementary sequence information. Unless the sequence is very simple, a complete nucleotide sequence of a specific region of DNA containing both purines and pyrimidines cannot be determined by these techniques alone. Furthermore, the location of the purine and pyrimidine tracts in the DNA molecule cannot be pinpointed by this approach.

i. Pyrimidine Tracts (Depurination Products)

When DNA is reacted with dilute sulfuric acid[132] or formic acid,[133] there is a slow acid hydrolysis of the glycosidic linkages between the various purines and their deoxyribose residues to give apurinic acids. In these products, the pyrimidine bases are still in place, but the purine bases are lost and the deoxyribose can exist in its open chain form. A more vigorous acid treatment or the addition of diphenylamine results in the cleavage of both the 3′ and 5′ phosphodiester bonds of the reacted deoxyriboses by a series of β-elimination reactions, giving a mixture of pyrimidine oligonuc-

leotides which carry a terminal phosphate at both the 3′ and 5′ positions. These pyrimidine tracts (or depurination products) have been isolated from many species of DNA and can be separated by a variety of chromatographic and electrophoretic procedures.[134-139] Studies such as these allow the comparison of the frequency of specific pyrimidine tracts present in related bacteriophages.

Studies on the pyrimidine tracts of fd DNA well illustrate the sequence information that can now be obtained using this method. By combining the two-dimensional electrophoresis and thin layer separation techniques with partial spleen and venom phosphodiesterase digestions, Ling[129] determined the complete sequence of a 20-nucleotide long pyrimidine oligonucleotide from fd DNA. In addition, the sequences of several other large (9 bases or longer) pyrimidine oligonucleotides from the DNA of bacteriophages fd, f1, and ϕX174 were determined by the same method.[140]

By analysis of the depurination products of the two separated strands of guinea pig α-satellite DNA, the -C-C-C-T- sequence was found to be predominant in the L-strand, and the -T-T sequence was predominant in the H-strand. Thus, it was deduced by Southern that -T-T-A-G-G-G- was the predominant sequence in the H-strand, and -C-C-C-T-A-A- in the L-strand.[141] Similar studies of the satellite DNA of mouse L cells showed a predominance of the sequences T-T-T-T-C-C, T-T-T-C-T-C, and T-T-T-T-T-C in the heavy strand and C-T-T in the light strand.[142]

ii. Purine Tracts (Depyrimidination Products)

When DNA is treated with anhydrous hydrazine followed by alkaline hydrolysis, a mixture of purine tracts is obtained.[143-147] The preparation of purine tracts is usually not as easy to reproduce as that of the pyrimidine tracts, and therefore extensive use of purine tracts for sequence analysis has not yet been made.

b. Enzymatic Degradation

i. Sequence Analysis of Labeled Oligonucleotides Synthesized from Native DNA in vitro

In general, the analysis consists of the digestion of the labeled oligonucleotide to mononucleotides, which are subsequently separated and quantitated by counting their radioactivity. Usually, a series of different types of labeling and digestion is required to unambiguously determine the sequence of an oligonucleotide. An example of how such labeled oligonucleotides may be obtained and analyzed is given as follows:

Segments of labeled DNA can be obtained by DNA polymerase I-catalyzed repair synthesis (see Section A-2). DNA polymerase, in the presence of radioactive deoxynucleoside triphosphates (^{32}P or ^{3}H labeled), is used to extend a primer, copying the DNA template to be sequenced. In this way, the sequence of the template is converted to a complementary sequence of radioactive DNA. Partial endonucleolytic degradation of the labeled DNA produces short, labeled oligonucleotides. The labeled oligonucleotides can be isolated, radiochemically pure, by a variety of chromatographic or electrophoretic techniques.

Figure 11 illustrates how the sequence of an oligonucleotide may be determined by enzymatic digestion. The tetranucleotide chosen for the illustration has all 4 bases labeled with ^{3}H, and the phosphate at the 5′ side of the cytidine has a ^{32}P label generated by the use of [α-^{32}P]dCTP for the repair reaction. Digestion with micrococcal nuclease produces endonucleolytic cleavages in the oligonucleotide, leaving 3′ phosphate terminated oligonucleotide products. Spleen phosphodiesterase degrades them sequentially from the 5′ termini to produce 3′ phosphate terminated mononucleotides.[8] In this case, the mononucleotides dAp, dGp* (where the asterisk indicates ^{32}P), and dCp as well as the nucleoside T-OH are produced. The presence of dG$\overset{*}{p}$ as a product indicates that there must have been a d-G$\overset{*}{p}$C sequence in the original oligonucleotide. This procedure is called nearest neighbor analysis.[148] The presence of the T-OH as a nucleoside indicates that T must have been the 3′ terminal nucleoside in the original oligonucleotide. We refer to this procedure as 3′ end analysis. The identification and quantitation of the tritium labeled nucleotides give the composition of the oligonucleotide. The sum of these data indicates a sequence of d(A-G-C-T) or d(G-C-A-T) for the original oligonucleotide.

A separate digestion from the 3′ termini with venom phosphodiesterase produced the 5′ phosphate terminated mononucleotides dp$\dot{\text{T}}$, d$\overset{*}{p}\dot{\text{C}}$, and dp$\dot{\text{G}}$ and the mononucleoside d$\dot{\text{A}}$. Thus, d$\dot{\text{A}}$ was the 5′ terminal nucleoside in the original oligonucleotide. We refer to this procedure as 5′ end analysis. The sum of data from both digestions defines the sequence d(A-G-C-T).

Sequences of certain larger oligonucleotides can

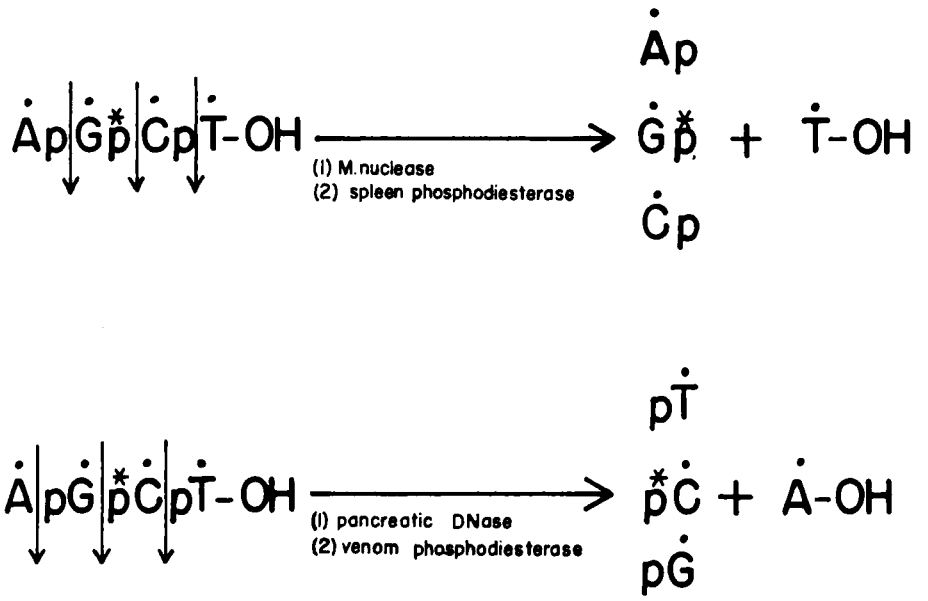


FIGURE 11. Standard procedure for sequence analysis of a radioactive oligonucleotide by enzymatic digestion. Nucleotides with a dot (Ṅ) are labeled with ^{3}H. Phosphates with an asterisk (p*) are labeled with ^{32}P. After digestion, products are separated on standard chromatographic systems.[5] Products are quantitated by counting of radioactivity. Other details of the analysis procedures are given in the text.

also be determined by this method, especially if it can be applied to a series of partial degradation products of the original oligonucleotide.[1,4,9]

These techniques were first used on DNA to determine the cohesive end sequence of λ DNA.[5,6,8] The single-stranded ends of λ DNA were used as templates for DNA polymerase I-catalyzed repair synthesis, using radioactive nucleoside triphosphates (see Figure 1). Random endonucleolytic digestion with micrococcal nuclease produced labeled oligonucleotides which were purified on electrophoresis. The complete sequences of many of the purified oligonucleotides were determined by these enzymatic digestion techniques. Other investigators have also used this method for analysis, but ^{3}H-labeled nucleotides were omitted and different ^{32}P-labeled nucleotides were used for different incubations. The nearest neighbor analysis on each short ^{32}P-oligonucleotide, labeled with each of the 4 nucleotides in turn, can often give the complete sequence of oligonucleotides 3 to 8 nucleotides long.

Whitcome, Fry, and Salser[2,3,2] have developed a simple method for sequence determination of short oligonucleotides with the use of micrococcal nuclease. Short, ^{32}P-labeled oligonucleotide fragments were generated from ribosubstituted DNA after cleavage at the ribonucleotide sites either by enzymatic digestion or alkali treatment, followed by separation on 2-D electrophoresis. The base composition of the ^{32}P-oligonucleotide could be deduced from the position on the 2-D fingerprint. The sequence of the oligonucleotide was determined after complete digestion with micrococcal nuclease to give mono- and dinucleotides, which were fractionated on 2-D polyethyleneimine thin layer chromatography. By using UV-absorbing markers of all possible mononucleotides and dinucleotides, the composition of the ^{32}P-labeled mono- and dinucleotides could be identified. The sequence of the original oligonucleotide could then be deduced from the ^{32}P counts in the mono- and dinucleotides from a specifically ^{32}P-labeled fragment. For example, an oligonucleotide which has a composition of $(TA_2)rG$ was generated by DNA polymerase-catalyzed repair synthesis using α-^{32}P]TTP, rGTP, dATP, and dCTP, followed by alkali cleavage. The micrococcal nuclease digest of this tetranucleotide yielded mainly ^{32}P-labeled ApAp* and some TprGp. The dinucleotide sequence ApAp* indicated the presence of ApAp*T, so that the oligonucleotide must have the sequence ApAp*TprGp. In this way, most short oligonucleotides can be sequenced, particularly if different labeled triphosphates are used in different experiments.

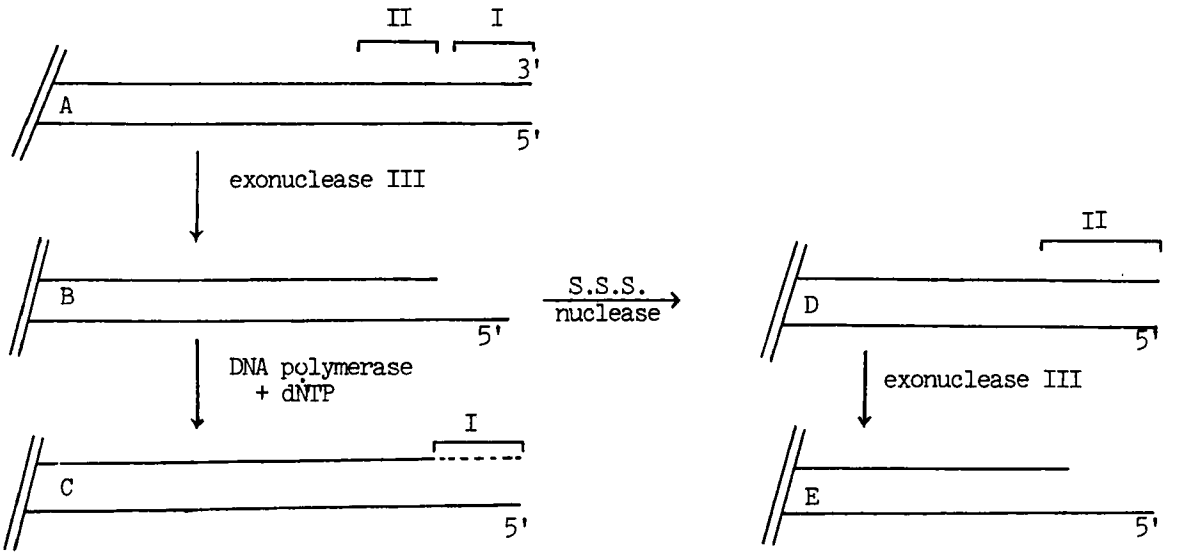


FIGURE 12. Terminal sequencing techniques involving exonuclease III and single-strand specific nucleases (S.S.S. nuclease). Regions I and II near the terminus of double-stranded DNA may be sequenced by these procedures. Region I in molecule (A) is rendered single-stranded by partial digestion with exonuclease III as described in the text. The single-stranded region in molecule (b) is repaired with radioactive nucleotides to produce (C) which can be sequenced. To obtain the sequence of region II, the single-stranded region in molecule (B) is removed by a S.S.S. nuclease to produce molecule (D). The DNA is again partially digested with exonuclease III, rendering region II single-stranded to give molecule (E), which can then be repaired with radioactive nucleotides for sequence analysis.

ii. Specificity of DNases

A number of nucleases have some cleavage specificities toward DNA which may be of value for sequence analysis. Several of the common endonucleolytic DNases have recently been reinvestigated to characterize more carefully their preferred cleavage sites. It now seems apparent that the specificity of these enzymes is influenced by a host of factors, including the concentration of activating cations, the degree of digestion,[150] etc.

Hog spleen acid DNase[151,152] seems to prefer certain linkages when salmon sperm DNA is digested to an average length of about ten residues, but this preference is not marked. Purines are found at the 3′ phosphate terminus of oligonucleotides from DNA digests 70% of the time, and the linkage cleaved least often is that between Cp and TpC in CpTpC.

The cleavage specificity of acid DNase from the hepatopancreas of the snail *Helix aspersa* (Müll)[153,154] and *E. coli* endonuclease I[154,155] have also been investigated. Oligonucleotides isolated from snail DNase digestions have greater than 90% A or T (almost 80% A) at their 3′ termini. The 5′ termini show a slight perference for G and C, possibly a statistical result of the 3′ perference. Oligonucleotides isolated from *E. coli* endonuclease I digestions show a preference for T, A, and C at the 3′ ends but no significant 5′ end preference. Pancreatic DNase I in the presence of Mg^{++} cleaves exclusively at the linkage between A and pT in a biosynthetic d(A-T) polymer[156] or crab d(A-T) polymer,[157] but no significant specificity is found in the hydrolysis of DNA.[123] The specificity of this enzyme is somewhat altered when Mn^{++} is used in place of Mg^{++} in the hydrolysis of DNA.[158]

Micrococcal nuclease cleaves DNA to produce oligonucleotides with predominantly T or A at the 5′ end,[159] and among many dinucleotides tested[160] cleaves TpT the fastest.

E. coli exonuclease III catalyzes the sequential removal of mononucleotides from the 3′ termini of duplex DNA.[161] At 5°C and in 67 m*M* Tris buffer, 70 m*M* NaCl, exonuclease III catalyzes the synchronous hydrolysis of approximately 6 nucleotides from each 3′ terminus at equimolar amounts of enzyme per DNA terminus.[162] At higher levels of exonuclease III, e.g., enzyme per DNA terminus equals 2, approximately 12 nucleotides can be removed from each 3′ terminus. Thus, it is possible to generate from any linear duplex DNA (Figure 12 molecule A) a molecule with 5′ terminated single-stranded ends (molecule B) which resemble the native termini of λ DNA

(Figure 2). With this DNA molecule, sequence analysis of the single-stranded ends can be accomplished by repair synthesis, followed by analysis of the complementary radioactive segment (see Section A-1). The sequence of this segment (dotted line in molecule C, Figure 12) is the same as segment I which was removed by exonuclease III.

iii. Single-strand Specific Deoxyribonucleases

Neurospora crassa endonuclease[155,163] is highly specific for single-stranded DNA, with a distinct preference for deoxyguanosine residues.

Aspergillus nuclease S_1 also has a high specificity for single-stranded DNA.[164] Since this enzyme is specific and relatively stable, it has been used extensively for degrading single-stranded DNA in the presence of double-stranded DNA.[165,166]

Mung bean nuclease I shows a strong preference for single-stranded DNA. However, at high nuclease concentration, it can cleave at a known A, T rich region of double-stranded DNA, such as that near the center of the λ DNA, presumably after local denaturation of the DNA.[167,168]

Micrococcus luteus UV-exonuclease degrades single-stranded DNA without affecting double-stranded DNA, unless the latter has been both irradiated and treated with the UV-endonuclease.[169]

These single-strand specific nucleases (S.S.S. nucleases) may be useful for DNA sequence analysis in the following way. The nuclease can be used to remove the single-stranded tail in molecule B (Figure 12) to produce a completely double-stranded DNA molecule D, which is shorter than the original molecule by the length equivalent to segment I. By removing this segment, its sequence need not be analyzed again when the sequence of segment II is to be determined. The sequence of segment II can be determined after exonuclease III digestion of molecule D to give E, followed by repair synthesis. In order to serve as a useful reagent for sequence analysis in the above scheme, the S.S.S. nuclease must fulfill at least two important criteria: First, the enzyme should be completely specific for single-stranded DNA under the experimental conditions. Second, the extent of hydrolysis of single-stranded DNA segments on different DNA molecules should be the same.

Using λ DNA as the test system, Hamilton and Wu[170] found that while the single-stranded cohesive ends can be almost completely removed by *Neurospora* endonuclease, some nicks, gaps, or double-stranded breaks are also created.[170] Therefore, this enzyme may not be useful for sequence analysis. On the other hand, *Aspergillus* nuclease S_1 was found to cleave off the single-stranded cohesive ends of λ DNA completely, without creating other nicks or breaks.[171] Thus, this enzyme is a good candidate for serving as the S.S.S. nuclease required in Figure 12. Although mung bean nuclease at 30°C also cleaved off the cohesive ends of λ DNA completely, some internal nicks or breaks were found.[171] In contrast, *M. luteus* UV-exonuclease cleaved off only part of the cohesive ends of λ DNA, leaving a 4-nucleotide-long, single-stranded tail.[171] This enzyme may also be useful, since the single-stranded tail may serve as overlap for aligning sequences in the analysis of segment II (see molecules D and E in Figure 12).

2. Sequencing by Mobility

This technique makes use of characteristic mobility differences on some separation systems between partial sequential degradation products of an oligonucleotide. The difference in mobility between an oligonucleotide and a homologous oligonucleotide one base longer can be used to identify the base by which the two oligonucleotides differ.

The first studies of the relationship between the nucleotide composition of an oligonucleotide and its characteristic mobility on a separation system were carried out by Sanger et al.[172,173] on two-dimensional (2-D) electrophoresis, using oligoribonucleotides. Later, these studies were extended to oligodeoxynucleotides.[6,123] In the standard 2-D electrophoresis system, cellulose acetate was used at pH 3.5 in the first dimension, and DEAE-cellulose paper at pH 1.9 or 3.5 was used in the second dimension. Comparative mobilities of homologous series of oligonucleotides in these systems often allow partial sequence determination of short oligonucleotides. It was observed by Sanger, Brownlee, and Barrell[172] that for any oligoribonucleotide having a 3′ phosphate, certain characteristic mobility shifts were caused by the addition of a single nucleotide, as shown:

$$(XpYpZp) + Np \rightarrow Np(XpYpZp) \text{ or } (XpYpZp)Np$$

The characteristic mobility shift caused by addi-

tion of Np was found to be approximately the same for the formation of any of the possible isomeric products. In general, the mobility shift was characteristic of Np and relatively independent of the identities of X, Y, and Z. However, small differences in the mobilities between sequence isomers of oligonucleotides having identical base composition were observed.

Electrophoretic separation in the first dimension on cellulose acetate at pH 3.5 often showed interpretable mobility shifts. When the added nucleotide was Cp, the lengthened oligonucleotide was observed to have a slower mobility than the original oligonucleotide. Addition of Ap produced relatively little change in mobility. Addition of Gp or Up produced an increase in mobility.[172]

Such behavior is, in fact, expected for electrophoresis on a neutral inert gel support, such as cellulose acetate, where mobility should be approximately proportional to the charge per mass of any oligonucleotide.[174,175] Addition of any nucleotide represents approximately the same addition of mass. At pH 3.5, however, the net charges of the 4 nucleotides differ widely because of the difference in the pK_a values for the amino groups on the bases. Addition of Up represents addition of one negative phosphate charge, since uracil is not protonated. Additions of Gp, Ap, and Cp, respectively, represent additions of progressively less net charges, since their bases are partially protonated. Since cytidine at pH 3.5 is 90% protonated, Cp has a net charge of only -0.1. When Cp is added, the charge per mass ratio of the oligonucleotide will be reduced, resulting in a decrease in mobility.

Somewhat less predictable behavior was observed for dephosphorylated oligonucleotides without a 3′ or 5′ phosphate. In this case, additions of any mononucleotide, except Cp, generally caused mobility increases.

In the second dimension, electrophoresis on DEAE-cellulose paper, Sanger, Brownlee, and Barrell[172] observed that the oligonucleotides exhibited a combination of electrophoretic and ion exchange behaviors. In the case of phosphorylated oligonucleotides, addition of single nucleotides to an oligonucleotide invariably caused a decrease in mobility. Successive additions of a single nucleotide to an oligonucleotide resulted in a rapid, approximately logarithmic decay of mobility until the product oligonucleotides had no detectable mobility. Because of this effect, the addition of a mononucleotide to a small oligonucleotide would cause a much greater change in mobility than the same addition to a large oligonucleotide. In order to compensate for this observation, Sanger et al.[172] proposed the use of the "m-value." The m-value is defined as [(mobility Xp . . . Zp) - (mobility Xp . . . ZpNp)]/(mobility Xp . . . ZpNp). This value is characteristic for each of the four mononucleotides represented by Np. The m-value is largest for Up and Gp, about 0.7 to 2.5, smaller for Ap, about 0.4 to 0.6, and smallest for Cp, about 0.05 to 0.25. The mobility shift values for the addition of Up or Gp have significant overlap in both the separation systems which have been described. This, however, presented no problem if the standard procedures for RNA sequence analysis were used to obtain oligonucleotides for sequencing by mobility. For a given fractionation of a T_1 RNase or pancreatic RNase digest, becuase of the specificity of the cleavages which produce the oligoribonucleotides, either Gp or Up, but never both, will be present in the internal positions of RNase fragments, and thus, differences in only three ribonucleotides have to be distinguished.

On the other hand, comparative mobilities of dephosphorylated oligonucleotides, particularly small ones on DEAE-cellulose electrophoresis, were sometimes anomalous.

Separation characteristics on the 2-D electrophoresis system were expected to be similar for oligodeoxynucleotides becuase of the similar structures and charges of the deoxy- and ribomononucleotides. Indeed, Murray[123] found that oligodeoxynucleotides behaved almost identically to oligoribonucleotides on the standard 2-D electrophoresis system. Also, 5′ phosphorylated oligonucleotides behaved similarly to 3′ phosphorylated oligonucleotides. He also introduced 2 new electrophoretic systems for fractionation of oligonucleotides: DEAE-cellulose electrophoresis at pH 9.7, and electrophoresis on AE-cellulose paper at pH 3.5.

Fractionation patterns of oligodeoxynucleotides were rather similar on DEAE-cellulose at pH 1.9 and 9.7. Differences in the nucleotide charges at the different pH values allowed fractionation in a 2-D system, employing DEAE-cellulose paper electrophoresis in both dimensions.

AE-cellulose electrophoresis at pH 3.5 also gave a fractionation pattern similar to that from DEAE-cellulose paper electrophoresis at pH 1.9. The significant advantage of using this system is in

its capability of distinguishing addition of dpT from addition of dpG. Additions of dpT generally gave m-values less than 2.0, while additions of dpG gave m-values greater than 2.0.

Szekely and Sanger[138] suggested that oligonucleotides could be partially degraded from the 5′ end with spleen phosphodiesterase and the products labeled with polynucleotide kinase. Adapting this idea, Murray[176] introduced a method of sequencing short oligonucleotides completely on the basis of mobility. Polynucleotide kinase was used to label the 5′ end of DNA or the oligonucleotide. Then, the DNA or the oligonucleotide was partially degraded with pancreatic DNase to produce a series of labeled oligonucleotides of decreasing length, all with the same labeled 5′ end. These oligonucleotides were separated by DEAE-cellulose electrophoresis at pH 1.9 and AE-cellulose electrophoresis at pH 3.5. For example, the oligodeoxynucleotide $\overset{*}{p}$XpYpZ, when partially digested, should yield the labeled products $\overset{*}{p}$XpYpZ, $\overset{*}{p}$XpY, and $\overset{*}{p}$X. The identity of Y was determined by comparing the mobilities of $\overset{*}{p}$X and $\overset{*}{p}$XpY, and the identity of Z was determined by comparing the mobilities of $\overset{*}{p}$XpY and $\overset{*}{p}$XpYpZ. By mobility shifts in these 2 systems, sequences from the 5′ termini of λ DNA were determined. Since both 5′ ends of λ DNA would be labeled by polynucleotide kinase, 5′ end analysis of each oligonucleotide was necessary to differentiate the 2 series in the digest. The sequence agreed with sequences obtained by DNA polymerase repair techniques.[5,6]

Unfortunately, both of the electrophoretic separations used for these sequence determinations are of limited value for sequence analysis of longer oligonucleotides. In the DEAE-cellulose electrophoresis system, because of the rapid decay of mobility with size, oligonucleotides longer than 6 to 12 nucleotides, depending on the composition, tend to cluster near the origin, preventing accurate differential mobility measurements. The AE-cellulose electrophoresis system has even poorer resolving capabilities. Oligonucleotides longer than 5 to 8 nucleotides, depending on composition, often cluster at the origin.[2] Recent efforts to improve the resolving power of the DEAE-cellulose and AE-cellulose electrophoresis systems[177] by modifications of pH and ionic strength of buffer have had little success. Wu and collaborators[17,19] have also investigated the possibility of 2-D DEAE-cellulose electrophoresis systems for direct sequence analysis of oligodeoxynucleotides by mapping and arrived at similar conclusions: These 2-D systems have rather limited use for direct sequence analysis, and uncertainties exist in some analyses.

In order to sequence longer oligonucleotides, new separation systems, with better fractionation capabilities as well as sequence differentiation capabilities, are required.

PEI-cellulose thin layer chromatography (cellulose impregnated with polyethyleneimine)[178] has been applied to the fractionation of oligonucleotides.[120,179] PEI is another ion exchange material with properties similar to DEAE- and AE-cellulose.

Using a 2-D system with cellulose acetate electrophoresis in the first dimension, and PEI-cellulose chromatography in pyridine formate buffer at pH 3.7 in the second dimension, Murray and Murray[14] obtained considerable sequence information from the coliphage cohesive end regions of 5 lambdoid phages (λ, φ80, 82, 21, and 424), as well as the nonlambdoid phages 186, P2, and 299. However, this could not be used as a general method because neither dimension is capable of distinguishing the addition of dpG from the addition of dpT. They were able to resolve G, T ambiguities because, as Morrison and Murray[177] point out, "the coliphage sequences were special cases, since each chromatogram contained two families of complementary sequences, providing internal comparisons of the effects of adding nucleotides."

This mobility method is faster than repair synthesis (see Section A-1) for the sequence analysis of cohesive ends of DNA such as λ or P2 DNA. However, there are two disadvantages in this mobility method. First, it cannot determine the length of the single-stranded cohesive end. Second, it cannot determine a sequence as long as that based on repair synthesis followed by sequence analysis.

It was also pointed out by Morrison and Murray[177] and Jay et al.[19] that the m-values for the four nucleotides vary considerably with the R_f values of the oligonucleotides being compared. This problem, as well as the problem of distinguishing G from T, limits the usefulness of the PEI-cellulose chromatography system using pyridine formate at pH 3.5 for purposes of sequence analysis by mobility shifts.

Jay et al.[19] have shown that in order to move larger oligonucleotides, a higher concentration of pyridine formate buffer must be used. Under such conditions, the smaller oligonucleotides, particu-

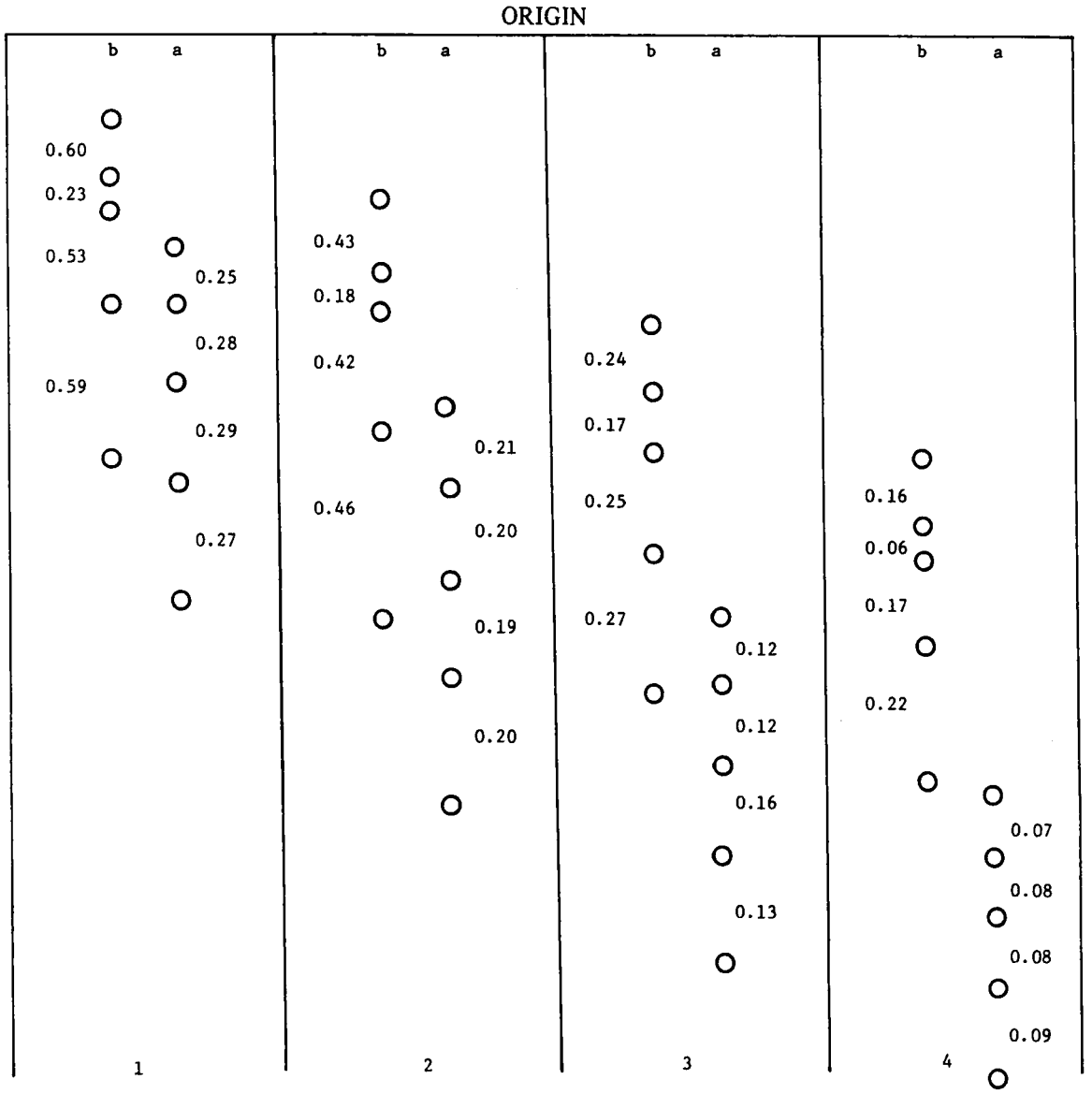


FIGURE 13. One-dimensional chromatography and the m-values for the partial venom phosphodiesterase digest of (a) d(^{32}pT-T-T-T-T), and (b) d(^{32}pG-G-G-C-G) on PEI-cellulose thin-layer plates in 7 *M* urea and (1) 0.6 *M* LiCl, (2) 0.7 *M* LiCl, (3) 0.8 *M* LiCl, and (4) 0.9 *M* LiCl.

larly those that are rich in C and A, either move with the solvent front or are compressed near the solvent front. They found that solvents of LiCl in 7 *M* urea can overcome most of the problems mentioned above. Figure 13 shows the mobilities of the partial venom phosphodiesterase digestion of 5′ ^{32}P-d(T-T-T-T-T) and 5′ ^{32}P-d(G-G-G-C-G) on a one-dimensional PEI-system in various concentrations of LiCl containing 7 *M* urea. It was observed that the m-values get smaller with increasing LiCl concentrations, but they remain rather constant for the addition of any particular nucleotide, regardless of the R_f values of the spots being compared. The m-value for the gain or loss of a dpG residue is twice that for a dpT or dpC. Thus, this system can be used for differentiating between dpG and dpT addition.

Since chromatography on PEI-cellulose in LiCl-7 *M* urea can distinguish dpG and dpT, and electrophoresis on cellulose acetate at pH 3.5 can distinguish dpC, dpA, and dpT or dpG, a 2-D system employing these 2 methods of separation

should provide unambiguous sequence information from mobility shifts of the partial degradation products of an oligonucleotide. Figure 14 shows the 2-D separation of a partial venom phosphodiesterase (3′ → 5′ exonuclease) digestion of a 5′ ^{32}P-labeled synthetic tetradecamer d(p*A-G-T-C-C-A-T-C-A-C-T-T-A-A). From the mobilities of the degradation products, it can be seen that the m-values for the gain of a dpT, dpC, or dpA are much smaller than the m-value for the gain of a dpG. Since dpT, dpC, and dpA have characteristic mobility shifts in the first dimension, addition of any of the four deoxynucleotides can be distinguished. Thus, this 2-D fractionation system can be very useful for sequence analysis by the mobility shift method.[19]

Another system which has shown unusually good fractionation capabilities is the so-called "homochromatography" of Brownlee and Sanger.[180] This is a thin layer chromatography technique on DEAE-cellulose paper or plate, using a solvent containing an RNA digest and 7 *M* urea in water. During development of the plate, the unlabeled oligonucleotides in the solvent displace the labeled nucleotides being investigated to different positions on the plate, approximately according to their size. This system showed early success in fractionation of large oligonucleotides in the 5- to 50-nucleotide size range. It was possible to control the mobility of the oligonucleotides by varying the extent of alkaline hydrolysis of the RNA or its concentration in the solvent. A long hydrolysis time produces short RNA segments in the solvent which can displace only shorter labeled oligonucleotides. A shorter hydrolysis time produces longer oligonucleotides which can displace short and long labeled oligonucleotides.

This system was coupled with electrophoresis on cellulose acetate to provide a 2-D system for fractionation of oligonucleotides. Sanger et al.[46] and other investigators in his laboratory used it for fractionating RNA and DNA digests. Ling[129] used it for fractionating partial digests of pyrimidine tracts obtained from fd DNA, and used mobility shifts on the cellulose acetate dimension to distinguish dpT and dpC. It soon became apparent that besides being a good fractionation system, the homochromatography separation was able to distinguish additions of pyrimidines from additions of purines. Since this allows distinction of dpG from dpT, the 2-D system with cellulose acetate in the first dimension should allow unambiguous sequencing of oligonucleotides solely by mobility. A number of reports have appeared which describe investigations of the sequence differentiating capabilities of homochromatography.[17,19,39,130,181]

Using dephosphorylated oligonucleotides derived from digests of R17 RNA, Rensing and Schoenmakers[181] were able to obtain some direct sequence information from relative mobilities of oligonucleotides on 2-D homochromatography. In some cases, however, resolution in the homochromatography dimension was insufficient to distinguish pA from pC.

We have developed a means of producing homochromatography solvent by reacting RNA with just the correct stoichiometric amount of KOH to produce the desired level of digestion.[19] In this way, the homochromatography solvents could be prepared in a very reproducible fashion. This method for producing homochromatography solvents also seems to produce better resolution and sequence differentiating characteristics in the homochromatography separations. Using a number of phosphorylated oligonucleotides, we were able to unambiguously differentiate the additions of purines from additions of pyrimidines in the homochromatography dimension. An example of a 2-D homochromatography separation of partial degradation products of an oligonucleotide is shown in Figure 15A. The synthetic oligonucleotide d(A-G-T-C-C-A-T-C-A-C-T-T-A-A) was 5′ terminal labeled using T_4 polynucleotide kinase and [γ-^{32}P]ATP. It was then partially digested from the 3′ end with venom phosphodiesterase. The d-value,[19] or distance between 2 neighboring spots, for addition of purine is usually about 1.5 times the d-value for addition of pyrimidine. Since the separation does experience some compression near the origin, d-values among neighboring spots are always compared.

It is also possible to label the 3′ end of a single-stranded oligonucleotide and sequentially degrade it from the 5′ end with spleen phosphodiesterase, or to use endonucleolytic cleavage with pancreatic DNase. In this way, a series of homologous oligonucleotides, all with the same labeled 3′ end, could be used to confirm data from separation of 5′ terminal labeled oligonucleotides. Deoxynucleotidyl terminal transferase[182] can be used to add a single ^{32}P-ribonucleotide to the 3′ terminus of oligonucleotides greater in size than trinucleotide.[183] The 3′ terminal ribonucleoside can be removed by treatment with $NaIO_4$,[39]

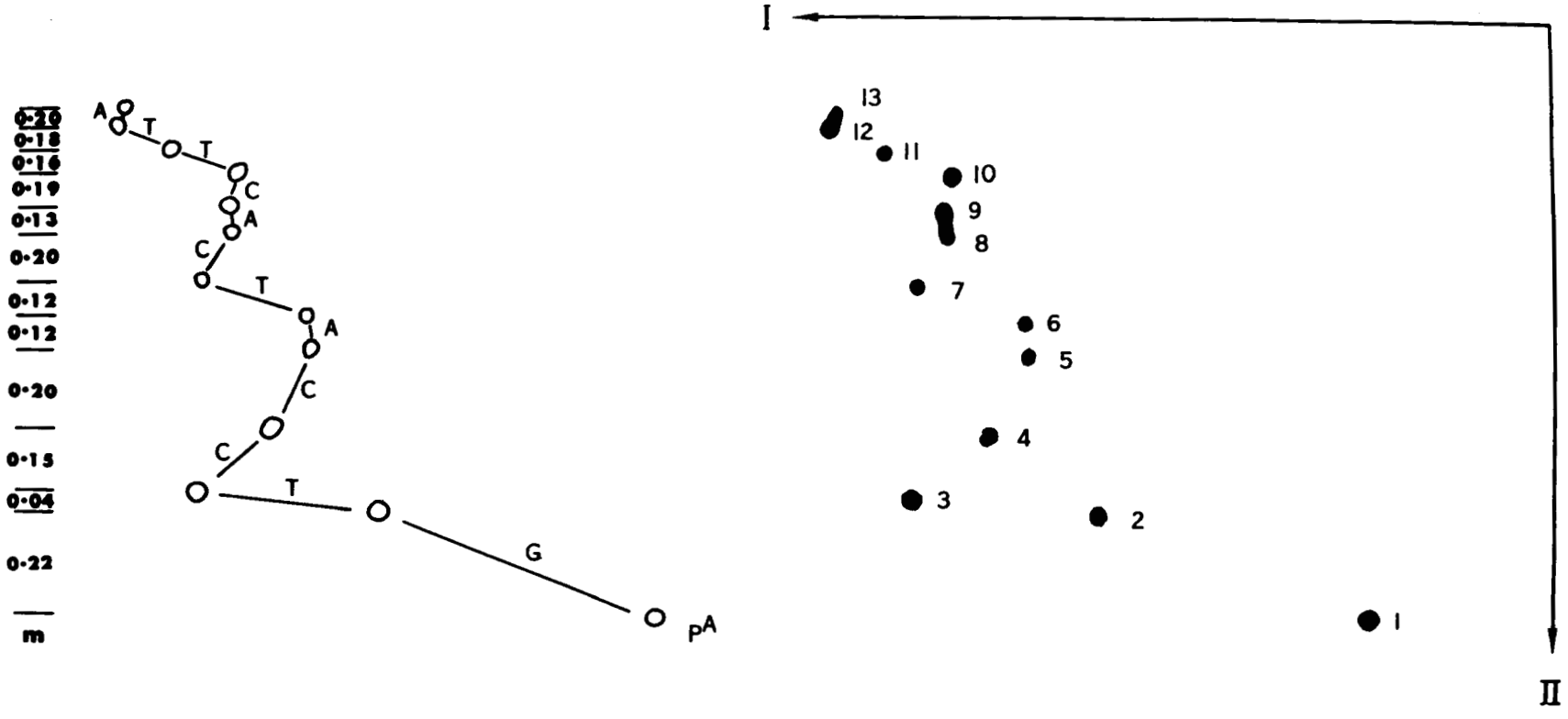


FIGURE 14. 2-D PEI-cellulose fingerprint of the partial venom phosphodiesterase digest of d(^{32}pA-G-T-C-C-A-T-C-A-C-T-T-A-A). The first dimension separation is electrophoresis on cellulose acetate gel at pH 3.5. The second dimension separation is on PEI-cellulose thin-layer plate in a solvent of 0.85 *M* LiCl – 7 *M* urea.

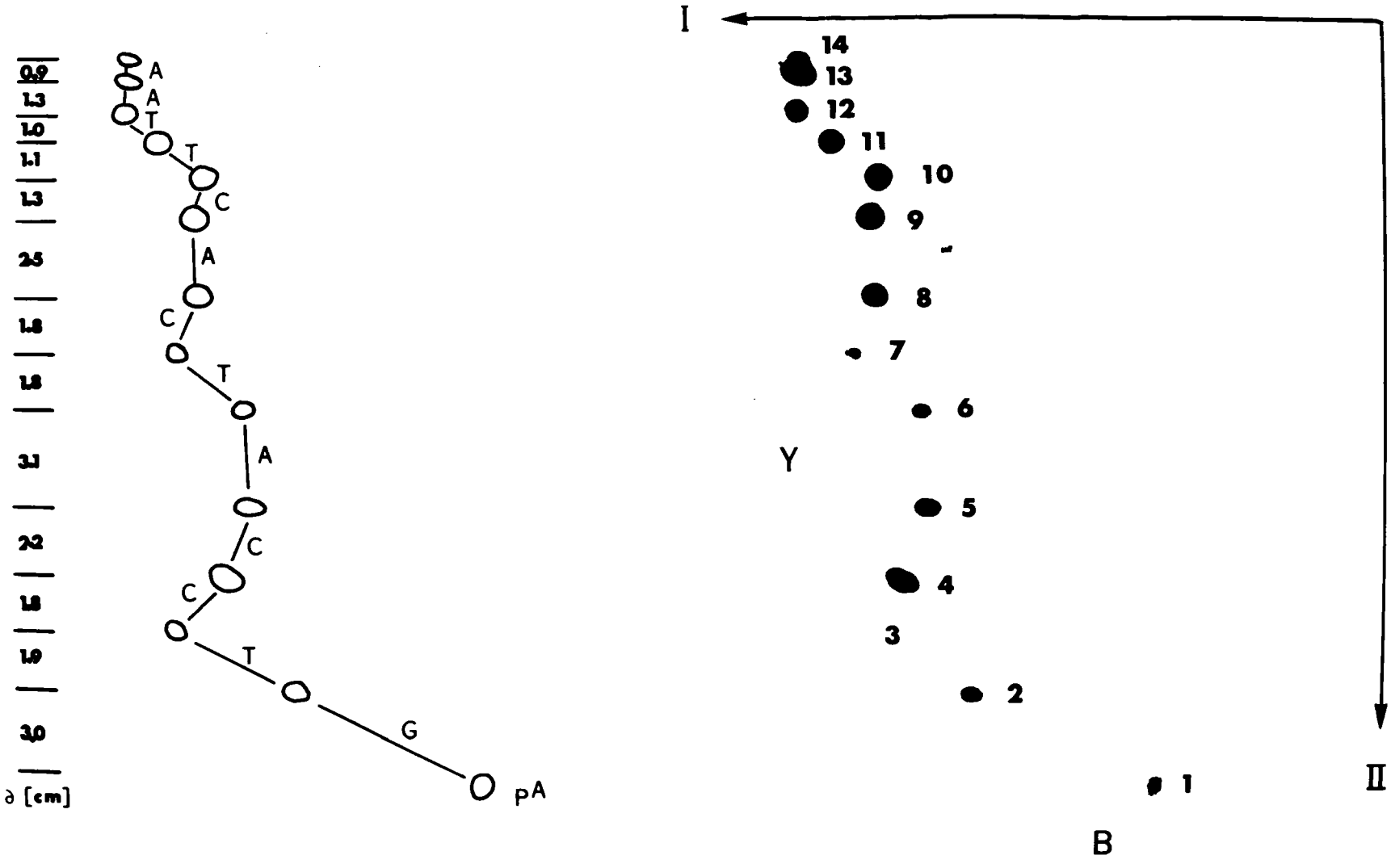


FIGURE 15A. 2-D homochromatography separations of partial digestions of an oligonucleotide. The first dimension separation is electrophoresis on cellulose acetate gel at pH 3.5. The second dimension separation is on DEAE-cellulose thin-layer plate, using the solvent Homomix V.[19] (A) Separation of the partial venom phosphodiesterase digest of d(^{32}pA-G-T-C-C-A-T-C-A-C-T-T-A-A). (B) Separation of the partial spleen phosphodiesterase digest of d(A-G-T-C-C-A-T-C-A-C-T-T-A-A^{32}p).

leaving a 3′ ^{32}P on the original oligonucleotide. Partial spleen phosphodiesterase digestion could then be used to produce the homologous series of labeled oligonucleotides for 2-D homochromatography separation (Figure 15B). Sequence information deduced from this map fully confirms that obtained with the 5′ labeled oligonucleotides shown in Figure 15A.

Similar techniques were used to obtain terminal sequences from longer DNA molecules. One 3′ terminus of λ DNA was specifically labeled using *E. coli* DNA polymerase I and appropriate [α-^{32}P]dNTP.[5,17] The terminally labeled DNA was then degraded to oligonucleotides, which were separated on 2-D homochromatography. With this method, a 2-D homochromatogram was produced from each 3′ terminus of λ DNA. Ten nucleotides from the double-stranded region adjacent to the left-hand 3′ terminus and 5 nucleotides from the right-hand 3′ terminus of bacteriophage φ80 DNA have also been sequenced.[19]

Many other DNA segments have now been sequenced using this technique.[51,53,60,130,131] Although 2-D homochromatography provides a promising means of sequence determination by inspection, sometimes the mobility shifts do not follow the simple rules which have been described. For example, addition of pA to pC would be expected to cause an increase in mobility, because at pH 3.5, the charge per mass of pCpA is much greater than the charge per mass of pC. Addition of pG to pC should also cause an increase in mobility probably much larger than that caused by addition of pA. However, if the nucleotides involved were unknown, the distinction between addition of pG and pA might be difficult. In order to resolve such problems, Bambara, Jay, and Wu[184] have proposed the use of an empirical formula to accurately predict mobility shifts caused by nucleotide additions to oligonucleotides on cellulose acetate electrophoresis.

Electrophoretic mobility (U), or movement in a standard field strength, is defined as

$$U = \frac{q}{f}$$

where q is the charge on the particle and f is the frictional drag on the particle.[185,186] In the case of polymers,

$$f = C(M)^a$$

where "a" is relatively constant for different sizes of any type of polymer, usually between 0.4 and 1.0, depending on the polymer used.[187]

In the case of deoxyoligonucleotides, the empirical formula

$$a = b(1 + K \ln N)$$

where b and K are constants, was found to adequately describe the frictional drag. For convenience, N, the number of nucleotides, could be substituted for M, with an appropriate change of the value of C. Mobility is expressed with respect to the mobility of dpT, which has a molecular charge of -1.0 at pH 3.5. If the mobility of dpT is defined as -1.0, then C = 1 and the mobility of any oligonucleotide with respect to dpT (U_T) is

$$U_T = \frac{q}{(N)^{b(1 + K \ln N)}}$$

From the values of pK_a for the dissociable groups of the four mononucleotides, dpT, dpG, dpA, and dpT, one can calculate their net charges at any pH. The calculated q at pH 3.5 for dpT = -1.0, dpG = -0.72, dpA = -0.41, and dpC = -0.08. The effective charge of each nucleotide will be slightly different from these values because the pK_a's for the nucleotides were determined in a different buffer, and also because of differential shielding of the charged species by ionic components of the electrophoretic buffer. Since the value of q for dpT is -1.0, the effective charges of the other nucleotides were determined by comparing their mobilities on cellulose acetate electrophoresis with the mobility of dpT. In this way, q for dpG = -0.85, dpA = -0.30 and dpC = -0.15.

For cellogel (cellulose acetate stored in methanol-water), the values of b and K were determined by fitting the equation to a large number of oligonucleotides. The best values seemed to be b = 0.50 and K = 0.165. With these values and the use of the mobility equation, the mobility shift for the addition of dpG can be clearly distinguished from that for addition of dpA. The complete sequence of an oligonucleotide such as that shown in Figure 15 can usually be determined by mobility shift alone.[184]

It should be emphasized that for successful application of this formula for accurate prediction of mobility shifts, the conditions for the electro-

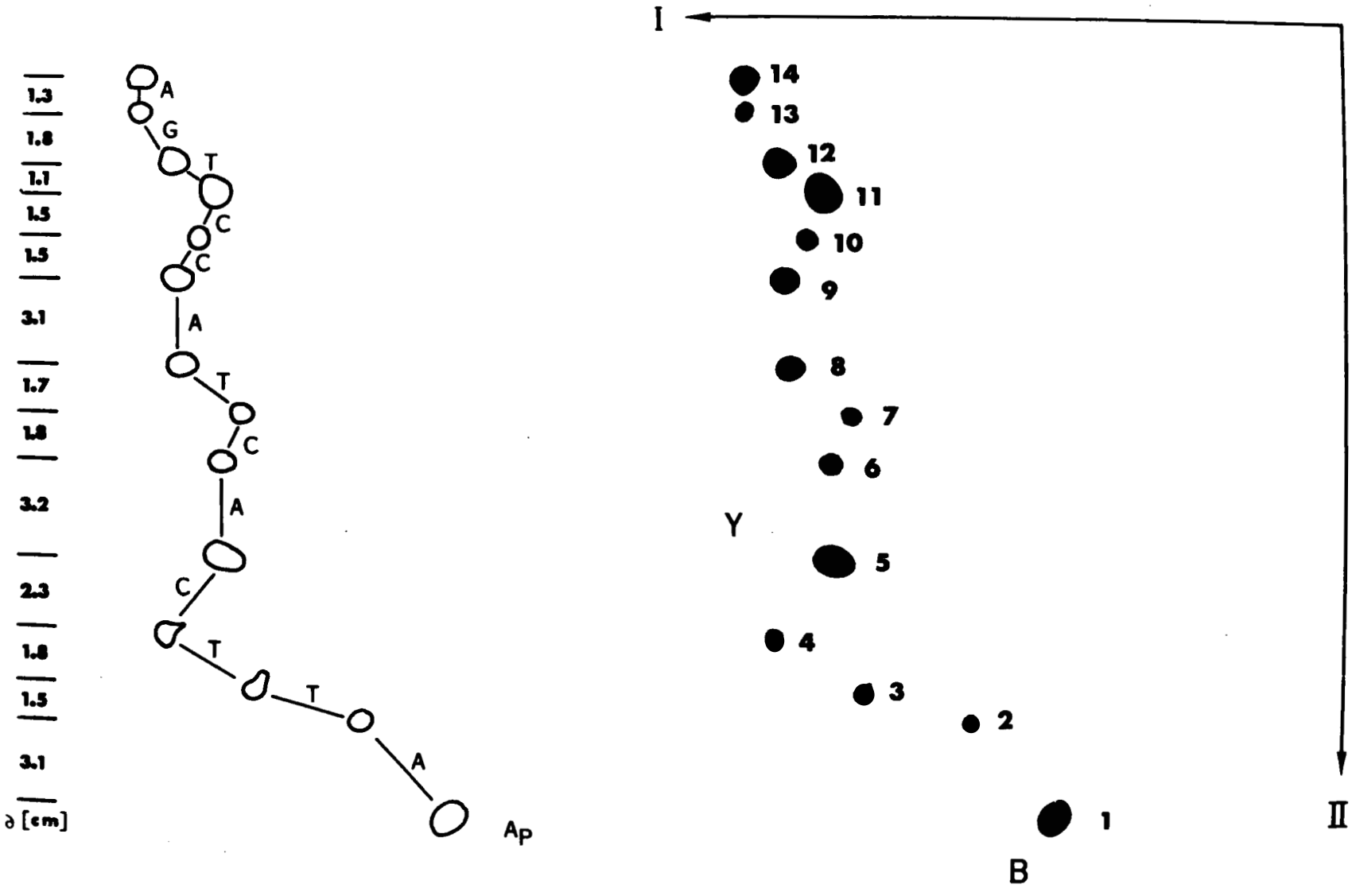


FIGURE 15B.

phoresis run should be kept constant and well controlled. For example, the pH of the buffer should be 3.5 ± 0.1, and the temperature of the electrophoretic tank should be relatively constant and should not exceed 25°C.

3. Sequencing DNA through its RNA Transcript

It is often advantageous to determine the sequence of medium length DNA molecules (e.g., 20 to 200 nucleotides long) from RNA produced by in vitro transcription. Since this complementary RNA can be made many times more highly radioactive than the DNA labeled in vivo, sequence analysis can be carried out with relatively small amounts of starting DNA. Another advantage is that all of the well-established RNA sequencing techniques can be applied to the analysis of complementary RNA. Also, the in vitro labeled RNA can be synthesized with a combination of different labeled and unlabeled ribonucleoside triphosphates to allow nearest neighbor analysis.

Potential problems with sequencing DNA through its RNA transcript include the possibility that errors may be made during transcription, and that multiple starts and random termination of transcription may occur, which may give an RNA transcript shorter than the DNA template. Furthermore, certain regions of DNA not transcribed in vivo may also be difficult to transcribe in vitro.

With the exception of the cases where a strong promoter site is present in the DNA segment, multiple initiation of RNA synthesis does seem to occur. Multiple starting and stopping points produce irregularly sized RNA segments which are difficult to sequence. Nonspecific or multiple initiation, however, can be minimized by using specific primers to initiate RNA synthesis. Even with this improvement, it is likely that some parts of the DNA are transcribed more frequently than others. Thus, meaningful quantitation of the RNA products (especially T_1 or pancreatic RNase digestion products) may be difficult or impossible unless the RNA transcripts are first fractionated or purified. The problem of missing a part of the DNA sequence because transcription may not start from the first nucleotide at the 3′ end of the DNA template might be remedied by determination of the sequence of the RNA transcript of both strands of DNA. Another solution is to do direct DNA sequence analysis from the termini. The DNA would be partially digested after labeling the 5′ or 3′ ends, and the products fractionated on a two-dimensional separation system. DNA sequences of 10 to 14 nucleotides from each end can be determined by the mobility shift method.[19] This DNA sequence should be long enough to overlap with the starting sequence of the RNA transcript.

For DNA molecules shorter than 20 nucleotides, direct sequence analysis by mobility shifts from both termini is a preferred method. For double-stranded molecules, if 12 nucleotides can be sequenced from the 5′ ends of each strand, the complete sequence of 20 nucleotides can be obtained with an overlap of 4 nucleotides. For DNA molecules much longer than 200 nucleotides, sequencing of the RNA transcript may become difficult and time-consuming, since much greater effort would be needed to obtain partial sequences and sufficient overlaps. In this case, sequencing of DNA after repair synthesis may be the method of choice. A long segment of DNA may be copied as ribosubstituted DNA through repair synthesis, and then specifically cleaved to shorter segments by restriction enzymes before each short segment is analyzed separately. In this instance, the original DNA is entirely copied, and therefore the product is most suitable for complete sequence analysis. If the long DNA molecule is first cleaved by restriction enzymes to three or four segments, and then each segment is transcribed into RNA sequence, uncertainties on both ends of each RNA segment are likely to be a problem in the complete sequence analysis of the original long DNA molecule.

Significant advances in DNA sequencing through in vitro transcription have been made in the last two years.

a. Use of Natural Promoter Sites for Specific Initiation of RNA Synthesis

If the segment of DNA to be sequenced includes a natural promoter site or a preferred RNA polymerase binding site, then rather specific initiation of RNA synthesis can be obtained. Several workers have taken advantage of this fact and have made impressive progress in the determination of long sequences starting from the 5′ end of a number of mRNAs synthesized in vitro from phage λ DNA.[188,189,190] These sequences, starting from different specific promoter sites, represent mRNAs which correspond to part of gene *N* and several other genes in λ DNA. In some of these

studies, the RNA synthesis was made more synchronous[189] by preincubation of the template DNA with RNA polymerase and nucleoside triphosphates at 20 to 25°C. The reaction was then started by the addition of $MgCl_2$, with or without rifampin. RNA synthesis was stopped at different time periods to allow production of RNA chains of different lengths in order to simplify the task of ordering RNA fragments.

Weissman and collaborators[54,55] have found a single preferred initiation site (strong RNA polymerase binding site) on intact SV40 DNA. This site was used to produce an in vitro transcript of the segment of SV40 DNA common to the nondefective $Ad2^+$ ND_1 and $Ad2^+$ ND_3 hybrid viruses. This radioactive RNA molecule was purified by annealing to the adeno-SV40 hybrid DNA (e.g., $Ad2^+$ ND_1). Only a small portion (about 5%) of the labeled SV40 RNA hybridized with the SV40 sequence in this hybrid viral DNA. A 180-nucleotide-long RNA (early strand transcript) was isolated, and its sequence was determined starting from the 5′ end. The RNA transcript from the late strand was obtained from fragment G_1 produced by digestion with the *Hemophilus influenzae* restriction endonuclease.[71] This transcript has also been sequenced.[57,58] Since the late strand sequence is complementary to that of the early strand transcript, the fidelity must be high in this in vitro system for RNA synthesis. Therefore, sequencing DNA by using it as template for in vitro RNA synthesis appears to be a valid approach. The nucleotide sequence of the SV40 DNA transcript which precedes the preferred initiation site for RNA polymerase has also been reported from analysis of both the in vitro late strand transcript[191] and the early strand transcript.[192] These two 45-nucleotide long sequences are, again, exactly complementary to each other, as expected. These sequences include a 17-residue-long true palindrome sequence, G-U-U-A-A-C-A-A-C-A-A-C-A-A-U-U-G, which may be part of the binding site of *E. coli* RNA polymerase. However, the significance of a possible recognition sequence for *E. coli* RNA polymerase on SV40 DNA remains uncertain.

Maizels[103] has sequenced the first 63 bases of mRNA transcribed in vitro from the UV5 promoter mutant of the *E. coli* lactose operon. When 1,000 base-pair-long sonicated DNA fragments were used as template, transcription usually started with pppA-A-U-U . . . , although about 15% of the molecules started with a G to give pppG-A-A-U-U . . . . When transcription was initiated with GpA as primer, most RNA molecules started with the sequence G-A-A-U-U . . . . During such primed RNA synthesis, RNA polymerase appeared to pause at particular sites along the DNA template. This effect generated several discrete sizes of RNA, ranging from 7 to over 100 bases long, that provided overlaps useful for sequencing. The in vitro synthesized mRNA included the transcript for the lactose operator and extended into the β-galactosidase gene starting at position 39 from the 5′ end of this RNA.

The 5′ terminal sequence of the in vitro synthesized mRNA from the galactose operon of *Escherichia coli* has been determined.[193] In this 77-nucleotide-long sequence, residues 1 to 26 represent the nontranslated "leader" sequence of galactose mRNA (see Figure 10C). There is a possibility that part of this sequence may correspond to part of the galactose operator. Residues 27 to 77 correspond to 17 amino acids, from the amino terminal, of UDP galactose-4-epimerase, the protein specified by the promoter proximal structural gene of the operon.

Interestingly, the sequences A-G-G-A and U-U-U-G-A found in many ribosome binding site regions[105] were not present in the leader sequence. However, the sequence does include a purine rich stretch that has been found in several other ribosome binding site sequences.

A region of 2-fold symmetry in the complementary DNA is indicated from symmetry in residues 4 to 18. The sequences of the symmetrical areas are dissimilar to sequences displaying two-fold symmetry in the *lac* control regions.

A sequence centered around the initiation codon, G-A-A-U-U-A-U-G-A-G-A-G, differs from the corresponding sequence in the *lac* operator, G-A-A-U-U-*G*-U-G-A-G-*C*-G, by only two bases. The significance of these similar sequences is not known.

Double-stranded DNA (RF_I), isolated from cells infected with bacteriophage f1, has been used as template for in vitro RNA synthesis.[104] Although these RNA molecules probably did not have specific starting points, they were purified by binding to ribosomes to form initiation complexes. The unbound RNA was digested away with pancreatic RNase. Three ribosome-protected RNA fragments (23 to 31 nucleotides long) were isolated and sequenced. As shown in Figure 10b, site #2, a true palindrome, A-U-U-A-A-A-G-U-U-G-A-A-A-U-U-A sequence is found to be located

immediately adjacent and to the 3′ side of the protein initiator triplet A-U-G. Another common feature in all the sequences[104,105] given in Figure 10, as well as in the ribosome protected fragments from 7 RNA phages,[100-104] is the presence of at least 1 termination triplet (U-A-A) at the 5′ side of the initiation triplet (A-U-G).

b. Use of Primers for Specific Initiation of RNA Synthesis

If the segment of double-stranded DNA to be sequenced does not contain a natural promoter site, which is likely to be the case in most future sequence analysis problems, then a primer may be used to direct more specific initiation of in vitro transcription. The RNA molecules synthesized in this way are likely to have relatively uniform starting points. Primers as short as dinucleotides, such as ApU, were found to stimulate in vitro RNA synthesis and to provide more specific initiation points during transcription, especially when low concentrations of ribonucleoside triphosphates were used.[194,195] Apparently, the K_m of ribonucleoside triphosphates for RNA polymerase is much higher for RNA chain initiation than that for chain elongation. Thus, when the substrate concentration is lowered to 10 μM or less, the enzyme can elongate chains, but cannot initiate new chains.[196] For single-stranded DNA, dinucleotide primers are useful only if the DNA to be transcribed is relatively short; otherwise, multiple starts can occur.[3] For example, if ApU is used as primer, there will be 4 initiation sites, on the average, on a template 64 nucleotides long.

Kleppe and Khorana[197] used short synthetic DNA fragments of known sequence as templates to study the initiation and termination of the in vitro transcription process. When a 30-nucleotide-long double-stranded DNA was used as template, it was found that initiation of RNA synthesis occurred at multiple sites, beginning with incorporation of purine nucleotides, and that both strands were transcribed. Furthermore, approximately one half of the total RNA product was larger than the DNA template. Attempts to control the initiation of transcription by the use of a complementary ribo-heptanucleotide primer and by carrying out the transcription in the presence of rifampicin were only partially successful. The RNA products were still rather heterogenous. When a 29-nucleotide-long, single-stranded DNA was used as template by Terao, Dahlberg, and Khorana,[198] RNA synthesis again did not begin at the 3′ end of the template, and initiation began with both ATP and GTP. When a ribo-heptanucleotide primer was added, the *de novo* chain initiation was abolished and the initiation of RNA synthesis became more specific. However, termination of transcription occurred at more than one point prior to the end of the template chain. These studies point out some of the problems involved in the use of DNA sequence analysis through in vitro transcription.

Gilbert and Maxam[112] have isolated a double-stranded DNA fragment (lactose repressor binding site) which interacts specifically with *lac* repressor.[111] Transcription of this lactose operator fragment (about 27 base pairs long) with RNA polymerase produced a mixture of products from many points of initiation and termination.[112] Primers were then used to direct synthesis from specific starting points. GpU was found to prime RNA synthesis from both template strands, while UpA and A-U-C-C-G primed synthesis from only one strand, and G-C-A-A-U primed synthesis from only the other strand. The operator DNA sequence (see Figure 6e) was finally deduced from the sequences of the RNA molecules isolated from primed synthesis.

c. Use of Satellite DNA Templates for in vitro RNA Synthesis with RNA Polymerase

Satellite DNAs are DNAs of highly repetitive sequence which have been found as a significant percentage of the cellular DNA content of many higher eucaryotes. It is realtively easy to purify these DNA molecules from other cellular DNA using neutral or alkaline CsCl equilibrium centrifugation.

Sequences of the major repeating units of some of these satellite DNAs have been determined by analysis of depurination products, as discussed in Section B-a-i, and others have been determined by transcription into RNA and sequencing of the transcript. The transcription technique has the advantage that overlapping may be used to confirm sequence information. Gall, Cohen, and Atherton[236] used this technique to determine the repeating sequences of three distinct satellites from *Drosophila virilis*. These sequences are: Sat. I., 5′-(A-C-A-A-A-C-T)$_n$-3′, Sat. II 5′-(A-T-A-A-A-C-T)$_n$-3′, and Sat. III 5′-(A-C-A-A-A-T-T)$_n$-3′. Fry et al.[234] and Salser et al.[199] determined the repeating sequence 5′-(A-C-A-C-A-G-C-G-G-G)$_n$-3′ found in the HS-β satellite DNA from kangaroo rat

(*D. ordii*). Skinner et al.[235] using the transcription method found that 86 to 92% of the hermit crab satellite DNA consisted of the repeating sequences $5'$-(A-U-C-C)$_n$-$3'$ and $5'$-(U-A-G-G)$_n$-$3'$. The purpose of satellite DNA, the significance of these sequences, and the variations between species and within the same species are still matters of speculation.

d. Use of Single-stranded DNA Templates for in vitro RNA Synthesis with Reverse Transcriptase

It is often difficult or impossible to use in vivo labeling techniques to produce DNA or RNA from higher organisms which is of sufficiently high specific activity for sequence analysis. Instead, it is necessary to use an in vitro system to produce this high specific activity DNA or RNA. One system makes use of RNA-dependent DNA polymerase (reverse transcriptase) to transcribe mammalian RNA of low specific activity into complementary DNA (cDNA). The cDNA can be sequenced directly if its specific activity is sufficiently high; or, alternatively, it can serve as template for *E. coli* RNA polymerase to produce high specific activity complementary RNA (cRNA) for sequence analysis. Thus, Salser and associates[199,206] transcribed rabbit reticulocyte mRNA with reverse transcriptase in the presence of oligo (dT) as primer, to produce cDNA. The latter was then used as template for in vitro synthesis of highly labeled RNA for sequencing. Two 11-nucleotide-long sequences and several shorter sequences have been analyzed. One of them was found to correspond to a tetrapeptide, and several of them to tripeptides and dipeptides found in the α-chain of the rabbit hemoglobin molecule. Thus, the transcription fidelity of this system seems adequate for sequence analysis. One possible complication in this system is that poly A and poly U were also synthesized.[3,200] The presence of these compounds may interfere with sequence analysis of certain oligonucleotides.

Marotta et al.[201] have reported similar types of experiments involving transcription of human globin mRNA into cDNA by use of reverse transcriptase. The cDNA was then transcribed into ^{32}P-labeled cRNA by *E. coli* RNA polymerase. The size of the globin cRNA was found to be in the range of 4S to 6S, considerably smaller than the natural globin mRNA. The fingerprint pattern of a RNase T_1 digest of the globin cRNA was somewhat simpler than, but rather similar to, that of the natural 10S globin mRNA. Thus, the fidelity of the transcription processes appears to be high. Nucleotide sequence information has been obtained from about 50% of the intermediate sized oligonucleotides (8 to 14 nucleotides long). Many of these oligonucleotide sequences can be matched to unique amino acid sequences in the α- or β-globin chains. The other 30% do not match known amino acid sequences and may correspond to untranslated portions of the mRNA.

e. Deoxysubstitution

The success of the method of ribosubstitution in *E. coli* DNA polymerase repair reactions has raised the question of whether *E. coli* RNA polymerase can be used to incorporate a mixture of one deoxy- and three ribonucleotides into RNA. This method would allow the newly synthesized deoxynucleotide-substituted RNA to be cleaved by various RNases at sites which are more specific, and the fractionated products to be specifically degraded further by DNA enzymes.

The essence of the procedure for deoxysubstitution[202-204] is that the presence of Mn^{++} ion, which allows DNA polymerase to synthesize ribosubstituted DNA, will also allow RNA polymerase to synthesize deoxysubstituted RNA. Paddock et al.[205] initially used M13 DNA as template, and found that the incorporation of 1 deoxy- and 3 ribonucleotides occurred as rapidly as the incorporation of 4 ribonucleotides. Complete alkaline hydrolysis produced ribomononucleotides and deoxyoligonucleotides with $3'$ ribonucleotides termini, such as dA-dA-rGp. Tests of fidelity of deoxysubstitution were done by comparing two-dimensional fingerprints of in vitro synthesized RNA and deoxysubstituted RNA molecules which were cleaved at the same base residue. For example, RNA was synthesized from either Hs-β satellite (*Dipodomys ordii*) DNA or Hb complementary DNA. In one experiment, using [α-^{32}P] UTP label, normal RNA was produced. In another experiemnt, using the same label, rC in the RNA was completely substituted by dC. U_1 RNase was used to cleave the RNA specifically after G residues. Standard 2-D electrophoretic fingerprints of both samples show the same positions and intensities of spots, even for spots of >50% cytidine content. These data strongly suggest that the enzyme is making essentially no transcriptional errors during deoxysubstitution synthesis.

Since the electrophoretic fingerprints of Hb

cRNA are well-characterized,[206] deoxysubstituted oligonucleotides of known sequence could be eluted from the maps to test enzymatic cleavage characteristics. Paddock et al.[205] found that oligoribonucleotides containing dC were cleaved only at U residues by pancreatic RNase A, and not at both the C and U residues as with normal RNA. Similar oligoribonucleotides substituted with dT were cleaved only after C residues. If dG is substituted, cleavage with U_2 RNase produced products ending in a 3′ A. Oligoribonucleotides terminating in G are, of course, available from T_1 RNase digests. Thus, with the use of the available RNases and deoxysubstitution, the substituted RNA can be cleaved specifically at any one of the four ribonucleotides.

Van de Voorde et al.[207] independently carried out similar types of deoxysubstitution experiments using in vitro synthesis of RNA. They found that all four deoxynucleotides can individually substitute for the corresponding ribonucleotide in RNA synthesis, and thus the deoxysubstituted RNA can give more specific cleavages by RNase U_2, pancreatic RNase, etc., as mentioned earlier. For different deoxynucleotides, a different optimal level of Mn^{++} was found for maximum rate of RNA synthesis. Using heat-denatured calf thymus DNA or SV40 Hind-fragment DNA, they found that deoxysubstituted transcription was only 5 to 10% as rapid as normal RNA transcription. This is in contrast to the data of Paddock et al.,[205] in which deoxysubstituted transcription of M13 DNA was very efficient.

One general problem with the deoxysubstitution technique is that small contaminations of other ribonucleoside triphosphates must be removed carefully from each ribonucleoside triphosphate used in the experiments, since the RNA polymerase still greatly favors synthesis with ribotriphosphates over deoxytriphosphates. This problem was recognized and successfully solved by purifying all the nucleoside triphosphates.[205]

Therefore, deoxysubstitution appears to be efficient and accurate. It is likely to prove very useful for RNA sequence analysis.

f. Sequence Analysis of mRNA

The following two pieces of work on mRNA sequence analysis are included here because of their unusual interest, even though they are somewhat beyond the scope of this review article.

i. The Trp Operon

Bronson, Squires, and Yanofsky[102] used a combination of deletion mutants and hybridization techniques for RNA sequence analysis of the control regions adjacent to the sequence coding for the first polypeptide of the tryptophan operon in *E. coli*. Three deletion mutants were used in these studies, *trp* ΔED102, *trp* ΔED2, and *trp* ΔED24. *Trp* ΔED102 has the region from just before the *trp* E gene (the first polypeptide) to within the D gene (the second polypeptide) deleted. *Trp* ΔED24 has the region from just within beginning of the E gene to within the D gene delted. *Trp* ΔED24 has the region from just within the E gene to within the D gene deleted. *Trp* mRNA transcribed from the mutant *trp* operons was expected to contain a "leader" sequence containing control region information, possibly some initial amino acid coding sequences (depending on the deletion), and then amino acid coding sequences for genes beyond the E gene in the transcription sequence.

Transcription was allowed to occur so that ^{32}P-labeled mRNA was produced. Then, the labeled *trp* mRNA was isolated from other bacterial mRNA by hybridization with the transducing phage λφ80*trp*, which carries the sequences of the *trp*E gene and some control region sequences. Nonhybridized mRNA and "tails," including transcription from beyond the E gene region, were removed by T_1 RNase digestion. Hybridized RNA was purified by nitrocellulose filtration, as has been described earlier.[208]

The number of oligonucleotides observed on fingerprints of the RNase digests of the mRNA which had been isolated was expected to increase as the length of the transcribed region increased. T_1 RNase digests from the ^{32}P-labeled mRNA were separated on 2-D homochromatography. *Trp* ΔED102 mRNA, *trp* ΔED2 mRNA, and *trp* ΔED24 mRNA gave respectively more complex fingerprints. Oligonucleotides were isolated from the fingerprints, and their sequences were analyzed. As expected, the largest group (group III from *trp* ED24 mRNA) contained all of the oligonucleotides found in the smaller groups. Oligonucleotides from the largest group were fitted together to produce the sequence corresponding to the first 11 amino acids of the *trp*E gene.

Additional tests were done to show that the nucleotide sequence obtained from alignment of

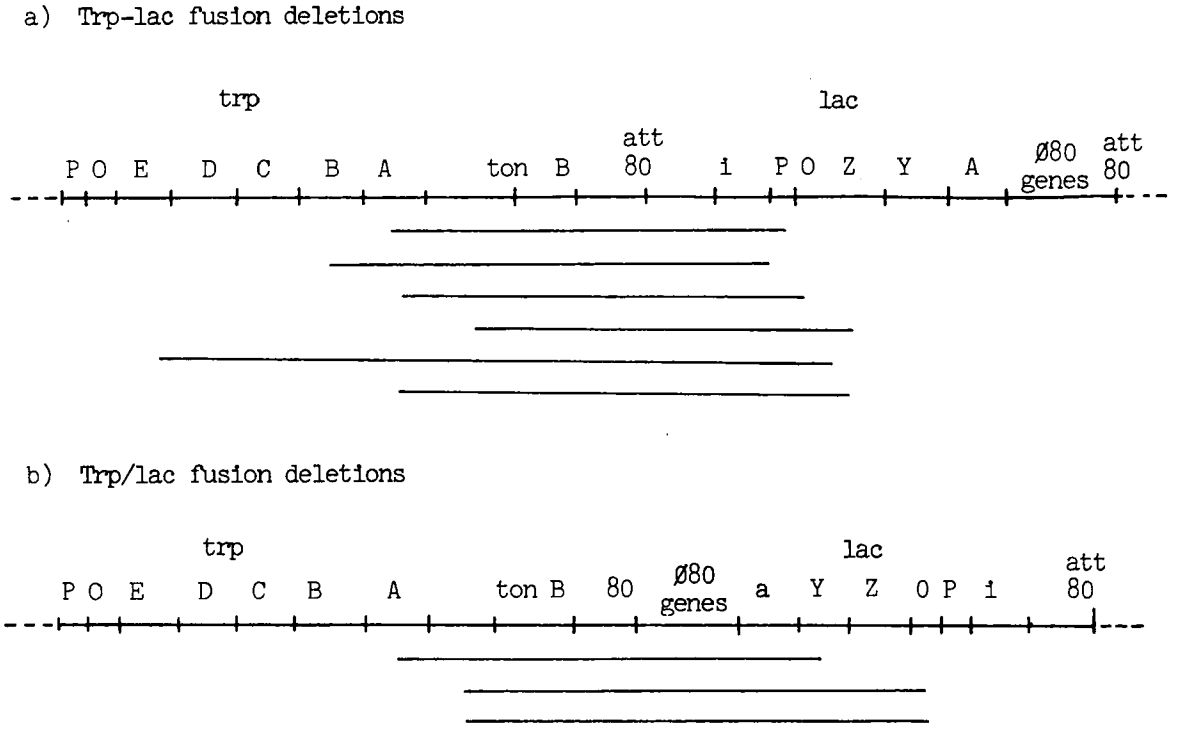


FIGURE 16. Fusion deletions used for preparation of purified *E. coli lac* control region DNA.[211] Details of the use of these mutants for purification of *lac* control region DNA are given in the text.

these oligonucleotides corresponds to the *trp*E amino acid sequence. Two mutants of *trp*E were characterized by RNA sequencing techniques.[209] *E. coli* strains W3110 *trp*E9914am, which carries a *trp*E frame-shift mutation, and W3110 *trp*E9914am, an amber mutation, were used to produce ^{32}P-labeled mRNA. Analysis of oligonucleotide fingerprints allowed determination of the sequence changes in the mRNA which produced these mutants.

A leader sequence of 11 nucleotides just prior to the A-U-G initiation codon has also been identified (see Figure 10d). The unassigned oligonucleotides from group III, which still included all oligonucleotides from the smaller groups, represented a sequence of about 120 nucleotides which precede the characterized leader. This brings the total leader size to about 140 nucleotides.

The positions of the control regions are still not defined. Possibly, studies of control region mutants will help investigators to locate the positions of the trp promoter and operator sequences, which may or may not be located in the leader region.

ii. The lac Control Regions

A significant achievement through the use of a combination of genetic and biochemical techniques is the sequence determination of the entire *E. coli lac* operon transcriptional control system. The complicated procedure was begun with the construction of a group of λ transducing phage mutants which could be used to isolate the *lac* control regions.[210] The function of these mutants was to define and subdivide the *lac* control regions in such a way that specific segments of RNA could be obtained from the control area for sequence analysis.

In order to obtain these mutants, Barnes et al.[210] began with two *E. coli* lysogens, a $\phi 80d_I$ *lac* lysogen and a $\phi 80d_{II}$ *lac* lysogen, In $\phi 80d_I$ *lac*, as shown in Figure 16a, deletion mutations caused by selecting against *tonB* usually originate near the terminus of the *trp* operon and enter the *lac* operon from the *i* gene side, ending near or within the *lac* control regions. In $\phi 80d_{II}$ *lac*, shown in Figure 16b, which has a reversed *lac* operon, mutations in *tonB* begin near the terminus of the *trp* operon and enter the *lac* operon from the *a* gene side, ending near the *lac* control regions.

These deletion mutants were then incorporated into plaque-forming transducing phages by recombination.

The procedure for sequence analysis by Dickson et al.[211] was to produce labeled RNA from λp*lac* DNA which contained the sequences from the *lac* control regions. In this case, transcription could start from a λ promoter and continue through the *lac* control regions and beyond into the *lac* structural genes.

The next step was to hydrolyze and remove the labeled RNA which was not produced by the *lac* control regions. This was done by hybridizing the RNA product to the appropriate strand of one of the λ*trp/lac* mutants (from the $\phi 80d_I$ *lac* lysogen) and then digesting the unhybridized "tails" with RNase. In general, sequences from the *i* gene and sections of λ prior to the *lac* control regions were degraded by this process. Then the RNA was isolated and rehybridized to the appropriate strand of one of the λ*trp/lac* mutants (from the $\phi 80d_{II}$ *lac* lysogen), and the unhybridized "tails" were again digested with RNase. In general, the translated genes of the *lac* operon and other later transcribed RNA were removed by this process. The remaining segment of RNA represented only the *lac* control region. Since a large number of deletion mutants were available for use in these experiments, the various sections of the *lac* control region obtained with different combinations of mutants aided in the ordering of oligonucleotides obtained from digestions of the control region fragments.

Standard RNA sequencing techniques were used to obtain a sequence of 122 base pairs from the *i* gene to the *z* gene which represents the *lac* control system. This sequence is shown in Figure 17. The mutations also define the approximate positions of control activity. In the region showing promoter activity, there are no large areas of twofold symmetry, as might be expected for control regions. There are, however, 3 alternating regions of high G-C and then high A-T content, each 12 base pairs in length. In addition, in the A-T rich region there is a repeating pentamer sequence G-A-A-A-T, which may also have some significance for RNA polymerase recognition. In the CAP binding site region there is an extensive area of twofold hyphenated symmetry similar to the symmetry in the operator region. Operator region sequences agree with those of Gilbert and Maxam[112] and Maizels.[103]

The binding sites for all control proteins are very close to each other and occupy most of the region between the *i* and *z* genes. Short distances between the binding sites suggest possible complicated interaction of control proteins and DNA.

iii. Recognition Sites on DNA Molecules for Protein Binding

Now that sequence information is available from several genetic control sites, the major problem is the determination of what aspects of the sequence are responsible for the control function at each site. Inspection of the sequences has provided some clues to their recognition by proteins. Figure 6 shows sequences from several promoter and operator sites. From the sequence of the *lac* control region expected to be a promoter site, Dickson et al.[211] have proposed that recognition may involve alternating concentrations of A-T pairs and G-C pairs. As indicated (sequence C), there is a G-C rich region (10 G-C pairs out of 12, or 10/12 G-C pairs), followed immediately by an A-T rich region (10/12 A-T pairs), and then another G-C rich region (9/12 G-C pairs).

We found that the pattern of alternating G-C rich and A-T rich regions is not restricted to the *lac* promoter site but, instead, is a common feature for all of the 5 sequences listed in Figure 6. The promoter site sequence of fd DNA has five rather short, alternating concentrations of A-T rich and G-C rich regions. The sequence starts with 7 A-T pairs out of 7 (7/7 A-T pairs), followed by 5/7 G-C pairs, 4/4 A-T pairs, 4/6 G-C pairs and 6/6 A-T pairs. The promoter site sequence of tyr-tRNA has three rather long, alternating concentrations of G-C rich and A-T rich regions. A region of 9/9 G-C pairs is followed by 10/12 A-T pairs and then 9/10 G-C pairs. In all three promoter sites (a,b,c) an A-T rich region is at the center. This A-T rich region may be partially denatured after the binding of RNA polymerase before the initiation of RNA synthesis.

The operator sites also contain the alternating G-C rich and A-T rich sites (see Figures 6d and e); however, the G-C rich region is at the center. In addition, symmetrical sequences about a specific axis are also present. The *lac* operator has extensive regions of twofold symmetry about the indicated axis. The λ operator region has a complex system of overlapping symmetries about three symmetry axes. In both cases, the central G-C rich region contains the symmetry.

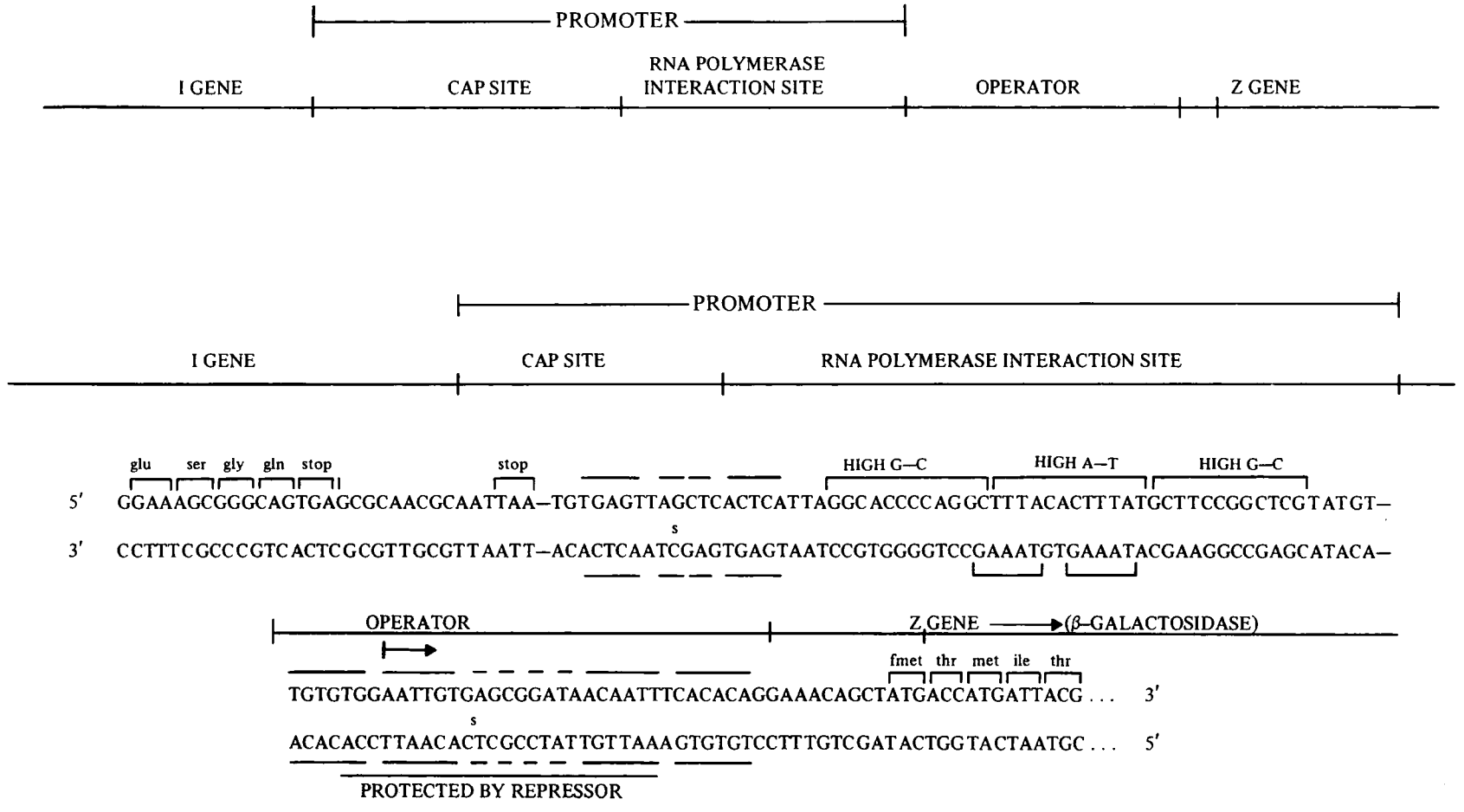


FIGURE 17. The *E. coli* lactose operon control region.[211] The topmost diagram shows the approximate positions for the various proteins which interact with the *lac* control regions. The lower portions of the figure show the complete nucleotide sequences which correspond to the *lac* control region.

Of the three promoter sites, the *lac* site shows some evidence of twofold symmetry, but does have the repeating sequence -TAAAG-TG-TAAAG in a central A-T rich region. The fd system has three twofold symmetry systems, two of them with an axis lying in the A-T rich region. The tyr-tRNA promoter has an extensive symmetrical region. The symmetry axis lies in the A-T rich region near the center of the promoter site.

The *lac* operator sequence has extensive regions in which nearly all nucleotides are symmetrical about the same axis. The λ operator sequence has some less extensive symmetries. The role of symmetry in the operator sequences for repressor binding was investigated for the *lac* operator through the use of operator mutants.[121,122] A group of operator constitutive mutants, which bind repressor poorly, were isolated under conditions of low Mg^{++} to increase the repressor-operator binding. Of the eight mutants which displayed single base pair changes in the operator region, five changes reduced the symmetry, one did not change the symmetry, and two extended the symmetry (Figure 6e). Since all of the mutants have impaired repressor binding, symmetry is clearly not the only important factor, but perhaps it is part of a more complex recognition pattern. This pattern may involve the A-T and G-C concentrations, and possibly other primary sequence information.

Another interesting fact is that five of the six mutations in the central G-C rich region of the operator convert a G-C to an A-T pair; in fact, six of the eight mutants are conversions of G-C to A-T.[121,122]

All of the sequences which display true twofold symmetry can be drawn in the form of a loop. It is possible that such loops provide a means of easy recognition of the binding sites. Other evidence[38] suggests that these structures are very unstable compared to the standard double-stranded configuration. Nevertheless, it seems possible that the repressor binding may significantly stabilize the loop formation. In incorrect binding positions, the loop may be drawn out momentarily, but the combination of DNA self-induced stabilization and protein induced stabilization would only be strong enough at the correct site.

It is puzzling that although the five promoter and operator sequences all contain symmetries, the number of symmetry axes, the sizes of the symmetrical regions, and the number and density of symmetrical nucleotides are all widely variable. The *lac* promoter site has little twofold symmetry when compared to the large, dense symmetry in the tyr tRNA promoter site. Assuming that all RNA polymerase binding sites share at least one major feature, extensive twofold symmetry would not appear to be this feature. On the other hand, it appears to us that alternating adjacent A-T rich and G-C rich regions may be essential, since this feature is shared in all five promoter and operator sequences. Possibly just single adjacent regions which are A-T rich and then G-C rich are sufficient for recognition. Additional A-T rich or G-C rich regions to make a triple concentration, or fivefold concentration, may be a more complex signal than the binding sequence itself. Also, it may be significant that symmetry in the promoter sites centers about A-T rich regions, while symmetry in the operator sites centers about G-C rich regions.

4. Use of Other Enzymes for DNA Sequence Analysis

a. Terminal Deoxynucleotidyl Transferase

Terminal deoxynucleotidyl transferase from calf thymus catalyzes the addition of mononucleotides from deoxynucleoside triphosphates to the 3′ hydroxy termini of DNA molecules.[212] It was shown by Roychoudhury and Kössel[213] that the enzyme can be made to add a single ribonucleotide to the 3′ terminus of a deoxyoligonucleotide. This result was used by Roychoudhury et al.[48] to partially sequence a synthetic octadeoxynucleotide. The oligonucleotide was partially degraded from the 3′ end with venom phosphodiesterase to yield a population of oligonucleotide products, each shorter by one nucleotide from the 3′ end. A ^{32}P-rA was attached to the 3′ termini of the products. The labeled products were separated on the basis of size. Then, the nearest neighbor of the ^{32}P-rA was determined for each product. The combination of size and nearest neighbor data gave the sequence. One problem with this technique is that terminal deoxynucleotidyl transferase does not label trimers and smaller oligonucleotides.

b. Exonuclease I

E. coli exonuclease I degrades an oligonucleotide to mononucleotides and a dinucleotide containing the original 5′ terminus.[214] Weiss and Richardson[215] first made use of this enzyme to analyze the 5′ terminal dinucleotides from T_7 DNA. Ziff, Sedat, and Galibert[130] and Galibert,

Sedat, and Ziff[131] recently used this enzyme to sequence oligonucleotides from ϕX174 DNA. Exonuclease I was used to digest a series of separated partial spleen phosphodiesterase digestion products of an oligonucleotide, each shorter by one nucleotide from the 5′ end. The exonuclease I digestion was performed individually on each partial degradation product, yielding the 5′ dinucleotide. The sequence of each dinucleotide was easily determined by complete spleen and venom phosphodiesterase digestion and standard chromatographic analysis. From the 5′ termini of all the dinucleotides, the sequence of the original oligonucleotide can be deduced. The 3′ termini of the dinucleotides represent the same sequence, frame shifted by one nucleotide. Although this method is rather time-consuming, one big advantage is that it contains an internal double check on the sequence being analyzed.

c. DNA Polymerase Catalyzed Exchange Reaction

The combined exonuclease-repair functions of phage T_4 polymerase have been used by Englund[216] to determine short DNA sequences at the 3′ termini of double-stranded DNA. At 37°C, 70 m*M* salt, linear duplex DNA is degraded extensively by the exonuclease function of T_4 polymerase if no deoxynucleoside triphosphates are present in the incubation mixture. The presence of a single deoxynucleoside triphosphate drastically reduces degradation. This is because it serves as a substrate to replace the corresponding loss of a nucleotide by degradation. Replacement can be made much faster than removal, so that the reaction becomes an alternating incorporation and removal of the 3′ terminal nucleotide. An important fact is that if the terminal nucleotide is not the same as the nucleotide in solution, it will be removed and not replaced. Subsequent nucleotides will be removed until the correct nucleotide for exchange incorporation is reached.

Weigel et al.[15] used this technique in conjunction with other techniques to obtain some 3′ terminal sequence information from λ DNA. The sequences which were obtained are given in Figure 3.

Using *E. coli* DNA polymerase I, Donelson and Wu[217] showed that at 5°C, 180 m*M* salt, only the 3′ terminal nucleotide was significantly exchangeable. The exchange reaction, under these conditions, followed by nearest neighbor analysis, was used to determine the dinucleotide sequences at the 3′ termini of T_7 DNA.

d. T_4 Polynucleotide Kinase Exchange

Since many native DNA molecules and oligonucleotide digestion products have 5′ phosphate groups, labeling of these molecules with T_4 polynucleotide kinase has often been a tedious operation. λ DNA, for example, had to be treated with bacterial alkaline phosphatase to remove the 5′ phosphates. The phosphatase then had to be removed by phenol extraction, and the phenol exhaustively dialyzed away, before ^{32}P could be incorporated onto the 5′ ends with T_4 polynucleotide kinase.

Recently, however, van de Sande et al.[218] have shown that polynucleotide kinase can be used to exchange ^{32}P from [γ-^{32}P] ATP with the 5′ phosphate group on an oligonucleotide in 1 step. In this exchange, a reverse kinase reaction removes the unlabeled 5′ phosphate from the oligonucleotide to phosphorylate ADP which is added in the solution. Simultaneously, the forward kinase reaction phosphorylates the 5′ end of the oligonucleotide in the normal fashion. It was found that in the presence of ADP, [γ-^{32}P] ATP, and a 5′ phosphorylated oligonucleotide, at pH 7.6, polynucleotide kinase can almost quantitatively exchange the unlabeled 5′ phosphate for a ^{32}P. Preliminary experiments suggest that this technique can also be used to label long DNA molecules.[219,220] This exchange method represents a significant advantage over the method previously used.

e. DNA Sequence Frequency Analysis

Besides direct DNA sequence analysis, another source of DNA sequence information is frequency analysis,[153,154,221] which can be considered to be an extension of nearest neighbor analysis.[148] This type of analysis involves comparison of the composition of the 3′ terminal, 5′ terminal, and 5′ penultimate nucleotides of oligonucleotides generated from 2 different DNA sources by 1 or several semispecific enzymes. The semispecific DNases which have been used in these investigations include hog spleen acid-DNase, acid-DNase from the hepatopancreas of the snail *Helix aspersa* (Müll), bovine pancreatic DNase, and *E. coli* endonuclease I. Compositional patterns from repetitive DNAs and eukaryotic DNAs have been found to deviate in characteristic ways from those of bacterial DNAs. This information will be discussed by Bernardi[222] in great detail in another article to be published in *CRC Critical Reviews in Biochemistry*.

f. Use of Venom Phosphodiesterase to Degrade 3′ Phosphorylated Oligonucleotides for Analysis

Although the pH optimum for venom phosphodiesterase activity on 3′ dephosphorylated oligonucleotides is 9.0, at this pH the degradation of phosphorylated oligonucleotides is strongly inhibited. Richards and Laskowski[223] have found that hydrolysis rates of short 3′ phosphorylated oligonucleotides can be significantly increased by lowering the pH of the reaction mixture to 6.0, where there is a lesser 3′ negative charge on the oligonucleotide. The venom phosphodiesterase digestion produces 3′,5′ phosphorylated mononucleotide products from the 3′ nucleotides of 3′ phosphorylated oligonucleotides.[224] Quantitation of these mononucleotides can be carried out for 3′ end analysis of small amounts of unlabeled oligonucleotides.

5. Other Methods for the in vitro Labeling of DNA

All of the methods developed thus far for DNA sequence analysis involve using radioactive compounds. There are three ways by which labeled DNA may be obtained:

a. In vivo labeling with ^{32}P or ^{33}P, using radioactive inorganic phosphates, or with ^{3}H or ^{14}C, using radioactive nucleosides or nucleotides.

b. In vitro labeling with ^{32}P, ^{33}P, ^{3}H, or ^{14}C, using various polymerases and radioactive nucleoside triphosphates to copy the DNA to give complementary products of high specific activity.

c. Labeling by chemical methods with radioactive reagents.

Until recently, almost all of the labeled DNA had been obtained by the first two methods. The first method gave DNA of rather low specific activity, so the second method was used whenever possible. In cases where the second method is not applicable, a third method, whereby purified DNA or polynucleotides can be labeled with high specific activity, would be most useful. In some of these cases, enzymatic terminal labeling using T_4 polynucleotide kinase or calf thymus deoxynucleotidyl terminal transferase has been used to provide some sequence information. Two methods have recently been introduced for internal labeling of whole DNA, RNA, or polynucleotides with high specific activity by means of chemical modifications.

Iodination of nucleosides and nucleotides can be effected by a variety of methods. Two methods had been adapted for iodination of polynucleotides using ICl[225] and *N*-iodosuccinimide.[226] However, neither has been found to be fully satisfactory. Commerford[227] introduced a simple method for iodination of nucleic acids using a solution of iodine or thallic trichloride and iodide at pH 5. Over 95% of the cytosine residues in DNA and RNA and a small amount of uracil in RNA were found to be iodinated. Both yielded 5-iodo derivatives. This method has also been successfully adapted for use in the iodination of nucleic acids with radioactive ^{125}I.

Prensky et al.[228] have described a modification of Commerford's iodination procedure by which RNA of high specific radioactivity (10^6 to 10^8 dpm/μg) can be obtained. Later, Robertson et al.[229] labeled several prokaryotic and eukaryotic RNAs by this method and examined their fingerprints after T_1 RNase digestion. They showed that although the iodinated oligonucleotides in the 2-D electrophoresis fingerprint show altered electrophoretic behavior, in comparison to their noniodinated counterparts, which were labeled with ^{32}P, the fingerprints were quite similar in complexity, and are useful for comparative fingerprinting studies. They have further shown that these oligonucleotides, labeled at the C and U residues, can be hydrolyzed by the common RNases: pancreatic RNase, RNase U_2, RNase T_2, and RNase T_1. However, the labeled forms are unstable to the alkaline conditions used for hydrolysis of RNA.

Another method for in vitro labeling of nucleic acids has been introduced by Dale, Livingston, and Ward.[230] Labeled 5-mercuriacetate derivatives of UTP, CTP, dUTP, and dCTP, prepared by an acetoxymercuration reaction with [^{203}Hg]mercuric acetate, have been successfully used as substrates for various nucleic acid polymerases. These modified nucleotides, in the absence of added mercaptan, are not incorporated, and in most instances are potent inhibitors. However, after their conversion to mercurithio derivatives by a variety of mercaptans, widely differing in structure and size, of which mercaptoethanol seemed most effective, the derivatives have been shown to be excellent substrates for various enzymes. *E. coli* and T_7 RNA polymerase, *E. coli* DNA polymerase I, and DNA polymerase from avian myeloblastosis virus have been tested. Extensive incorporation of mercurinucleotides by these

polymerases into calf thymus DNA has been achieved.

[^{203}Hg]-TTP and [^{203}Hg]-dUTP have also been demonstrated to be efficient substrates for calf thymus terminal transferase.[230] These derivatives, however, have more restrictive substrate requirements. Only the smaller methyl- and ethyl-mercapto derivatives are incorporated into oligo $(dT)_6$ and calf thymus DNA primers.

Although neither of these two in vitro chemical labeling methods has been used for DNA sequence analysis, in certain cases they may have a considerable advantage over the more common methods. The iodine-labeled RNAs have been shown to be degraded by the common RNases just as unmodified RNAs. However, it has not been shown whether the iodine-labeled DNas are susceptible to DNases and, in particular, exonucleases like venom and spleen phosphodiesterases. The fidelity of the polymerization of the ^{203}Hg-labeled compounds by various nucleic acid polymerases, and the susceptibility of the products to nucleases, have yet to be determined.

Because of the specificity of the acetoxymercuration reaction, which modifies only C residues in DNA, it can be used for future DNA sequence analysis by electron microscopy. Already, controlled conditions have been established for direct mercuration of polynucleotides, RNA, and DNA. Pyrimidine residues in single- and double-stranded polymers are modified at essentially the same rate. Apart from this obvious possible application of the acetoxymercuration reaction, the authors[231] have observed that these mercury containing compounds are retained by sulfhydryl-sepharose columns. Thus, by selective labeling of particular segments of DNA or RNA, they can be effectively separated from other segments of the molecule after endonuclease digestion.

Acknowledgments

We are very grateful to all our friends who kindly provided us with manuscripts and results in advance of publication, and to our colleagues for their helpful discussions. Research that originated in the authors' laboratory was supported by research grants GM-18887 and CA-14989 from the National Institutes of Health, NIH Training Grant 5TGM 00824-13 and Grant BMS 73-01859 A01 from the National Science Foundation.

REFERENCES

1. **Wu, R., Donelson, J., Padmanabhan, R., and Hamilton, R.,** Determination of primary nucleotide sequences in DNA molecules, *Bull. Inst. Pasteur,* 70, 203, 1972.
2. **Murray, K. and Old, R. W.,** The primary structure of DNA, in *Progress in Nucleic Acid Research and Molecular Biology,* Vol. 14, Cohn, W. E., Ed., Academic Press, New York, 1974, 117.
3. **Salser, W.,** DNA sequencing techniques, *Annu. Rev. Biochem.,* 43, 923, 1974.
4. **Kaiser, A. D. and Wu, R.,** Structure and base sequences in the cohesive ends of bacteriophage lambda DNA, *J. Mol. Biol.,* 35, 523, 1968.
5. **Wu, R.,** Nucleotide sequence analysis of DNA. I. Partial sequence of the cohesive ends of bacteriophage λ DNA and 186 DNA, *J. Mol. Biol.,* 51, 501, 1970.
6. **Wu, R. and Taylor, E.,** Nucleotide sequence analysis of DNA. II. Complete nucleotide sequence of the cohesive ends of bacteriophage λ DNA, *J. Mol. Biol.,* 57, 491, 1971.
7. **Wu, R. and Kaiser, A. D.,** Structure and base sequence in the cohesive ends of bacteriophage lambda DNA, *J. Mol. Biol.,* 35, 523, 1968.
8. **Wu, R., Padmanabhan, R., and Bambara, R.,** Nucleotide sequence analysis of bacteriophage DNA, in *Methods in Enzymology,* Vol. 29E, Grossman, L. and Moldave, K., Eds., Academic Press, New York, 1974, 231.
9. **Hedgpeth, J., Goodman, H. M. and Boyer, H. W.,** DNA nucleotide sequence restricted by the RI endonuclease, *Proc. Natl. Acad. Sci. USA,* 69, 3448, 1972.
10. **Padmanabhan, R. and Wu, R.,** Nucleotide sequence analysis of DNA. IV. Complete nucleotide sequence of the left-hand cohesive end of coliphage 186 DNA, *J. Mol. Biol.,* 65, 447, 1972.
11. **Fiandt, M., Hradecna, Z., Lozeron, H. A. and Szybalski, W.,** Electron micrographic mapping of deletions, insertions, inversions, and homologies in the DNA's, of coliphages, lambda and phi 80, in *The Bacteriophage Lambda,* Hershey, A. D., ed. Cold Spring Harbor Laboratory, New York, 1971, 329.

12. **Wang, J. C. and Kaiser, A. D.,** Evidence that the cohesive ends of mature λ DNA are generated by the gene A product, *Nat. New Biol.*, 241, 16, 1973.
13. **Bambara, R., Padmanabhan, R., and Wu, R.,** Complete nucleotide sequence of the cohesive ends of bacteriophage φ80 DNA, *J. Mol. Biol.*, 75, 741, 1973.
14. **Murray, K. and Murray, N. E.,** Terminal nucleotide sequences of DNA from temperate coliphages, *Nat. New Biol.*, 243, 134, 1973.
15. **Weigel, R. H., Englund, P. T., Murray, K., and Old, R. W.,** Terminal nucleotide sequences of bacteriophage λ DNA, *Proc. Natl. Acad. Sci. USA.*, 70, 1151, 1973.
16. **Brezinski, D. P. and Wang, J. C.,** The 3'-terminal nucleotide sequences of λ DNA, *Biochem. Biophys. Res. Commun.*, 50, 398, 1973.
17. **Ghangas, G. S., Jay, E., Bambara, R., and Wu, R.,** Nucleotide sequence analysis of DNA. XI. The 3'-terminal sequences of bacteriophage λ and φ80 DNA, *Biochem. Biophys. Res. Commun.*, 54, 998, 1973.
18. **Bambara, R., Ghangas, G. S., Jay, E., and Wu, R.,** Recognition sequences of lambda and φ80 DNA in the double stranded region adjacent to the cohesive ends, abstr. in *Fed. Proc.*, 32, 664, 1973.
19. **Jay, E., Bambara, R., Padmanabhan, R., and Wu, R.,** DNA sequence analysis: a general, simple and rapid method for sequencing large oligodeoxyribonucleotide fragments by mapping, *Nucleic Acids Res.*, 1, 331, 1974.
20. **Bernardi, G.,** Mechanism of action and structure of acid deoxyribonuclease, *Adv. Enzymol.*, 31, 1, 1968.
21. **Kelly, T. J. and Smith, H. O.,** A restriction enzyme from *Hemophilus influenzae*. II. Base sequence of the recognition site, *J. Mol. Biol.*, 51, 393, 1970.
22. **Murray, K.,** personal communication, 1974.
23. **Marians, K., Padmanabhan, R., and Wu, R.,** Complete nucleotide sequence of the right-hand cohesive end of coliphage 186 DNA, manuscript in preparation, 1974.
24. **Padmanabhan, R., Wu, R., and Calendar, R.,** Complete nucleotide sequence of the cohesive ends of bacteriophage P2 DNA, submitted to *J. Biol. Chem.*, 1974.
25. **Wang, J. C. and Brezinski, D. P.,** Alignment of two DNA helices: a model for recognition of DNA base sequences by the termini-generating enzymes of phage λ, 186 and P2, *Proc. Natl. Acad. Sci. USA*, 70, 2667, 1973.
26. **Gupta, N. K. and Khorana, H. G.,** Studies on polynucleotides. XC. DNA polymerase-catalyzed repair of short DNA duplexes with single-stranded ends, *Proc. Natl. Acad. Sci. USA*, 60, 215, 1968.
27. **Kleppe, K., Ohtsuka, E., and Khorana, H. G.,** Repair and replication of short synthetic DNA's by DNA polymerase, abstr. in *Fed. Proc.*, 29, 405, 1970.
28. **Kleppe, K., Ohtsuka, E., Kleppe, R., Molineux, I., and Khorana, H. G.,** Studies on polynucleotides. XCVI. Repair replication of short synthetic DNA's as catalyzed by DNA polymerase, *J. Mol. Biol.*, 56, 341, 1971.
29. **Goulian, M., Goulian, S. H., Codd, E. E., and Blaumanfield, A. Z.** Properties of oligodeoxynucleotides that determine priming activity with *Escherichia coli* deoxyribonucleic acid polymerase I, *Biochemistry*, 12, 2893, 1973.
30. **Gupta, N. K., Ohtsuka, E., Weber, H., Chang, S. H., and Khorana, H. G.,** Studies on polynucleotides. LXXVII. The joining of short deoxyribopolynucleotides by DNA-joining enzymes, *Proc. Natl. Acad. Sci. USA*, 60, 285, 1968.
31. **Wu, R.,** Nucleotide sequence analysis of DNA, *Nat. New Biol.*, 236, 198, 1972.
32. **Heimer, E. P., Ahmad, M., and Nussbaum, A. L.,** Chemical synthesis of the "sticky end" of lambda phage DNA r-strand, *Biochem. Biophys. Res. Commun.*, 48, 348, 1972.
33. **Padmanabhan, R., Padmanabhan, R., and Wu, R.,** Nucleotide sequence analysis of DNA. IX. Use of oligonucleotides of defined sequence as primers in DNA sequence analysis, *Biochem. Biophys. Res. Commun.*, 48, 1295, 1972.
34. **Oertel, W. and Schaller, H.,** A new approach to the sequence analysis of DNA, *FEBS Lett.*, 27, 316, 1972.
35. **Thomas, C. A., Jr.,** Recombination of DNA molecules, in *Progress in Nucleic Acid Research and Molecular Biology*, Vol. 5, Davidson, J. N. and Cohn, W. E., Eds., Academic Press, New York, 1966, 315.
36. **Crick, F. H. C.,** Codon-anticodon pairing: the wobble hypothesis, *J. Mol. Biol.*, 19, 548, 1966.
37. **Uhlenbeck, O. C., Martin, F. H., and Doty, P.,** Self-complementary oligoribonucleotides: effects of helix defects and guanylic acid-cytidylic acid base pairs, *J. Mol. Biol.*, 57, 217, 1971.
38. **Tinoco, I., Jr., Uhlenbeck, O. C., and Levine, M. O.,** Estimation of secondary structure in ribonucleic acids, *Nature*, 230, 362, 1971.
39. **Wu, R., Tu, C. D., and Padmanabhan, R.,** Nucleotide sequence anlysis of DNA. XII. The chemical synthesis and sequence analysis of a dodecadeoxynucleotide which binds to the endolysin gene of bacteriophage lambda, *Biochem. Biophys. Res. Commun.*, 55, 1092, 1973.
40. **Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzaghi, E., and Inouye, M.,** Frameshift mutations and the genetic code, *Cold Spring Harbor Symp.*, 31, 77, 1966.
41. **Padmanabhan, R., Jay, E., and Wu, R.,** Chemical synthesis of a primer and its use in the sequence analysis of the lysozyme gene of bacteriophage T4, *Proc. Natl. Acad. Sci. USA*, 71, 2510, 1974.
42. **Berg, P., Fancher, H., and Chamberlin, M.,** The synthesis of mixed polynucleotides containing ribo- and deoxyribonucleotides by purified preparations of DNA polymerase from *Escherichia coli*, in *Symposium on Information Macromolecules*, Vogel, H. Bryson, V., and Lampen, J. O., Eds., Academic Press, New York, 1963, 467.
43. **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. USA*, 68, 238, 1972.

44. **Lillehaug, J. R. and Kleppe, K.,** Effect of pH on incorporation of ribonucleotides into DNA by DNA polymerase I, *FEBS Lett.*, 40, 339, 1974.
45. **van de Sande, J. H., Loewan, P. C., and Khorana, H. G.,** A further study of ribonucleotide incorporation into deoxyribonucleic acid chains by DNA polymerase I of *Escherichia coli*, *J. Biol. Chem.*, 247, 6140, 1972.
46. **Sanger, F., Donelson, J. E., Coulson, A. D., Kössel, H., and Fischer, D.,** Use of DNA polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in phage f1 DNA, *Proc. Natl. Acad. Sci. USA*, 70, 1209, 1973.
47. **Asbeck, F., Beyruther, K., Kohler, H., Wettstein, G., and Braunitzer, G.,** Virusproteine. IV. Die Konstitution des hullproteins des phagen fd, *Hoppe-Seyler's Z. Physiol. Chem.*, 350, 1047, 1969.
48. **Roychoudhury, R., Fischer, D., and Kössel, H.,** A new method for the sequence analysis of oligodeoxynucleotides, *Biochem. Biophys. Res. Commun.*, 45, 430, 1971.
49. **Snell, D. T. and Offord, R. E.,** The amino acid sequence of the β-protein of bacteriophage ZJ-2, *Biochem. J.*, 127, 167, 1972.
50. **Loewen, P. C. and Khorana, H. G.,** The dodecanucleotide sequence adjoining the C-C-A end of the tyrosine transfer ribonucleic acid gene, *J. Biol. Chem.*, 248, 3489, 1973.
51. **Loewen, P. C., Sekiya, T., and Khorana, H. G.,** The nucleotide sequence adjoining the C-C-A end of an *Escherichia coli* tyrosine transfer ribonucleic acid gene, *J. Biol. Chem.*, 249, 217, 1974.
52. **Russell, R. L., Abelson, J. N., Landy, A., Gefter, M. L., Brenner, S., and Smith, J. D.,** Duplicate genes for tyrosine transfer RNA in *Escherichia coli*, *J. Mol. Biol.*, 47, 1970.
53. **Sekiya, T. and Khorana, H. G.,** The nucleotide sequence in the promoter region of the gene for an *E. coli* tyrosine transfer ribonucleic acid, *Proc. Natl. Acad. Sci. USA*, 71, 2978, 1974; Sekiya, T., Van Ormondt, H., and Khorana, H. G., personal communication.
54. **Ptashne, M.,** Repressor, operators and promoters in bacteriophage lambda, *Harvey Lect.*, in press, 1974.
55. **Kleid, D., Agarwal, K., and Khorana, H. G.,** personal communication, 1974.
56. **Schaller, H.,** personal communication, 1974.
57. **Zain, B. S., Dhar, R., Weissman, S. M., Lebowitz, P., and Lewis, A. M., Jr.,** Preferred site for initiation of RNA transcription by *Escherichia coli* RNA polymerase within the simian virus 40 DNA segment of the nondefective adenovirus-simian virus 40 hybrid viruses $Ad2^{+}ND_1$ and $Ad2^{+}ND_3$, *J. Virol.*, 11, 682, 1973.
58. **Dhar, R., Zain, S., Weissman, S. M., Pan, J., and Subramanian, K. N.,** Nucleotide sequences of RNA transcribed in infected cells and by *Escherichia coli* RNA polymerase from a segment of simian virus 40 DNA, *Proc. Natl. Acad. Sci. USA*, 71, 371, 1974.
59. **Gierer, A.,** Model for DNA and protein interactions and the function of the operator, *Nature*, 212, 1480, 1966.
60. **Maniatis, T., Ptashne, M., Barrell, B. G., and Donelson, J. E.,** Sequence of a repressor-binding site in the DNA of bacteriophage λ, *Nature*, 250, 394, 1974.
61. **Bambara, R. and Wu, R.,** unpublished results.
62. **Radding, C.,** The purification of β protein and exonuclease made by phage λ, in *Methods in Enzymology*, Vol. XXI D, Grossman, L. and Moldave, K., Eds., Academic Press, New York, 1971, 273.
63. **Harvey, C. L., Wright, R., and Nussbaum, A. L.,** Lambda phage DNA: joining of a chemically synthesized cohesive end, *Science*, 179, 291, 1973.
64. **Meselson, M. and Yuan, R.,** DNA restriction enzyme from *E. coli*, *Nature*, 217, 1110, 1968.
65. **Smith, H. O. and Wilcox, K. W.,** A restriction enzyme from *Hemophilus influenzae*. I. Purification and general properties, *J. Mol. Biol.*, 51, 379, 1971.
66. **Arber, W.,** DNA modification and restriction, in *Progress in Nucleic Acid Research and Molecular Biology*, Vol. 14, Cohn, W. E., Ed., Academic Press, New York, 1974, 1.
67. **Arber, W. and Linn, S.,** DNA modification and restriction, *Annu. Rev. Biochem.*, 38, 467, 1969.
68. **Meselson, M., Yuan, R., and Haywood, J.,** Restriction and modification of DNA, *Annu. Rev. Biochem.*, 41, 447, 1972.
69. **Sharp, P. A., Sugden, B., and Sambrook, J.,** Detection of two restriction endonuclease activities in *Haemophilus parainfluenzae* using analytical agarose-ethidium bromide electrophoresis, *Biochemistry*, 12, 3055, 1973.
70. **Smith, H. O. and Nathans, D.,** A suggested nomenclature for bacterial modification and restriction systems and their enzymes, *J. Mol. Biol.*, 81, 419, 1973.
71. **Danna, K. and Nathans, D.,** Specific cleavage of Simian virus 40 DNA by restriction endonuclease, *Proc. Natl. Acad. Sci. USA*, 68, 2913, 1971.
72. **Mulder, C. and Delius, H.,** Specificity of the break produced by restricting endonuclease R_1 in Simian virus 40 DNA, as revealed by partial denaturation mapping, *Proc. Natl. Acad. Sci. USA*, 69, 3215, 1972.
73. **Morrow, J. F. and Berg, P.,** Cleavage of Simian virus 40 DNA at a unique site by a bacterial restriction enzyme, *Proc. Natl. Acad. Sci. USA*, 69, 3365, 1972.
74. **Griffin, B., Fried, M., and Cowie, A.,** Polyoma DNA: a physical map, *Proc. Natl. Acad. Sci. USA*, 71, 2077, 1974.
75. **Danna, K. J., Sack, G. H., Jr., and Nathans, D.,** Studies on Simian virus 40 DNA. VII. A cleavage map of the SV40 genome, *J. Mol. Biol.*, 78, 363, 1973.
76. **Subramanian, K. N., Pan, J., Zain, S., and Weissman, S. M.,** The mapping and ordering of fragments of SV40 DNA produced by restriction endonucleases, *Nucleic Acids Res.*, 1, 727, 1974.

77. **Garfin, D. E. and Goodman, H. M.,** Nucleotide sequences at the cleavage sites of two restrictions endonucleases from *Hemophilus parainfluenzae, Biochem. Biophys. Res. Commun.*, 59, 108, 1974.
78. **Lee, A. S. and Sinsheimer, R. L.,** A cleavage map of bacteriophage φX174 genome, *Proc. Natl. Acad. Sci. USA*, 71, 2882, 1974.
79. **Mulder, C., Arrand, J. R., Delius, H., Keller, W., Pettersson, U., Roberts, R., and Sharp, P.,** Cleavage maps of DNA from adenovirus types 2 and 5 by restriction endonucleases *Eco*RI and *Hpa*I, *Cold Spring Harbor Symp. Quant. Biol.*, 39, 1974, in press.
80. **Inman, R. B. and Schnös, M.,** Partial denaturation of thymine- and 5-bromobracil-containing λ DNA in alkali, *J. Mol. Biol.*, 49, 93, 1970.
81. **Allet, B., Jeppesen, P. G. N., Karagin, V. J., and Delius, H.,** Mapping the DNA fragments produced by cleavage of λ DNA with endonuclease RI, *Nature*, 241, 120, 1973.
82. **Roberts, R. J., Arrand, J. R., and Myers, P. A.,** personal communication, 1974.
83. **Murray, K. and Morrison, A.,** 1974; Murray, K., Hughes, S. G., Brown, J. S., and Fleming, S. A., personal communication, 1974.
84. **Murray, K. and Murray, N.,** 1974; Davis, R., personal communication, 1974.
85. **Boyer, H. W., Chow, L. T., Dugaiczyk, A., Hedgpeth, J., and Goodman, H. M.,** DNA substrate site for the EcoRII restriction endonuclease and modification methylase, *Nat. New Biol.*, 224, 40, 1973.
86. **Middleton, J. H., Edgell, M. H., and Hutchison, C. A.,** Specific fragments of φX174 deoxyribonucleic acid produced by a restriction enzyme from *Haemophilus aegyptius*, Endonuclease Z, *J. Virol.*, 10, 42, 1972.
87. **Sugisaki, H. and Takanami, M.,** DNA sequence restricted by restriction endonuclease AP from *Haemophilus aphirophilus, Nat. New Biol.*, 246, 138, 1973.
88. **Takanami, M.,** Restriction endonucleases AP, GA, and H-I from three *Haemophilus* strains, *Methods in Molecular Biology*, 1974, in press.
89. **Bigger, C. H., Murray, K., and Murray, N. E.,** Recognition sites in bacteriophage λ DNA for a restriction endonuclease from an *Escherichia coli* f1 R-factor (RII), *Nat. New Biol.*, 244, 7, 1973.
90. **Bron, S., Trautner, T., and Murray, K.,** personal communication.
91. **Middleton, J. H., Shankus, D., Eagle, M. H., and Hutchison, C. A.,** personal communication, 1974.
92. **Landy, A.,** personal communication.
93. **Jay, E., Roychoudhury, R., and Wu, R.,** unpublished results.
94. **Takanami, M. and Kojo, H.,** Cleavage site specificity of an endonuclease prepared from *Haemophilus influenzae* strain H-I, *FEBS Lett.*, 29, 267, 1973.
95. **Gromkova, R. and Goodal, S. H.,** Action of Haemophilus endodeoxyribonuclease of biologically active deoxyribonucleic acid, *J. Bacteriol.*, 109, 987, 1972.
96. **Mulder, C. and Greene, R.,** personal communication.
97. **Wu, R., King, C., and Jay, E.,** unpublished results.
98. **Murray, K., Old, R. W., and Poizes, G.,** personal communication.
99. **Weissman, S.,** personal communication.
100. **Steitz, J. A.,** Polypeptide chain initiation: nucleotide sequences of the three ribosomal binding sites in bacteriophage R17 RNA, *Nature*, 224, 957, 1969.
101. **Arrand, J. R. and Hindley, J.,** Nucleotide sequence of a ribosome binding site on RNA synthesized *in vitro* from coliphage T7, *Nat. New Biol.*, 224, 10, 1973.
102. **Bronson, M. J., Squires, C., and Yanofsky, C.,** Nucleotide sequences from tryptophan messenger RNA of *E. coli*: the sequence corresponding to the amino-terminal region of the first polypeptide specified by the operon, *Proc. Natl. Acad. Sci. USA*, 70, 2335, 1973.
103. **Maizels, N.,** The nucleotide sequence of the lactose messenger ribonucleic acid transcribed from the UV5 promoter mutant of *Escherichia coli, Proc. Natl. Acad. Sci. USA*, 70, 3585, 1973.
104. **Pieczenik, G., Model, P., and Robertson, H. D.,** Sequence and symmetry in ribosome binding sites of bacteriophage f1 RNA, *J. Mol. Biol.*, in press, 1974.
105. **Robertson, H. D., Barrell, B. G., Weith, H. L., and Donelson, J. E.,** Isolation and sequence analysis of a ribosome-protected fragment from bacteriophage φX174 DNA, *Nat. New Biol.*, 241, 38, 1973.
106. **Bretscher, M. S.,** Ribosome initiation and the mode of action of neomycin in the direct translation of single-stranded fd DNA, *Cold Spring Harbor Symp. Quant. Biol.*, 34, 651, 1969.
107. **Ling, V.,** Partial digestion of ^{32}P-fd DNA with T4 endonuclease IV, *FEBS Lett.*, 19, 50, 1971.
108. **Robertson, H. D.,** Isolation of specific ribosome binding sites from single-stranded DNA, *J. Mol. Biol.*, submitted, 1974.
109. **Gilbert, W. and Muller-Hill, B.,** Isolation of the *lac* repressor, *Proc. Natl. Acad. Sci. USA*, 56, 1891, 1966.
110. **Gilbert, W. and Muller-Hill, B.,** The *lac* operator is DNA, *Proc. Natl. Acad. Sci. USA*, 58, 2415, 1967.
111. **Gilbert, W.,** The *lac* repressor and the *lac* operator, in *Polymerization in Biological Systems*, Ciba Foundation Symposium 7 (New Series), Elsevier, Amsterdam, 1972, 245.
112. **Gilbert, W. and Maxam, A.,** The nucleotide sequences of the *lac* operator, *Proc. Natl. Acad. Sci. USA*, 70, 3581, 1973.

113. **Riggs, A. D. and Bourgeois, S.,** On the assay, isolation and characterization of the *lac* repressor, *J. Mol. Biol.*, 34, 361, 1968.
114. **Ptashne, M.,** Isolation of the λ phage repressor, *Proc. Natl. Acad. Sci. USA*, 57, 306, 1967.
115. **Ptashne, M.,** Specific binding of the λ phage repressor to λ DNA, *Nature*, 214, 232, 1967.
116. **Ptashne, M. and Hopkins, N.,** Operators controlled by the λ phage repressor, *Proc. Natl. Acad. Sci. USA*, 60, 1292, 1968.
117. **Maniatis, T. and Ptashne, M.,** Multiple repressor binding at the operators in bacteriophage λ, *Proc. Natl. Acad. Sci. USA*, 70, 1531, 1973.
118. **Maurer, R., Maniatis, T., and Ptashne, M.,** Promoters are the operators in phage lambda, *Nature*, 249, 221, 1974.
119. **Heyden, B., Nüsslein, C., and Schaller, H.,** Single RNA polymerase binding site isolated, *Nat. New Biol.* 240, 9, 1972.
120. **Southern, E. M. and Mitchell, A. B.,** Chromatography of nucleic acid digests on thin layers of cellulose impregnated with polyethyleneimine, *Biochem. J.*, 123, 613, 1971.
121. **Gilbert, W., Gralla, J., Majors, J., and Maxam, A.,** Lactose operator sequences and the action of the *lac* repressor, personal communication, 1974.
122. **Gilbert, W., Maizels, N., and Maxam, A.,** Sequences of controlling regions of the lactose operon, *Cold Spring Harbor Symp. Quant. Biol.*, 38, 845, 1973.
123. **Murray, K.,** Nucleotide "maps" of digests of deoxyribonucleic acid, *Biochem. J.*, 118, 831, 1970.
124. **Anfinsen, C. B., Cuatrecasas, P., and Taniuchi, H.,** Staphylococcal nuclease, chemical properties and catalysis, in *The Enzymes*, Vol. IV, 3rd ed., Boyer, P. D., Ed., Academic Press, New York, 1971, 177.
125. **Sadowski, P. D. and Hurwitz, J.,** Enzymatic breakage of deoxyribonucleic acid. II. Purification and properties of endonuclease IV from T4 phage-infected *Escherichia coli*, *J. Biol. Chem.*, 244, 6192, 1969.
126. **Sanger, F.,** New techniques in DNA sequencing, in *Virus Research*, Fox, C. F. and Robinson, W. S., Eds., Academic Press, New York, 1973, 573.
127. **Sadowski, P. D.,** personal communication.
128. **Roychoudhury, R. and Wu, R.,** unpublished results, 1973.
129. **Ling, V.,** The fractionation and sequences of the large pyrimidine oligonucleotides from bacteriophage fd DNA, *J. Mol. Biol.*, 64, 87, 1972.
130. **Ziff, E. B., Sedat, J. W., and Galibert, F.,** Determination of the nucleotide sequence of a fragment of bacteriophage φX174 DNA, *Nat. New Biol.*, 241, 34, 1973.
131. **Galibert, F., Sedat, J. W., and Ziff, E. B.,** Direct determination of DNA nucleotide sequences: structure of a fragment of bacteriophage φX174 DNA, *J. Mol. Biol.*, in press, 1974.
132. **Shapiro, H. S. and Chargaff, E.,** Studies on the nucleotide arrangement in deoxyribonucleic acids. II. Differential analysis of pyrimidine nucleotide distribution as a method of characterization, *Biochim. Biophys. Acta*, 26, 608, 1957.
133. **Burton, K.,** Preparation of apurinic acid and of oligodeoxyribonucleotides with formic acid and diphenylamine, in *Methods in Enzymology*, Vol. 12A, Grossman and Moldave, K., Eds., Academic Press, New York, 1965, 222.
134. **Chargaff, E.,** Isolation and composition of the deoxypentose nucleic acids and of the corresponding nucleoproteins, in *The Nucleic Acids, Chemistry and Biology*, Vol. I, Academic Press, New York 1955, 307.
135. **Half, J. B. and Sinsheimer, R. L.,** The structure of the DNA of bacteriophage φX174. IV. Pyrimidine sequences, *J. Mol. Biol.*, 6, 115, 1963.
136. **Spencer, J. H. and Chargaff, E.,** Studies on the nucleotide arrangement in deoxyribonucleic acids. V. Pyrimidine nucleotide clusters: isolation and characterization, *Biochim. Biophys. Acta*, 68, 9, 1963.
137. **Peterson, G. B. and Reeves, J. M.,** An improved separation of pyrimidine oligonucleotides derived from DNA, *Biochim. Biophys. Acta*, 129, 438, 1966.
138. **Szekely, M. and Sanger, F.,** Use of polynucleotide kinase in fingerprinting non-radioactive nucleic acids, *J. Mol. Biol.*, 43, 607, 1969.
139. **Mushynski, W. E. and Spencer, J. H.,** Nucleotide clusters in deoxyribonucleic acids. VI. The pyrimidine oligonucleotides of strands r and l of bacteriophage lambda DNA, *J. Mol. Biol.*, 51, 107, 1970.
140. **Ling, V.,** Pyrimidine sequences from DNA of bacteriophages fd, f1 and φX174, *Proc. Natl. Acad. Sci. USA*, 69, 742, 1972.
141. **Southern, E. M.,** Base sequence and evolution of guinea-pig α satellite DNA, *Nature*, 227, 794, 1970.
142. **Harbours, K. and Spencer, J. H.,** Nucleotide clusters in deoxyribonucleic acids; pyrimidine oligonucleotides of mouse L-cell satellite deoxyribonucleic acid and main-band deoxyribonucleic acid, *Biochemistry*, 13, 1074, 1974.
143. **Takemura, S.,** Hydrazinolysis of nucleic acids. I. The formation of deoxyriboapyrimidinic acid from herring sperm deoxyribonucleic acid, *Bull. Chem. Soc. Jap.*, 32, 920, 1959.
144. **Habermann, V.,** The degradation of apyrimidinic deoxyribonucleic acid in alkali. A method for the isolation of purine nucleotide sequences from deoxyribonucleic acid, *Biochim. Biophys. Acta*, 55, 999, 1962.
145. **Sedat, J. and Sinsheimer, R. L.,** Structure of the DNA of bacteriophage φX174. V. Purine sequences, *J. Mol Biol.*, 9, 489, 1964.
146. **Turler, H. and Chargaff, E.,** Studies on the nucleotide arrangement in deoxyribonucleic acids. XII. A pyridiminic acid from calf-thymus deoxyribonucleic acid: preparation and properties, *Biochim. Biophys. Acta*, 195, 446, 1969.

147. **Shapiro, H. S.,** The preparation of purine oligonucleotides by hydrazinolysis of DNA, in *Methods in Enzymology,* Vol. 12A, Grossman, L. and Moldave, K., Eds., Academic Press, New York, 1967, 212.
148. **Josse, J., Kaiser, A. D., and Kornberg, A.,** Enzymatic synthesis of deoxyribonucleic acid. VII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid, *J. Biol. Chem.,* 236, 864, 1961.
149. **Penswick, J. R. and Holley, R. W.,** Specific cleavage of the yeast alanine RNA into two large fragments, *Prod. Natl. Acad. Sci. USA,* 53, 543, 1965.
150. **Laskowski, M., Sr.,** DNAses and their use in the studies of primary structure of nucleic acids, in *Advances in Enzymology,* Vol. 29, Nord, F. F., Ed., Interscience, New York, 1967, 165.
151. **Carrara, M. and Bernardi, G.,** Studies on acid deoxyribonuclease. V. The oligonucleotides obtained from deoxyribonucleic acid and their 3'-phosphate termini, *Biochemistry,* 7, 1121, 1968.
152. **Ehrlich, S. D., Torti, G., and Bernardi, G.,** Studies on acid deoxyribonuclease. IX. 5' Hydroxy terminal and penultimate nucleotides of oligonucleotides obtained from calf thymus deoxyribonucleic acid, *Biochemistry,* 10, 2000, 1971.
153. **Laval, J., Thiery, J., Ehrlich, S. D., Paoletti, C., and Bernardi, G.,** Studies on the specificity of an acid deoxyribonuclease for *Helix aspersa* (Müll), *Eur. J. Biochem.,* 40, 133, 1973.
154. **Bernardi, G., Ehrlich, S. D., and Thiery, J. P.,** The specificity of deoxyribonucleases and their use in nucleotide sequence studies, *Nat. New Biol.,* 246, 36, 1973.
155. **Lehman, I. R., Roussos, G. G., and Pratt, E. A.,** The deoxyribonucleases of *Escherichia coli.* II. Purification and properties of a ribonucleic acid-inhibitable endonuclease, *J. Biol. Chem.,* 237, 819, 1962.
156. **Scheffler, J. E., Elson, E. L., and Baldwin, R. L.,** Helix formation by dAT oligomers. I. Hairpin and straight-chain helices, *J. Mol. Biol.,* 36, 291, 1968.
157. **Simon, M., Chang, H. C., and Laskowski, M., Sr.,** Action of pancreatic deoxyribonuclease I on crab d(A-T) polymer, *Biochim. Biophys. Acta,* 232, 462, 1971.
158. **Junowicz, E. and Spencer, J. H.,** Studies on bovine pancreatic deoxyribonuclease A, *Biochem. Biophys. Acta,* 312, 1973, p. 72, 85.
159. **Van Ormondt, H. and Maagdonberg, J. W.,** Primary structure of micrococcal nuclease fragments of microbial DNA, *Biochem. Biophys. Acta,* 272, 141, 1972.
160. **Mikulski, A. J., Sulkowski, E., Stasiuk, L., and Laskowski, M., Sr.,** Susceptibility of dinucleotides bearing either 3'- or 5'-monophosphate to micrococcal nuclease, *J. Biol. Chem.,* 244, 6559, 1969.
161. **Richardson, C. C., Lehman, J. R., and Kornberg, A.,** A deoxyribonucleic acid phosphatase-exonuclease from *Escherichia coli, J. Biol. Chem.,* 239, 251, 1964.
162. **Donelson, J. E. and Wu, R.,** Nucleotide sequence analysis. VII. Characterization of *Escherichia coli* exonuclease III activity for possible use in terminal nucleotide sequence analysis of duplex DNA, *J. Biol. Chem.,* 247, 4661, 1972.
163. **Rabin, E. Z. and Frazer, M. J.,** Isolation of *Neurospora crassa* endonuclease specific for single-stranded DNA, *Can. J. Biochem.,* 48, 389, 1970.
164. **Ando, T.,** A nuclease specific for heat-denatured DNA isolated from a product of *Aspergillus oryzae, Biochem. Biophys. Acta,* 114, 158, 1966.
165. **Sutton, W. D.,** A crude nuclease preparation suitable for use in DNA reassociation experiments, *Biochim. Biophys. Acta,* 240, 522, 1971.
166. **Vogt, V. M.,** Purification and further properties of single-strand-specific nuclease from *Aspergillus oryzae, Eur. J. Biochem.,* 33, 192, 1973.
167. **Johnson, P. H. and Laskowski, M., Sr.,** Mung bean nuclease. I. Resistance of double stranded deoxyribonucleic acid and susceptibility of regions rich in adenosine and thymidine to enzymatic hydrolysis, *J. Biol. Chem.,* 245, 891, 1970.
168. **Ardelt, W. and Laskowski, M., Sr.,** Mung bean nuclease. I. An improved method of preparation., *Biochem. Biophys. Res. Commun.,* 44, 1205, 1971.
169. **Kaplan, J. C., Kushner, S. R., and Grossman, L.,** Enzymatic repair of DNA. I. Purification of two enzymes involved in the excision of thymine dimers from ultraviolet-irradiated DNA, *Proc. Natl. Acad. Sci. USA,* 63, 144, 1969.
170. **Hamilton, R. T. and Wu, R.,** unpublished observation, 1972.
171. **Ghangas, G. S. and Wu, R.,** Specific cleavage of cohesive ends of λ DNA by single-strand specific nucleases, *Fed. Proc.,* 33, 1493, 1974.
172. **Sanger, F., Brownlee, G. G., and Barrell, B. G.,** A two-dimensional fractionation procedure for radioactive nucleotides, *J. Mol. Biol.,* 13, 373, 1965.
173. **Sanger, F. and Brownlee, G. G.,** A two-dimensional fractionation method for radioactive nucleotides, in *Methods in Enzymology,* Vol. 12A, Grossman, L. and Moldave, K., Eds., Academic Press, New York, 1967, 361.
174. **Markham, R. and Smith, J. D.,** Structure of ribonucleic acid, *Nature,* 168, 406, 1951.
175. **Smith, J. D.,** Paper electrophoresis of nucleic acid components, in *Methods in Enzymology,* Vol. 12A, Grossman, L. and Moldave, K., Eds., Academic Press, New York, 1967, 350.
176. **Murray, K.,** Nucleotide sequence analysis with polynucleotide kinase and nucleotide mapping methods; 5'-terminal sequences of DNA from bacteriophage λ and 424, *Biochem. J.,* 131, 569, 1973.
177. **Morrison, A. and Murray, K.,** The behavior of oligodeoxynucleotides on thin-layer chromatography on polyethyleneimine-cellulose and ion-exchange paper electrophoresis. Applications in fractionating and sequencing terminally labeled oligodeoxynucleotides, *Biochem. J.,* 141, 321, 1974.

178. **Randerath, K. and Randerath, E.,** Thin-layer separation methods for nucleic acid derivatives, in *Methods in Enzymology*, Vol. 12A, Grossman, L. and Moldave, K., Eds., Academic Press, New York, 1967, 323.
179. **Griffin, B. E.,** Separation of ^{32}P-labeled ribonucleic acid components. The use of polyethyleneimine-cellulose (TLC) as a second dimension in separating oligoribonucleotides of 4.5S and 5S from *E. coli*, *FEBS Lett.*, 15, 165, 1971.
180. **Brownlee, G. G. and Sanger, F.,** Chromatography of ^{32}P-labeled oligonucleotides on thin layers of DEAE-cellulose, *Eur. J. Biochem.*, 11, 395, 1969.
181. **Rensing, F. E. and Shoenmakers, J.,** A sequence of 50 nucleotides from coli phage R17 RNA, *Eur. J. Biochem.*, 33, 8, 1973.
182. **Yoneda, M. and Bollum, F. J.,** Deoxynucleotide-polymerizing enzymes of calf thymus gland. I. Large scale purification of terminal and replicative deoxynucleotidyl transferase, *J. Biol. Chem.*, 240, 3385, 1965.
183. **Roychoudhury, R., Fischer, D., and Kössel, H.,** A new method for the sequence analysis of oligodeoxynucleotides, *Biochem. Biophys. Res. Commun.*, 45, 430, 1971.
184. **Bambara, R., Jay, E., and Wu, R.,** DNA sequence analysis: a formula to predict electrophoretic mobility of oligonucleotides on cellulose acetate, *Nucleic Acids Res.*, submitted, 1974.
185. **Van Holde, K. E.,** *Physical Biochemistry*, Prentice Hall, New Jersey, 1971.
186. **Tanford, C.,** *Physical Chemistry of Macromolecules*, John Wiley & Sons, New York, 1961.
187. **Tanford, C., Kawahara, K., and Lapanje, S. J.,** Proteins as random coils. I. Intrinsic viscosities and sedimentation coefficients in concentrated guanine hydrochloride, *J. Am. Chem. Soc.*, 89, 729, 1967.
188. **Lebowitz, P., Weissman, S. M., and Radding, C. M.,** Nucleotide sequence of a ribonucleic acid transcribed *in vitro* from λ phage deoxyribonucleic acid, *J. Biol. Chem.*, 246, 5120, 1971.
189. **Blattner, F. R. and Dahlberg, J. E.,** RNA synthesis startpoints in bacteriophage λ: are the promoter and operator transcribed? *Nat. New Biol.*, 237, 227, 1972.
190. **Lozeron, W., Szybalsky, W., and Dahlberg, J. E.,** cited by Salser, Reference 3.
191. **Zain, B. S., Weissman, S. M., Dhar, R., and Pan, J.,** The nucleotide sequence preceding a RNA polymerase initiation site on SV40 DNA. Part 1. The sequence of the late strand transcript, *Nucleic Acids Res.*, 1, 577, 1974.
192. **Dhar, R., Weissman, S. M., Zain, B. S., Pan, J., and Lewis, A. M., Jr.,** The nucleotide sequence preceding RNA polymerase initiation site on SV40 DNA. Part 2. The sequence of the early strand transcript, *Nucleic Acids Res.*, 1, 595, 1974.
193. **Musso, R., de Crombrugghe B., Pastan, I., Sklar, J., Yot, P., and Weissman, S.,** The 5′ terminal sequence of *E. coli* Gal mRNA, *Fed. Proc.*, 33, 1601, 1974; manuscript submitted to *Proc. Natl. Acad. Sci. USA*, 1974.
194. **Downey, K. M., Jurmark, B. S., and So, A. G.,** Determination of nucleotide sequences at promoter regions by the use of dinucleotides, *Biochemistry*, 10, 4970, 1971.
195. **Hoffman, D. J. and Niyogi, S. K.,** RNA initiation with dinucleotide monophosphates during transcription of bacteriophage T4 DNA with RNA polymerase of *Escherichia coli*, *Proc. Natl. Acad. Sci. USA*, 70, 574, 1973.
196. **Downey, K. M. and So, A. G.,** Studies on the kinetics of ribonucleic acid chain initiation and elongation, *Biochemistry*, 9, 2520, 1970.
197. **Kleppe, R. and Khorana, H. G.,** Transcription of short double-stranded deoxyribonucleic acids of defined nucleotide sequences, *J. Biol. Chem.*, 247, 6149, 1972.
198. **Terao, T., Dahlberg, J. E., and Khorana, H. G.,** On the transcription of a synthetic 29-unit long deoxyribopolynucleotide, *J. Biol. Chem.*, 247, 6157, 1972.
199. **Salser, W., Poon, R., Whitcome, P., and Fry, K.,** New techniques for determining nucleotide sequences from eucaryotic cells, in *Virus Research*, Fox, C. F. and Robinson, W. S., Eds., Academic Press, New York, 1973, 545.
200. **Poon, R. and Salser, W.,** personal communication.
201. **Marotta, C. A., Forget, B. G., Weissman, S. M., Verma, I. M., McCaffrey, R. P., and Baltimore, D.,** Nucleotide sequences of human globin messenger RNA, *Proc. Natl. Acad. Sci. USA*, 71, 2300, 1974.
202. **Cheng, T. Y. and Tso, P. O. P.,** Enzymatic synthesis of ribo-deoxyribopolynucleotides by the RNA polymerase, *Fed. Proc.*, 24, 602, 1965.
203. **Chamberlin, M.,** rAU C polymer, in *Procedures in Nucleic Acid Research*, Cantoni, C. L. and Davies, D. R., Eds., Harper and Row, New York, 1966, 513.
204. **Hurwitz, J., Yarbrough, L., and Wickner, S.,** Utilization of deoxynucleotide triphosphates by DNA-dependent RNA polymerase of *E. coli*, *Biochem. Biophys. Res. Commun.*, 48, 628, 1972.
205. **Paddock, G. V., Heindell, H. C., and Salser, W.,** Deoxysubstitution in RNA by RNA polymerase *in vitro*; a new approach to nucleotide sequence determinations, *Proc. Natl. Acad. Sci. USA*, in press, 1974.
206. **Poon, R., Paddock, G. V., Heindell, H., Whitcome, P., Salser, W., Kacian, D., Bank, A., Gambino, R., and Ramirez, F.,** Nucleotide sequence analysis of RNA synthesized from rabbit globin complementary DNA, *Proc. Natl. Acad. Sci. USA*, in press, 1974.
207. **Van de Voorde, A., Rogiers, R., Van Herrewerghe, J., Van Heuverswyn, H., Volckaert, G., and Fiers, W.,** Deoxynucleotide substitution: a new technique for sequence analysis of RNA, *Nucleic Acids Res.*, in press, 1974.
208. **Cohen, P. T., Yaniv, M., and Yanofsky, C.,** Nucleotide sequences from messenger RNA transcribed from the tryptophan operon of *E. coli*, *J. Mol. Biol.*, 74, 163, 1973.
209. **Bronson, M. J. and Yanofsky, C.,** Characterization of mutations in the tryptophan operon of *E. coli* by RNA sequencing, *J. Mol. Biol.*, submitted, 1974.

210. **Barnes, W. M., Seigel, R. B., and Resnikoff, W. S.,** The construction of λ transducing phages containing deletions defining regulatory elements of the lac and trp operons in *E. coli, Mol. Gen. Genet.*, 129, 201, 1974.
211. **Dickson, R., Abelson, J., Barnes, W. M., and Resnikoff, W. S.,** *Science*, in press.
212. **Yoneda, M. and Bollum, F. J.,** Deoxynucleotide polymerizing enzymes of calf thymus gland. I. Large scale purification of terminal and replicative deoxynucleotidyl transferases, *J. Biol. Chem.*, 240, 3385, 1965.
213. **Roychoudhury, R. and Kössel, H.,** Synthetic polynucleotides: enzymatic synthesis of ribonucleotide terminated oligodeoxynucleotides and their use as primers for the enzymatic synthesis of polydeoxynucleotides, *Eur. J. Biochem.*, 22, 310, 1971.
214. **Lehman, I. R. and Nussbaum, A.,** The deoxyribonucleases of *Escherichia coli.* V. On the specificity of exonuclease I (phosphodiesterase), *J. Biol. Chem.*, 239, 2628, 1964.
215. **Weiss, B. and Richardson, C. C.,** The 5′-terminal dinucleotides of the separated strands of T7 bacteriophage deoxyribonucleic acid, *J. Mol. Biol.*, 23, 405, 1967.
216. **Englund, P. T.,** The 3′-terminal nucleotide sequences of T7 DNA, *J. Mol. Biol.*, 66, 209, 1972.
217. **Donelson, J. E. and Wu, R.,** Nucleotide sequence analysis of deoxyribonucleic acid. VI. Determination of 3′-terminal dinucleotide sequences of several species of duplex deoxyribonucleic acid using *E. coli* deoxyribonucleic acid polymerase I, *J. Biol. Chem.*, 247, 4654, 1972.
218. **Van de Sande, J. H., Kleppe, K., and Khorana, H. G.,** Reversal of bacteriophage T4 induced polynucleotide kinase action, *Biochemistry*, 12, 5056, 1973.
219. **Van de Sande, J. H.,** personal communication.
220. **Bambara, R.,** unpublished results.
221. **Ehrlich, S. D., Thiery, J.-P., Devillers-Thiery, A., and Bernardi, G.,** A new approach to the study of nucleotide sequences in DNAs, *Nucleic Acids Res.*, 1, 87, 1974.
222. **Bernardi, G.,** personal communication.
223. **Richards, G. M. and Laskowski, M., Sr.,** Negative charge at the 3′-terminus of oligonucleotides and resistance to venom exonuclease, *Biochemistry*, 8, 1789, 1969.
224. **Richards, G. M., and Laskowski, M., Sr.,** Use of venom exonuclease at low pH for preparation of mononucleoside diphosphates, *Biochemistry*, 8, 4858, 1969.
225. **Ascoli, F. and Kahan, F. M.,** Iodination of nucleic acids in organic solvents with iodine monochloride, *J. Biol. Chem.*, 241, 428, 1966.
226. **Brammer, K. W.,** Chemical modification of viral RNA. II. Bromination and iodination, *Biochim. Biophys. Acta*, 72, 217, 1963.
227. **Commerford, S. L.,** Iodination of nucleic acids *in vitro*, *Biochemistry*, 10, 1993, 1971.
228. **Prensky, W., Steffensen, D. M., and Hughes, W. L.,** The use of iodinated RNA for gene localization, *Proc. Natl. Acad. Sci. USA*, 70, 1860, 1973.
229. **Robertson, H. D., Dickson, E., Model, P., and Prensky, W.,** Application of fingerprinting techniques to iodinated nucleic acids, *Proc. Natl. Acad. Sci. USA*, 70, 3260, 1973.
230. **Dale, R. M. K., Livingston, D. C., and Ward, D. C.,** The synthesis and enzymatic polymerization of nucleotides containing mercury: potential tools for nucleic acid sequencing and structural analysis, *Proc. Natl. Acad. Sci. USA*, 70, 2238, 1973.
231. **Dale, R. M. K., Martin, E., Livingston, D. C., and Ward, D. C.,** personal communication.
232. **Whitcome, P., Fry, K., and Salser, W.** The use of ribosubstitution techniques for determining DNA sequences, in *Methods in Enzymology*, Grossman, L. and Moldave, K., Eds., Academic Press, New York, 1974, 295.
233. **Salser, W., Fry, K., Wesley, M., and Simpson L.,** Use of nucleic acid fingerprints to estimate the complexity of minicircle DNA, *Biochim. Biophys. Acta*, 319, 277, 1973.
234. **Fry, K., Poon, R., Whitcome, P., Idress, J., Salser, W., Mazrimas, J., and Hatch, F.,** Nucleotide sequence of HS-β satellite DNA from kangaroo rat *dipodomys ordii, Proc. Natl. Acad. Sci. USA*, 70, 2642, 1973.
235. **Skinner, D. M., Beattie, W. G., Blattner, F. R., Stark, B. P., and Dahlberg, J. E.,** The repeat sequence of a hermit crab satellite deoxyribonucleic acid is (-T-A-G-G)n-(-A-T-C-C-)n, *Biochemistry*, 13, 3930, 1974.
236. **Gall, J. G., Cohen, E. H., and Atherton, D. D.,** The satellite DNAs of *drosophila virilis, Cold Spring Harbor Symp.*, 38, 417, 1973.
237. **Proudfoot, N. J. and Brownlee. G. G.,** Nucleotide sequence adjacent to polyadenylic acid in globin messenger RNA, *FEBS Lett.*, 38, 179, 1974.
238. **Besmer, P., Miller, R. C. Jr, Caruthers, M. H., Kumar, A., Minamoto, K., van de Sande, J. H., Sidarova, N. and Khorana, H. G.,** Studies on Polynucleotides, CXVII. Hybridization of Polydeoxynucleotides with Tyrosine Transfer RNA Sequences to the r-Strand of ϕ80psu$^{+}_{m}$ DNA, *J. Mol. Biol.*, 72, 503, 1972.