*Proc. Natl. Acad. Sci. USA*
Vol. 86, pp. 6917–6921, September 1989
Biochemistry

# A strategy for high-volume sequencing of cosmid DNAs: Random and directed priming with a library of oligonucleotides

(genome sequencing/octamers/nonamers/decamers)

F. WILLIAM STUDIER

Biology Department, Brookhaven National Laboratory, Upton, NY 11973

*Communicated by Richard B. Setlow, June 19, 1989 (received for review March 13, 1989)*

ABSTRACT **Direct sequencing of cosmid DNAs using a library of oligonucleotide primers of length 8, 9, or 10 is proposed. The statistics of priming indicate that a primer library sufficient for determining the sequence of the entire human genome (100,000 cosmids) would be small enough to be assembled and managed. Such a library would greatly reduce the cost and effort of high-volume sequencing: primers would be instantly available; the sequence of each cosmid DNA could be determined from a single DNA preparation without the necessity for mapping or subcloning; and, because each primer would be used repeatedly, the cost of primers would become a negligible fraction of other costs. A combination of random and directed priming could determine the sequence of a cosmid DNA in 1.2–1.5 times the minimum number of sequencing reactions required, and completely directed priming would be even more efficient. The success of this strategy requires that a considerable fraction of octamers, nonamers, or decamers be able to prime selectively in double-stranded DNAs 45,000 base pairs (bp) long; initial results indicate that this is likely to be the case. The strategy is not limited to cloned DNAs and would be useful for rapid identification and direct sequencing of viral nucleic acids.**

Techniques for rapidly determining nucleotide sequences (1, 2) have had enormous impact on biology. The relative ease of determining sequences and the importance of the information obtained has led to a rapid increase in the amount of sequence information generated. Almost 25 million nucleotides of sequence have been accumulated in the GenBank nucleic acid data base (Release 58.0, December 1988).

Nucleotide sequence information is so valuable that serious consideration is being given to determining the nucleotide sequence of the entire human genome, $3 \times 10^9$ base pairs (bp), and the genomes of other well-studied organisms. Such an enterprise represents a tremendous increase in scale over even the most ambitious sequencing projects that have been undertaken heretofore. Assuming that the human genome could be cloned as a set of ordered cosmids (3), and that each cosmid contains ≈40,000 bp of human DNA (inserted in 5000 bp of vector DNA), sequencing the entire genome would require sequencing ≈100,000 cosmids. Even if the cosmids were sequenced at the rate of one a day, a formidable task for a sequencing center using today's technology, centuries would be required to complete the task.

Consider the effort required to determine the nucleotide sequence of a single cosmid. Currently, the resolution of gel electrophoresis allows determination of perhaps 500 nucleotides (nt) of continuous sequence from a single set of sequencing reactions. Machines are being developed to carry out the sequencing reactions (4) and to read the nucleotide sequence from gel electrophoresis patterns (5, 6). However, a minimum of 160 individual blocks of sequence must be assembled to obtain the sequence of both strands of a 40,000-bp DNA, and a major part of the sequencing effort is to obtain and order all of the blocks of sequence needed.

Three basic strategies are used, alone or in combination, to obtain and order these blocks of nucleotide sequence from larger molecules. (*i*) Mapping strategies use restriction enzymes to obtain and map specific fragments of the DNA molecule; nucleotide sequences of appropriate fragments are determined, and the sequence of the entire molecule is assembled from the known positions of the fragments. (*ii*) Random strategies determine nucleotide sequences of randomly selected fragments of the molecule; these random blocks of sequence are accumulated until the sequence of the entire molecule can be assembled from overlaps. (*iii*) Primer-based strategies use already known sequence information to synthesize unique oligonucleotide primers that can be used to extend the sequence.

These strategies are capable of determining the nucleotide sequence of 40,000-bp DNA molecules, but they are labor intensive and expensive, and as such are not well suited to the task of sequencing large numbers of cosmids. I propose a strategy that would greatly simplify the problem of obtaining and assembling blocks of sequence from cosmid DNAs and which, when applied in large volume, would greatly reduce the cost. This strategy is not limited to cloned DNAs and could potentially be applied to DNAs considerably larger than cosmids.

## Statistics of Oligonucleotide Priming

The sequencing strategy I propose comes from a consideration of the statistics of oligonucleotide priming in the enzymatic sequencing technique. Assume that oligonucleotides of arbitrary size can be made to prime DNA synthesis at every perfectly complementary sequence in a DNA molecule but at no other sequence. Further assume that an average cosmid DNA contains 40,000 bp of cloned DNA and 5,000 bp of vector DNA, so that the single strands of such a DNA molecule will contain ≈90,000 potential priming sites. Oligonucleotides as short as hexamers are able to prime efficiently (7) but are too short to provide many unique priming sites in a molecule the size of cosmid DNA. For a DNA molecule of random sequence, the expected frequency of priming sites for a randomly selected oligonucleotide is approximated by the Poisson distribution

$$P(r) = \frac{n^r e^{-n}}{r!}, \qquad [1]$$

where $P(r)$ is the probability of having exactly $r$ priming sites in the DNA molecule and $n$ is the average number of priming sites for an individual oligonucleotide per DNA molecule = $2L/4^p$, where $L$ is the length of the DNA in base pairs and $4^p$ is the number of different combinations of the four nucleotides

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: nt, nucleotide(s).

that can form an oligonucleotide of length $p$. The probabilities that oligonucleotides between 6 and 12 nt long will have no priming site, exactly one priming site, or more than one priming site in a DNA of 45,000 bp are given in Table 1.

Primers useful for determining nucleotide sequence are those that have a single priming site in the DNA molecule. As seen in Table 1, the probability $P(1)$ that a randomly chosen oligonucleotide will have a unique priming site in a cosmid is appreciable only for octamers (0.348) and nonamers (0.244). The maximum possible value of $P(1)$ is 0.368, which would be found if the average number of priming sites per DNA molecule were 1. The average number of priming sites for a nonamer in a cosmid DNA is only 0.34, but using a mixture of two or three nonamers would increase this to 0.69 or 1.03, which would increase the probability of unique priming to 0.346 or 0.368. Cosmid DNAs can range in size between ≈37,000 and ≈52,000 bp, and octamers or a mixture of nonamers have about one chance in three of unique priming across this entire size range.

Priming in situations where it is not known whether (or where) the selected oligonucleotide will prime in the DNA molecule will be referred to as random priming. On the other hand, if the sequence of part of the DNA molecule is known, a primer unique to the known sequence can be chosen to direct DNA synthesis into the unknown region. The probability that such a selected primer will be unique to the entire molecule is simply the probability $P(0)$ that no priming site occurs in the unknown sequence. As seen again in Table 1, P(0) for cosmid DNAs becomes appreciable for primers 8 nt or longer and approaches 1 for primers longer than 12 nt. Priming in situations where the selected oligonucleotide is known to prime at a single site within the known sequence is referred to as directed priming, which, for the purposes of this paper, is assumed always to be directed toward an unknown sequence in the DNA molecule being sequenced.

## Sequencing Strategy

The statistics of unique priming suggests that priming with a library of octamers, nonamers, or decamers would be an effective strategy for high-volume sequencing of cosmids. Synthesis of each oligonucleotide produces enough material for perhaps $10^5$–$10^7$ primings, and the same set of primers could be used repeatedly to sequence any cosmid DNA. On the scale of human genome sequencing, the cost of even a very large library of primers would be a negligible fraction of total costs. Once a library were accumulated, all primers needed for sequencing a DNA would be instantly available. A single preparation of DNA could suffice for determining the entire sequence, and the considerable cost and effort of subcloning, mapping, and preparing multiple DNA samples for sequencing would be eliminated. Not all sequencing reactions primed with octamers, nonamers, or decamers would be productive, but, as shown below, the additional sequencing burden would be relatively small.

Table 1. Frequencies of priming sites in a 45,000-bp DNA for oligonucleotide primers of different lengths, assuming random nucleotide sequence and random selection of primer

| Primer length $p$ | Total possible primers ($4^p$) | Average no. of sites per molecule, $n$ | Probability of 0, 1, or >1 priming sites per 45,000-bp DNA molecule | | |
|---|---|---|---|---|---|
| | | | $P(0)$ | $P(1)$ | $P(>1)$ |
| 6 | 4,096 | 22.0 | $<10^{-9}$ | $<10^{-8}$ | 1.000 |
| 7 | 16,384 | 5.49 | 0.004 | 0.023 | 0.973 |
| 8 | 65,536 | 1.37 | 0.253 | 0.348 | 0.399 |
| 9 | 262,144 | 0.343 | 0.709 | 0.244 | 0.047 |
| 10 | 1,048,576 | 0.0858 | 0.918 | 0.079 | 0.003 |
| 11 | 4,194,304 | 0.0215 | 0.979 | 0.021 | $<10^{-3}$ |
| 12 | 16,777,216 | 0.00536 | 0.995 | 0.005 | $<10^{-4}$ |

$n = 90{,}000/4^p$.

Two strategies can be considered—a combination of random and directed priming, where initial blocks of sequence are generated by random priming and then extended by directed priming until they merge (Fig. 1), or completely directed priming, starting from within vector sequences at the ends of the cloned DNA. Completely directed priming is somewhat more efficient, particularly with the longer primers, but random priming would allow the sequence of a given DNA to be completed more quickly and would be an efficient way to begin if the sequence of the DNA were completely unknown. The same set of primers could be used to initiate random blocks of sequence in each new DNA, and, in a genome sequencing project, the initial blocks of sequence would be compared with already known sequence to look for any overlaps between the new and previously sequenced DNAs. In such a context, it might be more efficient to use the initial blocks of randomly primed sequence to establish overlaps between cosmids rather than to make the independent effort to order the cosmid DNAs by other means before sequencing them.

In the random phase of sequencing, only about one-third of all sequencing reactions will produce useful sequence information. Each successful reaction should produce about 500 bp of sequence, and the first 10 blocks of randomly primed sequence would be expected to have an average of about one overlap. In a double-stranded DNA, the complement of any unique primer will also be a unique primer, so each of the initial blocks of randomly primed sequence can be extended 500 bp in the opposite direction simply by priming with the complement of the original primer (or mixture of primers). A set of 30 primers and their complements would generate perhaps 20–25% of the sequence in 8–11 blocks of 1000 bp each.

The initial blocks of sequence can be extended in both directions by directed priming. If the switch from random to directed priming is made after the sequence of 10,000 bp of the cloned DNA is known, the probability of obtaining sequence information from a set of sequencing reactions is ≈0.400 when primed by an octamer, 0.795 when primed by a nonamer, and 0.944 when primed by a decamer, and these

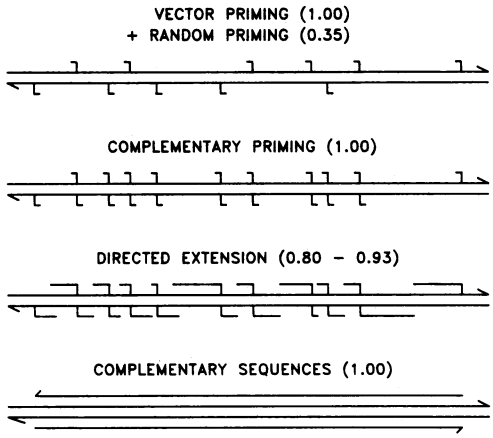


FIG. 1. Sequencing the cloned portion of a cosmid DNA by random and directed priming. Line lengths are to scale for an unknown sequence of 40,000 bp, vector sequences of 2,500 bp at each end, and primings that produce 500 nt of sequence each. The expected fraction of unique primings at each stage is given; during directed extension this fraction would be 0.40–0.74 for octamer primers, 0.80–0.93 for nonamers (as shown), or 0.94–0.98 for decamers (Table 2). Priming from within the vector sequences into the ends of the cloned DNA is assumed to use primers that are long enough to be unique.

probabilities increase as the amount of unknown sequence decreases (Table 2).

Using the probabilities of unique priming in the random and directed phases, making allowance for redundant sequencing when known blocks of sequence merge, and assuming that unique primers within the vector sequence are used to direct sequencing into the ends of the cloned DNA, the average numbers of sets of sequencing reactions needed to complete the sequence of a cosmid DNA have been estimated (Table 3). An average of 161 sets would be required to complete the first strand when directed priming is by octamers, 117 when by nonamers, or 109 when by decamers, compared with the required minimum of 80. Once the entire sequence is known on one strand, primers known to be unique in the molecule can be selected to complete the confirmatory sequences on the complementary strand. Assuming that these sequences could be completed in the required minimum 80 sets, the entire sequence of both strands could be completed in an average of about 1.51, 1.23, or 1.18 times the minimum of 160 sets required, using octamers, nonamers, or decamers, respectively.

Entirely directed priming, starting from within the vector portion of the cosmid and proceeding sequentially along each strand, would be somewhat more efficient than a combination of random and directed priming. An average of 153 sets of sequencing reactions primed by octamers, 93 by nonamers, or 83 by decamers would be required to complete the sequence of the first strand, compared with the required minimum of 80; completion of both strands would require 1.46, 1.08, or 1.02 times the required minimum of 160 sets (Table 3).

### Library Size

The sequencing efficiency of a primer library depends on the density of unique priming sites represented in the library. If ≈500 nt of sequence can be determined from a set of sequencing reactions, a unique primer must be available within a considerably smaller interval to be able to extend known sequences efficiently. The density of unique priming sites for octamers, nonamers, and decamers in a cosmid DNA is high enough that priming from a complete library would allow any known sequence to be extended with little overlap. Reducing library size increases the average overlap because the primer that would initiate synthesis with minimal overlap may not be present in the library. Any increase in overlap decreases the amount of new sequence that can be obtained from a set of sequencing reactions.

The probability $P(x)$ that a library of oligonucleotides will contain no primer that initiates DNA synthesis uniquely within an interval of $x$ nucleotides in a DNA molecule is

$$P(x) = [1 - P(S)P(0)]^x \quad [2]$$

Table 2. Probabilities, $P(0)$, of having no priming sites for an octamer, nonamer, or decamer in different lengths of unknown nucleotide sequence

| | Probability $P(0)$ of no priming site | | |
|---|---|---|---|
| $L$, bp | Octamer | Nonamer | Decamer |
| 45,000 | 0.253 | 0.709 | 0.918 |
| 40,000 | 0.295 | 0.737 | 0.927 |
| 35,000 | 0.344 | 0.766 | 0.935 |
| 30,000 | 0.400 | 0.795 | 0.944 |
| 25,000 | 0.466 | 0.826 | 0.953 |
| 20,000 | 0.543 | 0.858 | 0.963 |
| 15,000 | 0.633 | 0.892 | 0.971 |
| 10,000 | 0.737 | 0.927 | 0.981 |
| 5,000 | 0.858 | 0.963 | 0.991 |

$L$, length of unknown sequence.

Table 3. Average numbers of sets of sequencing reactions needed to complete the nucleotide sequence of a 45,000-bp cosmid DNA

| | Sets of sequencing reactions | | | |
|---|---|---|---|---|
| | Minimum possible | Primers of length 8 | 9 | 10 |
| Random and directed priming | | | | |
| Priming from vector | 2 | 2 | | |
| Random priming | 9 | 25.9 | | |
| Complementary priming | 9 | 9 | | |
| Directed priming (10,000 bp) | | | | |
| 30,000 bp unknown | 20 | 50.0 | 25.2 | 21.2 |
| 20,000 bp unknown | 20 | 36.8 | 23.3 | 20.8 |
| 10,000 bp unknown | 20 | 27.1 | 21.6 | 20.4 |
| Overlaps | | 10 | 10 | 10 |
| First strand total | 80 | 160.8 | 117.0 | 109.3 |
| Complement | 80 | 80 | 80 | 80 |
| Both strands | 160 | 240.8 | 197.0 | 189.3 |
| Ratio | 1.00 | 1.51 | 1.23 | 1.18 |
| Completely directed priming | | | | |
| Priming from vector | 2 | 2 | 2 | 2 |
| Rest of first strand | 78 | 151.1 | 91.0 | 81.0 |
| Complement | 80 | 80 | 80 | 80 |
| Both strands | 160 | 233.1 | 173.0 | 163.0 |
| Ratio | 1.00 | 1.46 | 1.08 | 1.02 |

It is assumed that the cosmid DNA contains 40,000 bp of unknown sequence and 5,000 bp of known vector sequence, that 500 nt of sequence is obtained from each set of productive sequencing reactions, and that the following strategies are used: (*i*) random priming with octamers, followed by directed priming with octamers, nonamers, or decamers after 25% of the sequence has been determined; or (*ii*) completely directed priming with octamers, nonamers, or decamers. Priming from within the vector sequence is assumed to be unique.

where $P(S)$ is the probability that an oligonucleotide that primes at a particular site will be found in the library, and $P(0)$ is the probability that the selected primer will be unique in the entire DNA molecule. For a library of oligonucleotides of length $p$, $P(S)$ is simply the number of oligonucleotides in the library divided by $4^p$, the total possible number; $P(0)$ is given by the Poisson distribution and is a function of both the length of the DNA molecule and the length of the primer.

Two quantities derived from the above expression are useful for estimating and comparing efficiencies of libraries, the 90% priming interval and the average priming interval. The 90% priming interval is the nucleotide length in a cosmid DNA within which the library has a 90% chance of having a primer that initiates uniquely. This length provides a convenient estimate of the longest overlap likely to be needed to extend a known sequence, since a library has a 99% chance of priming uniquely within twice the 90% interval, a 99.9% chance within three times the 90% interval, etc. The average priming interval, on the other hand, is the average nucleotide length needed to reach the first useful priming site represented in the library and is given by

$$\sum (x - 1)\,[P(x - 1) - P(x)] = \frac{1}{P(S)P(0)} - 1. \quad [3]$$

This is the average reduction in new sequence obtained per reaction set when the library is used to extend known sequences.

Library sizes of octamers, nonamers, or decamers needed to achieve different priming densities are given in Table 4. Libraries with a 90% priming interval of 25 nt are very

Table 4. Library sizes of octamers, nonamers, or decamers required for different priming intervals in cosmid DNAs

| Priming interval | | Library size required | | |
|---|---|---|---|---|
| 90% | Average | Octamer | Nonamer | Decamer |
| 12.5 | 4.9 | 43,532 | 62,167 | 192,219 |
| 25 | 10.4 | 22,768 | 32,514 | 100,532 |
| 50 | 21.2 | 11,646 | 16,631 | 51,423 |
| 100 | 42.9 | 5,890 | 8,411 | 26,008 |
| 200 | 86.4 | 2,962 | 4,230 | 13,079 |

efficient for sequencing cosmids: the length of new sequence obtained per reaction set would be reduced an average of only 10.4 nt, ≈2% of the 500 nt of sequence typically obtained, and fewer than one cosmid in five would require an overlap as long as 75 nt. This level of efficiency requires ≈22,800 octamers, 32,500 nonamers, or 100,000 decamers. Libraries containing as few as 5,900 octamers, 8,400 nonamers, or 26,000 decamers would still be reasonably efficient: the length of new sequence obtained per reaction set would be reduced by ≈43 nt, <10%, but an average of about two sequence extensions per cosmid would require an overlap of >200 nt. Optimum library size will presumably reflect some balance between the costs of analyzing sequencing reactions or of adding more oligonucleotides to the library.

To have an equivalent sequencing efficiency, a library need contain only ≈43% more nonamers than octamers but >3-fold more decamers than nonamers, the difference reflecting the effect of $P(0)$. Equivalent sequencing efficiencies with longer primers require ≈4-fold increase in library size for each additional nucleotide of primer length.

### Cost of Primers

In a large-scale sequencing project, each preparation of a primer would ultimately be completely used for generating sequence and then replaced. The cost of primers in such a situation can be estimated to be $<\$10^{-5}$ per nucleotide of sequence obtained. [Custom synthesis of 0.2 μmol of an octamer can currently be purchased commercially for $60, a nonamer for $65, or a decamer for $70 (from Genetic Designs, Houston, TX, for example). Assume that 1 pmol is used to prime each sequencing reaction, that an average of 500 nt of sequence is obtained from each set of four reactions, and that the entire sample is used for priming sequencing reactions; under these conditions, the cost of primer would be $<\$10^{-5}$ per nt of sequence obtained.] Libraries of 22,800 octamers, 32,500 nonamers, or 100,000 decamers would have more than enough priming capacity to sequence the entire human genome, and even if purchased at current commercial prices the most expensive would cost only ≈$0.001 per nt of human genome, a small fraction of other costs. By contrast, equivalent costs for primers in conventional primer-based sequencing would be ≈$0.20 per nt of sequence obtained, assuming primers are 16 nt long and used only once.

An extensive library of primers and its attendant expense need not be provided at the outset. A limited set of perhaps 30–60 primers and their complements could be used repeatedly to prime the initial sequencing reactions on different cosmids. Primers for extending known sequences in each cosmid could be synthesized as required and added to the library. Initially, the cost of sequencing would be similar to that of conventional primer-based sequencing, but as the library increased, fewer new primers would have to be synthesized and the cost of sequencing would steadily decline. A highly efficient library would be accumulated by the time a few hundred cosmids had been sequenced. No matter how a library were accumulated, it would have so much priming capacity that it could be divided among multiple sequencing centers.

### Potential Problems

The success of this sequencing strategy depends on finding conditions for error-free priming by a wide range of octamers, nonamers, or decamers. M. Blewitt and X. Zhang of this laboratory have tested about a dozen octamers for ability to prime DNA synthesis by modified T7 DNA polymerase (Sequenase, United States Biochemical) at sites known to be unique in T7 DNA or λ DNA, double-stranded molecules of known sequence that are about the size of a cosmid (8–10). Some of these octamers primed very clean sequencing ladders, equivalent to the best obtained with any primer (for example, see Fig. 2). However, others seemed to prime poorly, to prime at more than one site, or to prime at an incorrect site. About one-third to one-half of the octamers gave good sequence information under at least one condition tested so far. Nonamers and decamers would be expected to be somewhat better primers than octamers.

Inaccurate or inefficient priming could have several causes. Some oligonucleotides might not prime selectively or efficiently for reasons intrinsic to their sequence or structure, or might prime effectively under some reaction conditions but not others. For example, primers that are palindromes would not be useful for sequencing double-stranded DNA because if they prime at all they must prime at sites in both strands. Priming might also be affected by folded structures in the template DNA, which could mask or expose priming sites. If we can learn general rules for predicting which primers will be useful and for optimizing reaction conditions, a defined subset of octamers, nonamers, or decamers would be sufficient for the success of this sequencing strategy.

### Implementation of the Sequencing Strategy

The first step in implementing this sequencing strategy is to establish conditions where a suitable fraction of octamers, nonamers, or decamers can be used reliably to prime sequencing reactions. Once conditions for reliable priming are established, it may be worth trying to develop a simple test for identifying those sequencing reactions where the oligonucleotide did not prime DNA synthesis or, if possible, to distinguish whether DNA synthesis had been primed at zero, one, or more than one site. Such a test might be based on the yield of acid-precipitable or hybridizable radioactivity or fluorescence. If a reliable test could be developed, the efficiency of the sequencing process could be increased by eliminating the need for gel electrophoresis of samples that will not give useful sequence information. Such a test would be particularly useful for increasing the efficiency of random priming, which can provide useful sequence information in a maximum of 37% of reactions.

Information such as base composition, dinucleotide frequencies, or the sequences of repeated elements in the DNA to be sequenced can be used to optimize the composition or use of a primer library. For example, primers for known highly repeated sequences of a genome might be excluded from the set used for random priming. If the same 5,000-bp cosmid vector were used for all cosmids to be sequenced, an octamer library could be reduced 14%, a nonamer library 4%, and a decamer library 1% without loss of priming efficiency, simply by excluding oligonucleotides that would prime within the known vector sequence.

Computer programs tailored to this strategy will also need to be developed. Initial blocks of sequence produced by random priming must be compared with each other and to the emerging genomic sequence to identify overlaps. A computer will be needed for selecting primers for extending sequences

in a given cosmid and to identify overlaps when blocks of sequence merge. Inventory and maintenance of large libraries of primers might also benefit from special programs.

### Sequencing Other Nucleic Acids

Once an appropriate primer library is available, it can be used for direct sequencing of DNA of any size up to 45,000 bp, and presumably for RNAs as well. The DNA can be from any source, need not be cloned, and can be single- or double-stranded. A strategy of random and directed priming could be particularly useful for direct sequencing of viral nucleic acids of completely unknown sequence. The random priming phase itself, using an appropriate set of primers, could provide rapid and specific identification of viral nucleic acids.

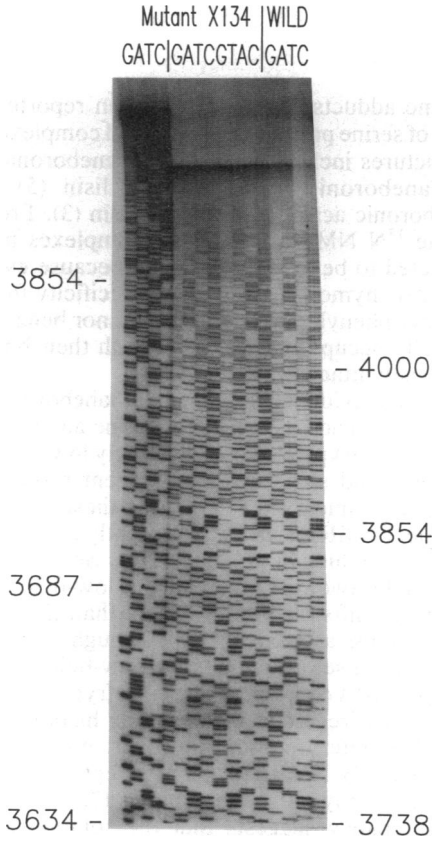


FIG. 2. Sequencing T7 DNA with a unique octamer primer. T7 DNA was prepared from phage particles that had been purified by isopycnic centrifugation in CsCl solution: wild-type DNA was released from the phage particles by phenol extraction followed by chloroform extraction; mutant X134 DNA was released by heating for 5 min at 65°C in 10 mM $Na_3EDTA$ solution. Each DNA was precipitated with 2 vol of 95% (vol/vol) ethanol and redissolved in 10 mM Tris·HCl/0.1 mM $Na_3EDTA$, pH 8.0 at 800 μg/ml. For sequencing reactions, 10 μl of DNA solution was heated for 3 min in a boiling-water bath and placed on ice; 1 μl containing 1 pmol of octamer primer (GGCCATTG) plus 2 μl of 200 mM Tris·HCl, pH 7.5/100 mM $MgCl_2$/250 mM NaCl were added. After 30 min on ice, the two-step sequencing protocol for sequencing with Sequenase (United States Biochemical) was followed, except that the 5-min labeling reaction, using [$\alpha$-$^{32}$P]dATP, was on ice rather than at room temperature. Electrophoresis was through a 0.4-mm thick 6% polyacrylamide gel, with two loadings. After electrophoresis, the gel was soaked in 10% (vol/vol) acetic acid/12% methanol (vol/vol) for 15 min and then dried under vacuum before autoradiography. Priming extended rightward from nt 3607 of T7 DNA (9); nt numbers in T7 DNA are indicated. Mutant X134 DNA contains C → T mutations at nt 3687 and 3854.

In principle, this sequencing strategy could work for molecules longer than cosmids, provided that oligonucleotides of an appropriate length would serve as unique primers for sequencing reactions. However, the longer the template the more of it is needed to provide an equivalent molar concentration of substrate for sequencing reactions, and this limitation will eventually prevail. (For example, direct sequencing of the human genome would theoretically be possible with primers 16 nt long, but sequencing reactions comparable to those of Fig. 2 would require 600 mg of DNA at a concentration of 60 g/ml.) Unique priming is clearly feasible for DNAs 40,000 bp long, as shown in Fig. 2; the upper limit of DNA length will depend ultimately on how sensitively the products of DNA sequencing reactions can be detected.

In random priming, more than one-third of sequencing reactions will be informative as long as the average number of priming sites per molecule is between 0.6 and 1.5, corresponding to double-stranded DNAs 20,000 to 50,000 bp long for octamer primers, 80,000 to 200,000 bp for nonamers, and 320,000 to 800,000 bp for decamers. For directed priming, the library size and the length of primer needed for a given sequencing efficiency will increase with the length of the DNA to be sequenced. For example, a library having a 90% priming interval of 25 nt for a DNA of 200,000 bp is not possible with octamers and would require 106,000 nonamers or 135,000 decamers. Determining the complete nucleotide sequence of both strands by a random and directed priming strategy would require 1.67 or 1.14 times the minimum possible number of sequencing reactions, depending on whether nonamers or decamers were used.

The economy of the sequencing strategy outlined here comes from repeated use of primers from a library. Random priming on any DNA requires only a limited set of primers, but directed priming requires a much larger library. Cosmids provide large numbers of DNAs of a size to make efficient use of a library of octamers, nonamers, or decamers. Unless a comparable source of larger DNAs becomes available (from a new type of cloning vector or an interesting collection of viral DNAs, for example), larger libraries of longer primers may not be competitive with simply subcloning or fragmenting longer DNAs to about cosmid size for sequencing, or sequencing them directly with a library of primers developed for cosmid DNAs, but with reduced efficiency.

I thank M. Blewitt and X. Zhang for testing the priming ability of octamers on T7 and λ DNAs and K. Thompson for confirming some of the calculations. This work was supported by the Office of Health and Environmental Research of the U.S. Department of Energy.

1. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
2. Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 560–564.
3. Collins, J. & Hohn, B. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 4242–4246.
4. Wilson, R. K., Yuen, A. S., Clark, S. M., Spence, C., Arakelian, P. & Hood, L. E. (1988) *BioTechniques* **6,** 776–787.
5. Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., Heiner, C., Kent, S. B. H. & Hood, L. E. (1986) *Nature (London)* **321,** 674–679.
6. Prober, J. M., Trainor, G. L., Dam, R. J., Hobbs, F. W., Robertson, C. W., Zagursky, R. J., Cocuzza, A. J., Jensen, M. A. & Baumeister, K. (1987) *Science* **238,** 336–341.
7. Feinberg, A. P. & Vogelstein, B. (1983) *Anal. Biochem.* **132,** 6–13.
8. Dunn, J. J. & Studier, F. W. (1983) *J. Mol. Biol.* **166,** 477–535.
9. Moffatt, B. A., Dunn, J. J. & Studier, F. W. (1984) *J. Mol. Biol.* **173,** 265–269.
10. Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. & Petersen, G. B. (1982) *J. Mol. Biol.* **162,** 729–773.