This paper describes a novel method of primer walking using octamer oligonucleotides to prime DNA sequencing reactions. Octamer sequencing is compatible with isotopic and fluorescent sequencing chemistry, reaction conditions are optimized such that the samples can be processed in parallel, and the procedure has the potential to be automated. This strategy is faster than the traditional primer walking sequencing strategy, as the existence of a primer library allows immediate access to a primer for the next sequencing reaction, eliminating delays associated with designing and synthesizing gene-specific primers. The octamer library is comprised of optimized sequencing primers, such that octamer sequencing yields results equivalent to or better than traditional primer walking. This technology is more economical because gene-specific sequencing primers, the major cost in the reaction, are replaced by an optimized subset of frequently occurring octamers that are able to prime multiple reactions.

Primer walking is a very common and effective sequencing strategy. In this procedure a primer designed from a known sequence is used to extend sequence information into a previously unknown region. The new sequence information is used to design the next primer, and the process is continued until the entire sequence of the region of interest is determined. Primer walking has advantages over other sequencing strategies: It does not require multiple cloning steps or subsequent template preparations, and it is a directed and methodical approach. The primary disadvantages of primer walking include the cost of synthesizing gene-specific primers to extend the sequence (mainly because the synthesis of each oligonucleotide produces ~1000-fold excess of primer) and the time involved in designing and synthesizing these new primers. These disadvantages could be eliminated if a comprehensive library of pre-existing primers were available. A strategy for high-volume sequencing of cosmid DNAs using a primer library composed of 8-, 9-, or 10-mers was proposed (Studier 1989), whereas others have proposed synthesizing a library containing a subset of useful octamers or nonamers (Siemieniak and Slightom 1990; Burbelo and Iadarola 1994).

RESULTS AND DISCUSSION

The primary consideration for designing a sequencing strategy using short oligonucleotides is defining the minimal primer length that will be recognized by the DNA polymerase and provide sufficient specificity in the sequencing reaction. Additional considerations include optimizing reaction procedures such that they are amenable to automation, effective with single- or double-stranded templates, and consistently yield high quality sequence data. Assuming random sequence composition, the size of the template directly determines the probability that a priming site will be present. The size of many double-stranded sequencing templates typically ranges from 5 to 10 kbp, or 10 to 20 kb, of target sequence. Single hexamers are insufficient to routinely prime cycle sequencing reactions (data not shown) and are likely to be present multiple times in clones of this size class [probability of existence (P) = 4.88/20 kb of random sequence target DNA]. Strings of hexamers or modules comprised of hexamers and pentamers can prime DNA sequencing reactions (Kieleczawa et al.
HARDIN ET AL.

1992; Kolter et al. 1993, 1994; McCombie and Kieleczawa 1994); however, as these reactions involve multiple pipetting steps of oligonucleotides and require single-stranded templates, these sequencing strategies are not optimal. Although certain heptamers are able to prime cycle sequencing reactions (data not shown), it is possible that the priming site may not be unique \( (P = 1.22/20 \text{ kb of target}) \), making this class of primer suboptimal for primer walking. Octamers, however, can routinely prime cycle sequencing reactions, and it is likely that an octamer priming site will be unique \( (P = 0.30/20 \text{ kb of target}) \). Nonamers can also efficiently prime sequencing reactions (Slightom et al. 1994), and the target priming site will almost certainly be unique \( (P = 0.08/20 \text{ kb}) \), although a nonamer library must be four times the size of an octamer library to achieve a similar quality of coverage.

The main obstacle associated with octamer sequencing is the size of the complete octamer library: \( 4^8 = 65,536 \) primers. Ideally, a primer library should be relatively small in size (~1000 members) and include primers that are predicted to occur frequently, thus minimizing the distance between priming sites. The use of a primer library for high-throughput sequencing projects requires that all reactions be processed in parallel throughout the sequencing procedure and that each step in the procedure has the potential to be automated. Optimization of the octamer library used for sequencing depends on determining the rules for optimal octamer design and construction of a reasonably sized primer library.

To define the rules for primer selection, we assayed the ability of different octamer sequences to generate sequence data using cycle sequencing chemistry. This type of chemistry was chosen because it enables one to identify reaction conditions that allow only perfectly matched primer-template duplexes to form and prime sequence data, it works effectively with single- or double-stranded templates, and it is less sensitive to DNA secondary structure problems because the reaction is carried out at higher temperatures. Complete, double-stranded sequence information was determined for a 3.4-kb cDNA insert from Arabidopsis using octamers containing 50% GC, (Fig. 1A). Sequence information was obtained from 44 of the 51 reactions primed by octamers, a sequencing success rate of ~86%. Four octamers primed the reaction from two positions, producing uninterpretable sequence data, and three primers did not produce a sequence ladder. The

**Figure 1** Representative data generated in octamer primed cycle sequencing reactions using a double-stranded DNA template. (A) Sequencing ladders produced using a 20°C annealing temperature. The signal-to-noise ratio for octamer 42 (CCCAC1TI-) was 3.2, and the ratio for octamer 61 (TTAGAGGC) was 5.7. (B) Sequencing ladders produced using a 30°C annealing temperature. The signal-to-noise ratio for octamer 42 was 5.4, and the ratio for octamer 61 was 8.1. Note that octamer 42 is a member of the optimized octamer library, whereas octamer 61 was removed because it contains TTA in octamer positions 1–3. Lanes were loaded from left to right as A, C, G, and T.

44 successful sequencing reactions produced an average read length of ~340 bases using octamer primers, which was similar to that obtained using standard ~17-base primers. Likewise, at ~99%, the accuracy obtained using octamer primers was similar to that obtained using standard primers. Given the average read length from these reactions and an average primer-walking redundancy of 1.5-fold, complete double-stranded information for this clone would typically require 30 sequencing reactions (3400 bases × 2 strands × 1.5 redundancy/340 average bases per sequence read).

Each of the four instances of double priming encountered in this project was correlated with mismatched priming, with a single mismatch occurring at the 5’ position in the octamer. Therefore, to increase the success rate of this technology by minimizing the frequency of double prim-
ing, priming specificity was made more stringent by increasing the annealing temperature from 20°C to 30°C. At this higher temperature, octamers selectively primed reactions from perfectly matched templates. In addition to minimizing double-priming problems, increasing the annealing temperature also increased the signal strength. The primers that had produced less robust sequence data at 20°C typically produced a 50% to 150% increase in signal strength using the 30°C annealing step (Fig. 1B). Although most octamer sequencing technology development has used isotopic sequencing methods, similar results were obtained when this technology was transferred to an automated sequencing platform using fluorescently tagged dye terminator chemistry. In fact, similar increases in signal strength were observed using the 30°C annealing step (Fig. 2). Perhaps the polymerase is better able to compensate for nearest neighbor effects on duplex stability by clamping the octamer onto the template and/or by increasing nucleotide incorporation rate at the increased annealing temperature, resulting in more efficient extension from the octamer. We predict that virtually any octamer primer composed of 50% GC will produce high quality data using similar reaction conditions.

As octamers can efficiently prime isotopic and fluorescent cycle sequencing reactions, we designed a library of optimal octamers that could be used to sequence virtually any template. Because octamers having a 50% GC content efficiently primed sequencing reactions, our first criterion for the library was that each primer contains 50% GC. To increase the library's functionality with protein coding sequences, octamers were optimized with regard to amino acid and codon use. Considering octamer positions 1–3 as a codon, positions 4–6 as an adjoining codon, and positions 7 and 8 as the first two bases of a third codon, octamers containing termination codons, less frequently used amino acids, or infrequently used codons were removed to reduce the library to 4717 octamers. Among this subset of octamers were 958 complementary octamers, which provide the opportunity to determine sequence information in either direction, thus facilitating the completion of double-strand sequence information or amplification of sequences via PCR (J. McCollum and S. Hardin, 2000).

Figure 2  Data generated using octamers to prime cycle sequencing, dye terminator chemistry from a double-stranded *Anabaena* sequencing template. Chromatograms were aligned using Sequencher 3.0. The highlighted base, G, is 475 bases from the 3' end of octamer 5 (GTGCAAGT). (A) Sequencing reaction using a 20°C annealing temperature. The signal strengths for the reactions are 132 (A), 50 (C), 103 (G), and 74 (T). (B) Sequencing reaction using a 30°C annealing temperature. The signal strengths for the reactions are 214 (A), 79 (C), 183 (G), and 124 (T).
unpubl.). Finally, this set of 958 complementary octamer primers was reduced to 566 primer pairs by eliminating those that would prime sequencing reactions within common cloning vectors. This vector subtraction step could be performed on a more limited or restricted basis, resulting in a larger or smaller octamer library, respectively. An electronic version of this optimized octamer library can be accessed through the World Wide Web (URL is http://www.bchs.uh.edu/octamer_library/).

The coverage of these 566 primer pairs that comprise the optimized octamer library was tested by searching for perfectly matched priming sites in human p53 (P. Chumakov, V.P. Almazov, and J.R. Jenkins, GenBank accession no. X54156) and retinoblastoma (Rb) (Toguchida et al. 1993) genomic or protein-coding DNA sequences (Table 1). The 1132 member octamer library produced close spacing, averaging 37–51 bases between primer matches, which should allow flexibility in choosing the next primer to extend sequence information. We found that as the GC content of the template sequence deviated from 50%, the design of the octamer library became increasingly important (Table 1). For sequences that were ~50% GC, such as the p53 genomic and protein-coding DNAs, the average distances between octamer primers were very similar. However, differences in the average distance between primers were seen as the GC composition diverged from 50%, as in the case of Rb genomic and protein-coding DNAs (Table 1). Closer average spacing was observed in protein-coding sequences, and this is reflected in the sizes of the largest gaps observed in the sequences analyzed (Table 1).

We next compared the coverage of octamers in the optimized library to the coverage of 566 randomly selected octamer pairs containing 50% GC (Table 1). If the GC content of the target DNA is ~50%, as is the case for the p53 gene, then the library comprised of the randomized octamers produced a similar distribution of primer matches as the optimized library. However, as the GC content of the target sequence diverged from 50%, as in the Rb genomic and protein-coding regions, the optimized library produced closer spacing between primers than the randomized library. The optimized octamer library produced ~30% closer spacing of primer matches on both the Rb genomic and protein-coding regions. However, using the randomized library, the largest gaps between primer matches were 637 bases in the genomic (i.e., noncoding) sequence and 159 bases in the coding sequence. Given the fact that this distance is between pairs of primers, even this largest gap would be covered in typical sequencing runs.

We have shown that octamers with 50% GC content can be used to effectively prime sequencing reactions and that an octamer library of only 1132 members can provide sufficient coverage to sequence genomic and cDNA sequences. If necessary, this coverage could be enhanced by selecting primers with 1 base mismatch at the 5' end and by lowering the annealing temperature to 20°C. The octamer library may have additional

| Table 1. Distribution of Optimized and Randomized Octamer Primer Pair Libraries on Two DNA Sequences |
|-------------------------------------------------|-------------------------------------------------|
| **p53**                                        | **Rb**                                          |
| base pairs | GC (%) | avg distance | largest gap | base pairs | GC (%) | avg distance | largest gap |
| Optimized library | | | | |
| genomic sequence | 20,303 | 50 | 39 ± 34 | 227 | 180,388 | 37 | 51 ± 44 | 504 |
| coding sequence | 1,182 | 57 | 38 ± 34 | 204 | 2,787 | 39 | 36 ± 30 | 243 |
| Randomized library | | | | |
| genomic sequence | 20,303 | 50 | 58 ± 44 | 365 | 180,388 | 37 | 77 ± 66 | 637 |
| coding sequence | 1,182 | 57 | 39 ± 31 | 185 | 2,787 | 39 | 50 ± 30 | 159 |

The number of base pairs and the percent GC composition in both the genomic and coding regions for p53 and retinoblastoma DNA sequences are shown. To indicate the variability in distance between primer pairs, the average (avg) distances between primers in these DNAs are shown with the average of the absolute deviations of these distances from their mean. The largest gap between pairs of primers is indicated for each DNA.
uses. For example, the high frequency of occurrence of the octamers in the library may be useful for speeding the completion of the sequencing project by identifying additional internal entry sites into the project. For this, several octamers could be pooled and used to prime a sequencing reaction. If a single octamer in the pool primes the reaction, then an additional entry into the reaction. If a single octamer in the pool primes the reaction, then an additional entry into the DNA is identified. During sequencing of the 3.4-kb Arabidopsis cDNA, an octamer primed a sequencing reaction from an internal region and enabled us to proceed from an additional site into the insert sequence.

Octamer sequencing technology can be used by the typical laboratory or by genome sequencing centers to generate high-quality sequence from virtually any source of DNA. This strategy will work most efficiently in conjunction with an octamer library; however, an octamer library is not required to perform the sequencing reactions. This strategy should make sequencing faster, better and cheaper, and should facilitate analysis of specific genes and whole genomes.

**METHODS**

**Sequencing Methods**

Isotopic sequencing reactions using $^{32}$P-labeled dATP (800 Ci/m mole, New England Nuclear) were performed using an EpiCentre Technologies, Inc., SequiTherm Cycle Sequencing Kit according to the “Cycle Sequencing Using Internal Labeling with alpha-Labeled Nucleotides” procedure, except that 25 pmoles of primer were added to the reaction, and a 1-min annealing step at 20°C or 30°C was included between the 95°C denaturation and 70°C synthesis steps. Signal-to-noise ratios were determined using a Fuji BAS1000 or BAS2000 PhosphorImaging system to quantify the relative amounts of radioactivity incorporated into the sequence ladder versus the amount of background radioactivity. Isotopic sequencing gels were placed against a BAS-III-S imaging plate for 2 hr, the plate was scanned, and the data were analyzed. Analysis was performed by determining signal intensities from two representative sequence bands, averaging these numbers, and dividing this number by the background signal intensity value. Automated sequencing reactions were performed using a Perkin-Elmer PRISM Dye Terminator Cycle Sequencing Ready Reaction Kit containing AmpliTaq DNA Polymerase, FS, according to the manufacturer’s directions, except that 25 pmoles of an octamer were added to the reaction. The reactions were cycled on a GeneAmp PCR System 9600 as per manufacturer’s instructions, except that they were annealed at 20°C or 30°C for 1 min. Sequencing reactions were ethanol precipitated, pellets were resuspended in 8 μl of loading buffer, 1.5 μl was loaded on a sequencing gel, and the data were collected by an ABI PRISM 377 DNA Sequencer.

**Library Design**

Starting with all octamers containing 50% GC, selected octamers were removed that contained either (1) termination codons (TAG, TAA, TGA) in positions 1–3 or 4–6 or a TA in positions 7–8; (2) codons specifying less frequently occurring amino acids [Cys (TGC/TGT); His (CAC/CAT); Tyr (TAC/TAT); Phe (TTC/TTC); Asn (AAC/AAT); Ile (ATA/ATC/ATT)] at positions 1–3 or 4–6; or (3) infrequently used codons (Gly, CAA; Thr, ACG; Pro, CCC; Arg, CGA; Arg, CGT; Leu, CTA; Leu, CTT; Ala, GCC; Val, GTA; Ser, TCG; Leu, TTA). CodonUse 3.5 analysis software provided the human codon frequency information (C. Halling and R. Hazelkorn, pers. comm.). Within this subset of 4717 octamers, 958 complementary octamer pairs were identified to facilitate double-stranded sequencing efforts. Next, Sequencher 3.0 (GeneCodes, Inc.) was used to identify and subtract octamers that matched vector sequences [pBlue5SK(−), M13mp18, pBR322, pEMB19(+), pICEM19Rp, pT7T3A18, pBlI30, pMAC78, pMBL604, pRSV-Neo, pSP64, pUC19]. Sequencher search assembly parameters were set at 100% match and 8-base overlap.

**Library Analysis**

We have developed procedures for tracking and screening individual octamers using Sequencher 3.0 (GeneCodes, Inc.) and the Wisconsin Package Genetics Computer Group (GCG) analysis software, v. 8. A data base consisting of all 566 octamer sequences was constructed, and individual octamer sequences within the library were aligned with various DNA sequences. Using Sequencher, the location of octamers identified by this comparison can be graphically presented on the sequence, using alignment settings that require a 100% match and an 8-base overlap. If an octamer primer is present more than one time in a sequence, Sequencher displays the positions of the first match. The Wordsearch program (GCG) was used for the analysis of the human genomic and cDNA sequences, as it was critical to identify all matches in the sequence. After the positions of the octamer matches were identified, the information was exported to a Microsoft Excel spreadsheet for further analysis.

**ACKNOWLEDGMENTS**

We thank Paul Hardin, Beth Allen, and Arnold Eskin for careful reading of the manuscript; Terry Thomas for providing the Arabidopsis clone; James Golden for providing the Anabaena clone; David Needleman for use of the ABI PRISM 377 DNA Sequencer, Zhong Chen and Michael Rice for computer assistance; and members of the Texas A&M University Gene Technologies Laboratory for oligonucleotides and helpful discussions. This work was supported by National Institutes of Health grant R29-HG01151, National Science Foundation grant 9217251, and the Departments of Biology at University of Houston and Texas A&M University.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 USC section 1734 solely to indicate this fact.
REFERENCES


Received January 29, 1996; accepted in revised form April 24, 1996.
Octamer-primed cycle sequencing: design of an optimized primer library.

S H Hardin, L B Jones, R Homayouni, et al.

*Genome Res.* 1996 6: 545-550
Access the most recent version at doi:10.1101/gr.6.6.545

References

This article cites 9 articles, 3 of which can be accessed free at: http://genome.cshlp.org/content/6/6/545.full.html#ref-list-1

License

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.

To subscribe to *Genome Research* go to: http://genome.cshlp.org/subscriptions

Copyright © Cold Spring Harbor Laboratory Press