

Random Subcloning of Sonicated DNA: Application to Shotgun DNA Sequence Analysis¹

PRESCOTT L. DEININGER²

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

Received September 7, 1982

A method for producing random subclones using sonication to fragment the DNA is presented. The sonication is combined with enzymatic repair of the fragment ends and a rigorous size fractionation step to prepare subclones of relatively homogeneous and specific size. Under some conditions sonication is shown to shear A + T-rich sequences preferentially, although under most conditions it will create a random subclone library. The use of these subclone libraries for an improved "shotgun" DNA sequencing strategy is tested on a 17.2-kb (kilobase) fragment of Epstein-Barr virus.

Random subcloning of a large DNA fragment is an essential step in the application of several rapid, shotgun DNA sequencing techniques (1-3), as well as in potential studies of DNA structure and function. In most previous studies, subcloning has been carried out using restriction enzyme digestions to fragment the DNA. There are three major disadvantages to the use of restriction enzymes in this way. The first is that some regions of DNA sequence have very few restriction sites and would be underrepresented in the resulting clone banks. Second, it is necessary to make clone banks using a number of different restriction enzymes to obtain fragments which overlap each other properly to complete the DNA sequence analysis. The third disadvantage is that many fragments produced by restriction digestion are quite small. Sequencing these small clones produces very little information for the work carried out.

Many of these difficulties have been overcome recently by fragmentation of DNA with DNase I in the presence of Mn^{2+} , which

cleaves with very little sequence specificity (4). This cloning procedure allowed the complete sequence analysis of a 4257-bp DNA fragment from a single subclone library.

I have explored the use of physical shearing of DNA to provide an alternate and simpler means of creating such random DNA libraries. Sonication was chosen as the best fragmentation method for several reasons. It can be carried out in small volumes and with small amounts of DNA with high reproducibility. Sonication is also capable of producing a very good size distribution for DNA sequence analysis, leaving only short overhanging ends after the shearing process (5).

These ends are repaired easily as they contain predominantly 5'-phosphate and 3'-hydroxyl groups (6,7), both of which make excellent substrates for many repair enzymes. To demonstrate the effectiveness of this cloning protocol, clone libraries were constructed to sequence the 17.2-kb *EcoRI*-C fragment of Epstein-Barr Virus (8).

MATERIALS AND METHODS

DNA preparation and sonication. The cloned *EcoRI*-C fragment of Epstein-Barr Virus (8) was grown and prepared from chlor-

¹ Supported by a Fellowship from the North Atlantic Treaty Organization.

² Present address: Department of Biochemistry, LSU Medical Center, 1901 Perdido Street, New Orleans, La. 70112.

amphenicol-amplified *Escherichia coli* HB101 as previously described (9,10). The *Eco*RI-C fragment was cleaved from the vector by digestion with *Eco*RI and then separated and isolated from the vector by agarose gel electrophoresis in an 0.8% low-gel-temperature agarose (Bethesda Research Laboratories) gel. The DNA band was excised and the fragment was eluted by melting the gel at 65°C; the removal of the agarose by three successive phenol extractions was followed by ether extraction and ethanol precipitation (11).

To avoid any potentially nonrandom shearing effects near the ends of molecules, the 4 μ g of fragment was concatenated and circularized by incubation for 2 h in 25 μ l of ligation buffer (66 mM Tris, pH 7.6, 6.6 mM MgCl_2 , 10 mM DTT,³ and 1 mM rATP) with dilute T₄ DNA ligase (T₄ DNA ligase was a gift of S. Fields and D. Bentley). This solution was then brought up to 250 μ l in a 1.5-ml Eppendorf tube with 100 mM Tris (pH 7.4), 10 mM EDTA in preparation for sonication.

Sonication was carried out with the microtip of an MSE sonicator, code 1-71. Samples were kept in ice water and the sonicator probe inserted as far as possible into the solution without contacting the sides of the tube. An amplitude of 5 μ m with bursts of 5 s spaced by 15 s cooling was used in most experiments. The number of bursts is described for each experiment in the text, although four bursts are now used routinely. It should be noted that although sonication is effective on supercoiled DNAs, they require about twice as many bursts to give the same size distribution as relaxed DNA molecules. The DNA concentration was typically 16 μ g/ml, but concentrations between 4 and 100 μ g/ml gave similar results. DNA samples were ethanol precipitated before the end-repair reactions.

The *Micrococcus lysodeikticus* and *Clostridium perfringens* DNAs (Sigma) used in the experiment described in Fig. 1 were resuspended in 10 mM Tris, pH 7.4, 1 mM EDTA,

phenol extracted once, and ether extracted before sonication.

Repair of fragmented DNA ends. The fragment ends were repaired with T₄ DNA polymerase by a modification of the conditions of Challberg and Englund (12). The fragments were suspended in 25 μ l of a solution containing 67 mM Tris, pH 8.0, 6.7 mM MgCl_2 , 10 mM 2-mercaptoethanol, and 25 μ M of dGTP, dCTP, and dTTP. To incorporate radiolabel during the repair process, this solution was used to dissolve 5 μ Ci of [α -³²P]dATP (400 Ci/mmol), incubated with 10 units of T₄ DNA polymerase (P-L Biochemicals) at 15°C for 2 h, and then followed by a final chase for 30 min with 100 μ M of each of the four deoxynucleoside triphosphates. The reaction was terminated by chilling on ice and phenol extraction. The DNA was ethanol precipitated prior to size fractionation or ligation. Alternatively, if radiolabel was not used (Fig. 2), the repair was carried out for the indicated times with 25 μ M concentrations of the four cold deoxynucleoside triphosphates.

Size fractionation. Size fractionations were carried out on either 1% low-gel-temperature agarose (Bethesda Research Laboratories) or 1.5% high-gel-temperature agarose in 1X TBE (90 mM Tris, 90 mM boric acid, 3 mM EDTA, pH 8.3) containing 0.5 μ g/ml ethidium bromide in a minigel apparatus (Uniscience). The gels were loaded almost to the point of overloading, 4 μ g of sonicated DNA in a well 0.9-cm wide \times 0.3-cm deep and electrophoresed into the gel until a bromophenol blue marker had electrophoresed 1 to 2 cm. This short electrophoresis allowed a reasonable fractionation without diluting the DNA through a large volume of agarose. The DNA was visualized using uv irradiation and ethidium bromide fluorescence. For the elution of DNA fragments from high-gel-temperature agarose, a trough elution procedure was used (13). A 2-mm-wide trough was cut across the DNA track at about 350 nucleotides. The trough was filled with running buffer and the gel run for a short period. The buffer, containing eluted DNA, was removed from the trough

³ Abbreviations used: DTT, dithiothreitol; PEG, polyethyleneglycol 6000.

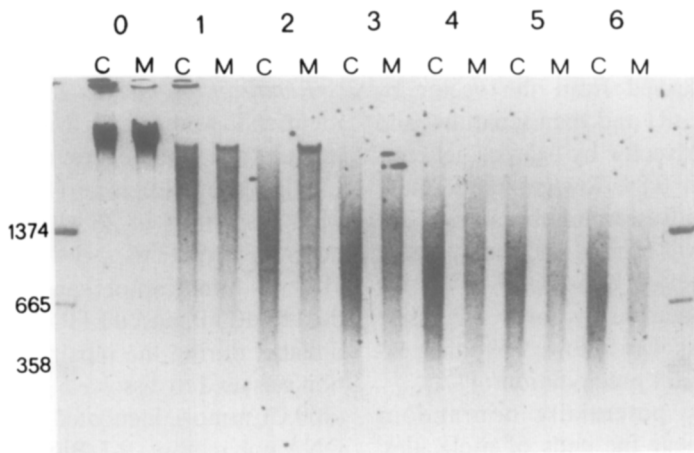


FIG. 1. The effects of a time course of sonication on DNAs of differing base compositions. Sonication was carried out on *M. lysodeikticus* DNA (M), which has a 72.6% G + C content, and *Cl. perfringens* DNA (C) with a G + C content of 30.9%. Sonications were at an amplitude of 5 μ m in 5-s bursts and an aliquot was removed from the mix for electrophoresis on a 1.5% agarose gel after each burst. The samples were run in pairs with the number of bursts indicated over each pair. The two outside tracks represent PBR322 cleaved by *Sau*3A as a marker.

and the process repeated until all the desired fractions were collected. The time of each elution step into the trough should be carefully calibrated so as to avoid the appearance of DNA on the far side of the trough after elution, as observed by the ethidium bromide fluorescence. The DNA fractions were then pooled into fractions 100 to 200 nucleotides broad and ethanol precipitated. Using the low-temperature gel, size fractions were excised from the gel and the DNA eluted as described previously (11).

Cloning and sequencing. Blunt-ended DNA fragments were cloned by ligation directly into *Hinc*II (a gift of G. Winter) and calf alkaline phosphatase (Boehringer-Mannheim)-treated M13 mp7 replicative form DNA as previously described (14), or by ligation in the same manner into M13 mp8 (a gift of Dr. J. Messing) replicative form which had been digested with *Sma*I (New England Biolabs) to produce a blunt-end vector and phosphatase treated to stop religation of the vector and to reduce background plaques. A typical ligation reaction contained 10 ng of blunt-end vector, 5 to 10% of the DNA from a size fraction (estimated to be 10–20 ng of DNA), ligation

buffer, and T_4 DNA ligase. Ligations were carried out in a 10- μ l volume in sealed capillary tubes overnight at 15°C. Using size-fractionated DNA, the DNA concentration was only estimated. In this case, ligations were carried out with various concentrations of insert, es-

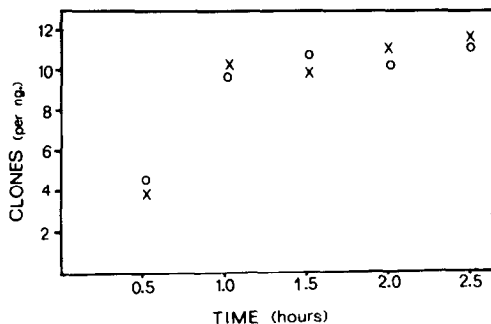


FIG. 2. Time course of repair of sonicated ends with T_4 DNA polymerase. Four micrograms of DNA were sonicated for four 5-s bursts at 5 μ m amplitude. One-half microgram aliquots were removed from a repair reaction every 30 min, phenol extracted, ether extracted, and ethanol precipitated. Ten nanograms of each of these aliquots was then ligated to 10 ng of M13 mp8 blunt-end cloning vector, transformed into *E. coli* JM101, and the white plaques containing inserts counted. The two symbols (O,X) represent two separate repair reactions on the same sonicated DNA sample.

timated from 0.5 to 10 ng. The lowest concentration of insert DNA which yields a reasonable number of clones (10–30) should then be used in further ligations to create the subclone library. The transformation and selection of clones in the host strain JM101 have been described in detail previously (1,2); the strain JM103 was handled in the same way. The clones were sequenced by the method of Sanger (1,2,15) with the modification of Messing (2), which allows the sequence analysis to be carried out in Eppendorf tubes instead of capillaries.

RESULTS

Fragmentation of DNA

The DNA fragment size produced by sonication is influenced by several factors, primarily the amplitude, the time of sonication, and certain components of the DNA solution (6). The initial fragmentation of DNA occurs quite rapidly, but the smaller fragments produced are then resistant to further shearing. The actual size of the DNA, when it becomes resistant to fragmentation, varies with the amplitude used. Thus, if smaller DNA fragments are desired, a higher amplitude will produce them more quickly. A typical time course of sonication is shown in Fig. 1. After about 15 s of sonication at an amplitude of 5 μ m, all the DNA has been fragmented to some extent and further sonication only slowly lowers the resulting fragment size distribution. At this amplitude, sonication for between 15 and 45 s will produce a fragment distribution containing a large proportion of 300- to 1000-nucleotide-long DNA fragments. At different amplitude settings the shearing rate can be quite different, but this can be calibrated rapidly on any DNA sample. Figure 1 shows sonication of DNAs of widely differing G + C contents, *Cl. perfringens* DNA (30.9% G + C) and *M. lysodeiketicus* (72.6% G + C), as a function of time. There is an apparent difference in initial rate of shearing of these two DNA samples. The A + T-rich DNA reproducibly shears more rapidly, as can be seen

particularly well in the 10-s time point. This difference disappears after 15 or 20 s when the distributions are essentially indistinguishable between the different DNAs.

Repair of Sonicated Ends

The DNA fragments produced from sonication do not contain blunt ends capable of direct ligation and attempts to directly clone this material have produced no clones containing inserted DNA. T₄ DNA polymerase was used to repair the ends of the sonicated DNA. Essentially, the procedure of Englund *et al.* (12) for end-labeling with T₄ DNA polymerase was used with this enzyme, as described under Materials and Methods. In order to study the kinetics of this repair, aliquots of DNA were taken out of a repair reaction every 30 min, phenol extracted, ether extracted, and ethanol precipitated. This material was cloned directly into M13 mp7 or mp8 using blunt-end ligation. Figure 2 shows the efficiency of cloning as measured by ligations using 10 ng of sonicated DNA and 10 ng of the blunt-end M13 vector. The cloning efficiency rose to about 12 clones per nanogram of insert DNA (Fig. 2). This compares favorably with about 300 clones obtained for the same mass of *AluI* cleaved λ -phage DNA under these same conditions. The repair also seemed to be complete in 1 h, although longer incubation times may have resulted in a slight improvement.

Size Fractionation of Repaired, Sonicated DNA

When DNA having a size distribution essentially that of the 20-s sonication in Fig. 1 is cloned directly after repair, inserts of varying sizes are produced. About half of the clones are rather small (between 100 and 200 nucleotides). For many projects it will be desirable to remove these smaller fragments prior to insertion into the cloning vector. To obtain clones which will be used for sequence analysis, an agarose gel electrophoresis fractionation step is routinely used. Agarose impu-

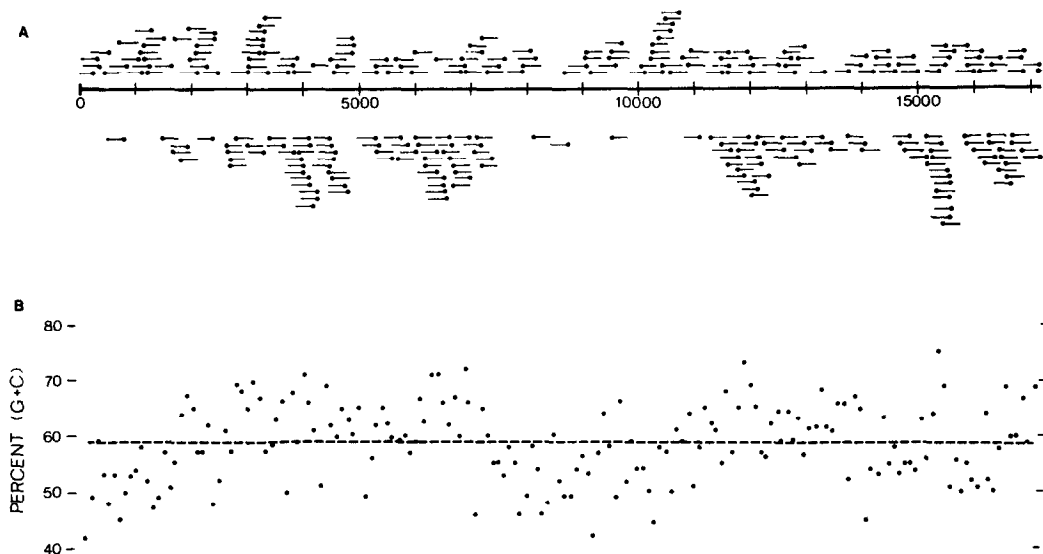


FIG. 3. (A) The distribution of subclones in the EBV and *EcoRI*-C fragment. The axis represents the nucleotide map of the total restriction fragment in base pairs. There are two separate groups of clones which were prepared independently. The group I clones, below the axis, show a nonrandom distribution and the group II clones, above the axis, are distributed more randomly. The group I clones were produced from DNA having an equivalent treatment to the two-burst sonication in Fig. 1 and group II from DNA sonicated for four bursts. (B) The G + C distribution in the EBV *EcoRI* fragment. The percentage G + C is presented for every 100 bases of the percentage of the overall fragment. Note the strong correlation between the distribution of the group I clones in (A) and high G + C content.

rities seem to inhibit the ligation somewhat, but from 4 μ g of initial DNA, 500 to 2000 clones in the range of 400 to 1000 nucleotides in length can be obtained. These longer clones also minimize the probability of sequencing across a religation junction between two fragments. In this study, sequence analysis has shown two religation events out of 300 clones in spite of the size fractionation. These created no difficulty, however, as in the random cloning strategy almost every region of the genome is covered by more than one clone and such religations are located in the computer analysis (16).

The Distribution of Clones within the EBV EcoRI-C Fragment

The distribution of the group I of 150 clones derived from the EBV *EcoRI*-C fragment is represented schematically as the lower group of lines in Fig. 3A. As can be seen, the distribution of these clones was markedly non-

random. Group II, a second clone bank produced using twofold longer sonication, proved to be much more random (top set of lines in Fig. 3A). Analysis of the G + C distribution as a function of position in the fragment as predicted by the final sequence revealed that the clones in group I were being preferentially produced from the G + C rich regions of DNA. The sonication conditions used in the group I cloning were similar to the 10-s sonication in Fig. 2, which shows a marked bias in fragmentation of DNAs of different base composition. The group II clone library, which appears to have a more random distribution, was produced from DNA after 20 s of sonication when there was little G + C bias (Fig. 1). Both groups of clones were prepared from size fractions between 400 and 600 nucleotides long. It might improve the randomness of the library to use a broader size fraction to compensate for any minor nonrandomness in the DNA size distribution.

The 300 clones in groups I and II were enough to determine the sequence of all but 70 bases of the 17.2-kb fragment. Also, 85% of the DNA sequence was determined on both strands using this totally random approach. The last 70 bases of sequence were actually determined using a nonrandom approach. Basically, a restriction map of the whole fragment was predicted from the almost completed sequence. From this map, a *Pst*I restriction fragment which would span the undetermined sequence was located. This fragment was isolated by agarose gel electrophoresis, cloned, and the sequence determined from the *Pst*I site across the gap in the sequence at about base 2500 (Fig. 3). This final 70 bases of sequence was found to contain an *Eco*K restriction site. Since the majority of clones were created in a host strain, JM 101, which is R_K^+ , it is not surprising that this fragment was not cloned readily.

DISCUSSION

There are a number of requirements to produce an optimal subclone bank for DNA sequence analysis by the shotgun strategy. The most important is that the clones are produced randomly from a fragment so that every region of the DNA is readily cloned into the subclone bank. Even a few small regions which do not clone well, for one reason or another, can greatly increase the work required to complete the sequence. A second important factor is that the clones preferably should be within a fairly narrow size range from 400 to about 1000 nucleotides. This is the optimal range for several reasons. Clones smaller than 400 nucleotides are not desired since a single sequencing experiment can readily produce over 300 nucleotides and it is possible to extend a sequence to over 500 nucleotides. Thus, if clones include appreciably less than this, fewer data are produced from each experiment. Small fragments are also a disadvantage because it is possible to clone two fragments in the same M13 recombinant. If the insert fragments are small, it

will be possible to sequence entirely through one into the other, producing an artificial join in the DNA sequence. A maximum fragment size of 1000 bases is also mentioned because this is well beyond the ability to sequence in a single experiment. Fragments produced much longer than this would decrease the clone yield by obtaining less fragments from a given mass of DNA and because longer fragments do not clone as well in M13 (2).

The subcloning protocol described in this paper has a number of advantages which make it ideal for shotgun sequence analysis. Sonication readily produces fragments with the approximate size distribution described above. At higher amplitudes even a large number of clones as small as 50 bases can be obtained for other experiments. A simple size-fractionation step can select for the exact size fraction desired. Furthermore, fragmentation by sonication gives a reproducible size distribution which is fairly independent of DNA concentrations from 4 to 100 μ g/ml. Satisfactory results have been obtained using sonication to subclone nanogram quantities of influenza virus cDNA (G. Winter and S. Fields, personal communication). This protocol is also very fast and an isolated DNA fragment can be prepared for ligation into the vector in a single day.

The cloning efficiency is quite good using this procedure, although it is difficult to judge quantitatively. Without the size fractionation, cloning of λ DNA by this procedure results in over one-third as many clones as the cloning of restriction fragments resulting from digestion with *Alu*I (data not shown). Therefore, a large portion of the fragments has been repaired and the repair may be almost complete, since the relative proportion of fragments when sonicated versus the restriction digestion is not known in this case. It has been reported that approximately 90–95% of all sonicated DNA ends contain a 5'-phosphate (6,7); therefore, it should be possible to repair as much as 80–90% of the fragments at both ends.

In the process of developing this protocol,

a number of other shearing and repair procedures were attempted, none of which worked as well as the protocol presented. Shearing by high-speed stirring was used (17), but it was difficult to obtain small enough fragments in a small volume. This sheared DNA could be repaired and cloned with reasonable efficiency. Two other methods of end repair were also explored extensively. The large fragment of DNA polymerase I repairs the ends reasonably well, but gave yields 4- to 10-fold lower than when T_4 DNA polymerase was used. Nuclease *Bal31* digestion was also effective to repair sheared fragment ends for ligation. However, this enzyme has rapid, processive 3' and 5' double-stranded exonuclease activities (18) which made it difficult to control on differing DNA samples. Thus, degradation of the DNA was sometimes a problem and this method was not continued.

The rigorous size-fractionation step used in this subcloning protocol is an important contribution to its effectiveness for DNA sequence analysis. This size fractionation does decrease the cloning efficiency. To a large extent this is due to the removal of the small fragments which would clone very efficiently in M13. However, contamination with impurities from the agarose gel and loss of material during elution also causes a loss in overall cloning efficiency. Even with this loss, it is still possible to prepare several hundred clones of this size range starting from a microgram of DNA. In cases where the clone yield must be higher than this, there are alternatives to agarose gel fractionation. One such approach would be to use polyethyleneglycol 6000 (PEG) precipitation (19). In the presence of 7% PEG and 0.5 M NaCl, only the larger DNA fragments will precipitate. These conditions yield essentially no clones under 300 nucleotides (G. Winter, personal communication). Other PEG conditions may also be used to provide other size fractions of DNA.

The size fractionation used in this study was not 100% effective in eliminating all small fragments, although only two clones out of

300 sequenced contained religations of two separate insert fragments as detected by sequence analysis. These religations were readily detected because the random sequencing approach allows most of the DNA to be sequenced several times. Thus, religation events are located during the computer analysis which puts the random fragments of DNA sequence together (6). Also, a careful calibration of insert concentration so that the ratio of vector DNA to insert is maximized as described under Materials and Methods will decrease the probability of such religation events.

The clone banks used in this study demonstrate that sonication will produce essentially random sequence libraries, but that some care should be taken to avoid a potential G + C bias (Figs. 1 and 3). The reason for this possible bias is not completely clear. It appears that A + T-rich DNA sonicates more quickly than G + C-rich DNA, so that under certain conditions different size fractions can be relatively enriched for either G + C- or A + T-rich DNA. However, most sonication conditions do not appear to produce this bias and random libraries can be readily produced. The cloned banks in Fig. 3 were produced from a narrow size range of 400 to 600 nucleotides. This narrow size range may also have helped select for a biased population of fragments. It should be noted that even though the group I clones showed a markedly non-random distribution, over 17 kb were sequenced using only 300 clones. Since it is possible to sequence and handle the data from over 30 clones a week, sequence data accumulates very rapidly.

The random DNA sequence approach used here and elsewhere (1-4) has the potential disadvantage that, although data accumulate rapidly at first, eventually most of the data produced only confirms previously determined sequences. This duplication of data is ordinarily quite useful, as in most sequencing studies it is desirable to sequence both strands of the DNA. Also it is advisable to determine the sequence several times to completely eliminate the possibility of sequencing through

multiple fragments that were religated together. There are several approaches to finishing a DNA sequencing project in a non-random way which can overcome the slow data accumulation toward the end. The one used to complete the sequence of this fragment was to predict a restriction map of the DNA from the almost-completed sequence and to use this map to isolate a fragment containing the desired region of DNA. Alternatively, the final clones may be selected from a large clone bank using previously sequenced M13 clones as probes for overlapping or opposite strand clones (F. Sanger, personal communication). The size fractionation used in creating these clones can also be put to advantage in completing or confirming portions of the sequence. First, the clones will be large enough that often they will not have been completely sequenced in the first experiment. Therefore, further efforts to extend the sequence obtained from specific clones can be useful. Second, since the approximate size of any insert is known, the position of the far end of the clone can be predicted to help determine whether sequencing from the other end of the clone by one of several methods (19,20) is likely to produce sequence in the region desired.

ACKNOWLEDGMENTS

I would like to thank Dr. F. Sanger and Dr. B. Barrell for space in their laboratories at the MRC, Cambridge, England, and for discussions during the course of this work. I would also like to thank Dr. G. Winter and Dr. S. Anderson for critical discussions and A. Bankier for his skilled technical assistance and constant criticism.

REFERENCES

1. Sanger, F., Coulson, A. R., Barrell, B. G., Smith, A. J. H., and Roe, B. A. (1980) *J. Mol. Biol.* **143**, 161-178.
2. Messing, J., Crea, R., and Seeburg, P. H. (1981) *Nucl. Acids Res.* **9**, 309-321.
3. Ruther, U., Koewen, M., Otto, K., and Muller-Hill, B. (1981) *Nucl. Acids Res.* **9**, 4087-4098.
4. Anderson, S. (1981) *Nucl. Acids Res.* **9**, 3015-3027.
5. Pyeritz, R. E., Schlegel, R. A., and Thomas, C. A., Jr. (1972) *Biochim. Biophys. Acta* **272**, 504-509.
6. Richards, O. C., and Boyer, P. D. (1965) *J. Mol. Biol.* **11**, 327-340.
7. Richardson, C. C. (1966) *J. Mol. Biol.* **15**, 49-61.
8. Arrand, J. R., Rymo, L., Walsh, J. E., Bjorck, E., Lindahl, T., and Griffin, B. (1981) *Nucl. Acids Res.* **9**, 2999-3014.
9. Redloff, R., Bauer, W., and Vinograd, J. (1967) *Proc. Nat. Acad. Sci. USA* **57**, 1514-1521.
10. Clewell, D., and Helinski, D. (1969) *Proc. Nat. Acad. Sci. USA* **62**, 1159-1166.
11. Wieslander, L. (1979) *Anal. Biochem.* **98**, 305-309.
12. Challberg, M. D., and Englund, P. T. (1980) in *Methods in Enzymology* (Colowick, S. P. and Kaplan, N. C., eds.), Vol. 65, pp. 39-43. Academic Press, New York.
13. Yang, R., Lis, J., and Wu, R. (1979) in *Methods in Enzymology* (Wu, R., ed.), Vol. 68, pp. 176-183, Academic Press, New York.
14. Winter, G., Fields, S., Gait, M., and Brownlee, G. (1981) *Nucl. Acids Res.* **9**, 237-245.
15. Sanger, F., Nicklen, S., and Coulson, A. R. (1977) *Proc. Nat. Acad. Sci. USA* **74**, 5463-5467.
16. Staden, R. (1980) *Nucl. Acids Res.* **8**, 3673-3694.
17. Britten, R. J., Graham, D. E., and Neufeld, B. R. (1974) in *Methods in Enzymology* (Grossman, L. and Moldave, K., eds.), Vol. 29, pp. 378-381, Academic Press, New York.
18. Lau, P. P., and Gray, H. B. (1979) *Nucl. Acids Res.* **6**, 331-357.
19. Lis, J. T. (1980) in *Methods in Enzymology* (Grossman, L., and Moldave, K., eds.), Vol. 65, pp. 347-353, Academic Press, New York.
20. Hong, G. F. (1981) *Biosci. Rep.* **1**, 243-252.
21. Winter, G. and Fields, S. (1980) *Nucl. Acids Res.* **8**, 1965-1974.