A computer program for selection of oligonucleotide primers for polymerase chain reactions

Todd Lowe, John Sharefkin, Shi Qi Yang and Carl W.Dieffenbach*
Departments of Surgery and Pathology, Uniformed Services University of the Health Sciences, Bethesda, MD 20814, USA

Received December 18, 1989; Revised and Accepted March 2, 1990

ABSTRACT

We have designed a computer program which rapidly scans nucleic acid sequences to select all possible pairs of oligonucleotides suitable for use as primers to direct efficient DNA amplification by the polymerase chain reaction. This program is based on a set of rules which define in generic terms both the sequence composition of the primers and the amplified region of DNA. These rules (1) enhance primer-to-target sequence hybridization avidity at critical 3'-end extension initiation sites, (2) facilitate attainment of full length extension during the 72°C phase, by minimizing generation of incomplete or nonspecific product and (3) limit primer losses occurring from primer-self or primer-primer homologies. Three examples of primer sets chosen by the program that correctly amplified the target regions starting from RNA are shown. This program should facilitate the rapid selection of effective and specific primers from long gene sequences while providing a flexible choice of various primers to focus study on particular regions of interest.

INTRODUCTION

Use of the polymerase chain reaction (PCR) employing the thermostable DNA polymerase derived from the bacterium Thermus Aquaticus (*Taq*) has allowed sequence-specific and highly sensitive detection and amplification of individual DNA sequences from very small initial sample quantities (1,2). When combined with use of reverse transcriptase (RT) to create cDNA copies of cellular mRNA, the PCR method can also be used to detect and to some extent quantitate specific mRNA sequences, and thus to characterize cellular gene expression phenotype (3–5). Fully achieving the extraordinary sensitivity and high specificity of which the PCR method is capable, however, depends upon several conditions being met to ensure the efficient amplification of the sequence to be detected. These conditions include the availability of highly purified RNA, use of highly purified RT and *Taq* polymerase enzymes, and the use of sense and antisense oligonucleotide primers chosen to give efficient amplification of a uniquely identifiable product. The primers, primer annealing temperatures and times used for the PCR cycle must be chosen to make amplification efficient with a minimum of incompletely amplified product, nonspecific product, or extraneous 'primer-dimer' products arising from primer-primer self- or cross- homologies (6).

At the present time, the methods for preparation of highly purified RNA are well established, and the provision of enzymes of high purity and specific activity has been largely ensured by the availability of both recombinant Moloney murine leukemia virus reverse transcriptase (MMLV-RT) and *Taq* polymerase. The design of appropriate PCR sense and antisense primers, however, is not yet part of any standardized algorithm, and several different sets of rules have been proposed to govern primer sequence selection (4,6,7). These include (a) suggestions to choose primers of similar melting temperatures with 40–60% GC content, (b) to choose primers flanking target sequences containing an intron-exon border, and (c) to choose primers specifying an amplified segment of no more than 0.2 to 0.5 kb in size near the 3' end of a gene sequence. Another suggested rule has been to choose amplified regions containing unique restriction sites to permit better identification of the amplified products (4).

In this report we describe a set of rules for the selection of sense and antisense primers for efficient DNA amplification in PCR reactions. We have written a computer based algorithm to rapidly scan entire gene sequences for all possible primer pairs obeying these rules. The program also displays information such as the length and melting temperature of the amplified product, which may be of additional help in choosing primers. The program also allows the stringency of certain criteria to be varied to better suit the needs of particular experiments.

COMPUTER PROGRAM

Primer Selection Algorithm

The following conditions were applied to the selection of sense and antisense primers by the computer program. To maintain the flexibility of the program, but prevent inappropriate values from being entered, very loose limits on some of the variables have been included.

1. The length of both the sense and antisense primers should be between 18 to 22 nucleotides.

* To whom correspondence should be addressed

2. All primers should contain a GC-type sequence pair (i.e., either a CC, GG, GC, or CG) at their 3′ end. These bonds will facilitate the initiation of complementary strand formation by the RT or by the *Taq* polymerase acting at the 3′ end of the hybridized primer. However, since the antisense primer will be used in its complementary form in relation to the original DNA, the required GC-type sequence for the antisense primers will appear at the 5′ end of the positive (coding) strand of the DNA being searched by the program. There is an on screen reminder that the sequence of the antisense primer must be converted to the complementary form.

3. Each primer should have a GC-type sequence content of between 45% and 55% of its total bases. The program will accept GC content to vary from 10−90%.

4. Sense and antisense primer pairs should specify amplified products of between 100 and 600 base pairs in length. These are the default lengths; different sizes of amplified regions can be chosen by the user. The lower and upper limits for the amplified region the program will allow are 36 bp and 20 kb respectively.

5. No primer should contain more than four contiguous base pairs of homology to itself or to its respective sense or antisense counterpart. If two or fewer contiguous complementary pairs of the primer-primer homology are GC-type bonds, however, then a total of four or fewer bases of self- or sense-antisense homologies is acceptable in primer selection.

6. Parameter number 6 is not an automatic selection rule but a computation of the predicted melting temperature (Tm) of the amplified region defined by any sense-antisense primer pair obeying the above rules. Tm is computed from the standard equation (8):

$$Tm = 81.5°C + 16.6 \log M + 0.41(G+C\%) - 500/\text{amplified length},$$

using the M=0.070 M, the equation reduces to:

$$Tm = 62.3°C + 0.41(G+C\%) - 500/\text{amplified length}.$$

Rule 6 limits primer selection to primer pairs whose amplified product Tm is in the range from 76°C to 82°C (inclusive) for the overall product. The program allows the researcher the flexibility of an upper limit temperature of 100°C. Application of this rule is designed to facilitate the complete denaturation of double stranded DNA product during the 92°C temperature phase of the PCR cycle to guarantee a full doubling of the entire length of the product at each cycle.

Computer Algorithm

The program was written in Turbo Pascal 5.5 (Borland Inc.), on an IBM AT type machine. Gene sequences to be analyzed are first entered in ASCII file format, and saved with the suffix .seq as follows, geneX.seq. Sequences from DNA databases such as GenBank, or files accessed via Microgenie software (Beckman) can be used after translation to the ASCII format. Since the program will recognize only the capital letters A,T,G,C and U as sequence for use by the program, comments can be saved with the sequence as long as they remain all lower case.

The user first enters the file name of the target sequence. Prompts are then given to choose the range and stringency of the primer search. Options include: (a) a choice of which region of the original sequence to search for PCR primers; (b) choice of upper and lower limits on the allowable melting temperature (Tm) of the amplified region, and (c) the allowable range of%GC content in the primers to be chosen.

At this point, the DNA sequence file is read into memory from the disk. The program then builds a bank of possible antisense primers. Beginning at the 5′ end of the region to be searched, the algorithm finds each occurrence of a GC-type sequence and concatenates these two bases with the next 20 bases lying on the 3′ side of the pair, producing a 22-mer antisense primer 'candidate'. The GC content of the candidate 22-mer is then calculated; if the GC content is acceptable (i.e., within a specified range near 50%), the program moves on to the next step. But, if the GC content is too high or to low, the primer is shortened by one base pair at a time at its 3′ end until either (a) it becomes less than 18 bases long and is discarded, or (b) its GC ratio becomes acceptable. The primer is then checked for any self-homology consistent with rule 5 as stated above. If the antisense candidate passes this test, it is added to a bank of other anti-sense primers to be later paired with sense primers. The program continues to choose additional antisense primers as it scans in the 5′ to 3′ direction until it reaches the end of the specified search region.

Once the complete antisense 'bank' has been found, the program begins a similar process of choosing sense primers with the GC type sequence pair at their 3′ ends. This is again done by scanning the desired region from 5′ to 3′. Similar GC content and self-homology standards are applied.

Each suitable sense primer selected in this way is then checked for cross homologies against the transformed version (i.e., the lower strand sequence) of all complementary primers suitable for matching with it. (An antisense primer is suitable for matching only if its distance from a sense primer falls within a user-specified range of desired length for the amplified product.) If no unacceptable cross-homologies are found, the Tm for the amplified region specified by the primer pair is calculated. If the Tm is within the user-specified bounds, the primer set is 'approved' and immediately displayed in a format which demonstrates the sense and antisense base sequences, as well as the Tm and length of their amplified products. A sample of program output for human preproendothelin is shown in Table I. The program then continues to check the sense primer against the rest of the antisense primers in the bank until all compatible antisense primers for that particular sense primer have been found. The search for the next sense primer then continues in the 5′ to 3′ direction, again matching each new sense primer against all the appropriate antisense primers. The program ends when the entire specified region has been scanned.

A final primer set is then selected by the user from the multiple pairs selected by the program. A suitable oligonucleotide probe can usually be chosen from the list of other primers that fall within the amplified region. As an additional guarantee of primer specificity it is advantageous to then check the selected sense, antisense, and probe oligomers against other sequences from the same genome for cross-homologies of 12 or more contiguous base pairs. This task can be accomplished by using the data bank search function of a computer program like Microgenie (Beckman) which is capable of scanning the primate section of GenBank within ten minutes, or all of GenBank in under fifty minutes (when using IBM PC-compatible machines equipped with 80386 processors).

METHODS

RNA Preparation

Total cellular and cytoplasmic RNA prepared by several different methods (9,10) has been used successfully for amplification.

TABLE I. Display of program output for human preproendothelin

SN183= **AGAGTGTGTCTACTTCTGCC**	ASN356= GCTGGAATTTTTGCCAAGCAGG	Tm=79.5 195bp
	ASN359= GGAATTTTTGCCAAGCAGG	Tm=79.5 195bp
	ASN368= GCCAAGCAGGAAAAGAACTCAG	Tm=79.3 207bp
	ASN376= GGAAAAGAACTCAGGGCTGAAG	Tm=79.6 215bp
	ASN389= GGGCTGAAGACATTATGGAGA	Tm=79.6 227bp
	ASN433= GGAAAAGACTGTTCCAAGC	Tm=79.0 269bp
	ASN474= GCAGTTAGTGAGAGGAAGA	Tm=78.9 310bp
	ASN513= **GGAACACCTAAGACAAACCAGG**	Tm=78.8 352bp
	ASN519=*CCTAAGACAAACCAGGTCGGA	Tm=79.0 357bp
	ASN536= CGGAGACCATGAGAAACAG	Tm=79.0 372bp
	ASN542=*CCATGAGAAACAGCGTCAA	Tm=79.1 378bp
	ASN596=*CCTCCAGAGAGCGTTATGTGA	Tm=79.4 434bp
	ASN599=*CCAGAGAGCGTTATGTGAC	Tm=79.4 435bp
	ASN606= **GCGTTATGTGACCCACAAC**	Tm=79.5 442bp
	ASN625= CGAGCACATTGGTGACAGACTT	Tm=79.7 464bp

Printout format for the results of primer search done by the computer program. Primers were chosen from human preproendothelin cDNA by the Primer Selection Program (14). The table displays each sense primer chosen in the left hand column; all possible antisense primers compatible with that sense primer are shown on the right, along with the size and melting temperature of each predicted amplified product. A reminder appears on the screen and at the beginning of the printout that the complementary form of the antisense primers must be used, since the antisense primer is shown as it exists in the coding strand of the original nucleic acid sequence. Antisense primers marked with an asterisk, ASN519,542, 596, and 599 have the potential to form 'primer-dimers' with the sense oligonucleotide due to a two base match at the 3'-ends. This mark warns the researcher of the potential hazard. The final sense (SN183) and antisense (ASN606) primers chosen for human preproendothelin are in bold type, underlined and specified by their initial (5' end) number in the original cDNA sequence (13). An antisense primer shown in bold type, lying between these (ASN513) was chosen as an oligonucleotide probe.

However, all RNA samples employed in this study were processed through one round of LiCl precipitation prior to reverse transcriptase. LiCl precipitation was preformed by adjusting the aqueous solution of RNA to a final concentration of 2.5 M LiCl. The samples were then placed in an ice water bath for 2 hr, the centrifuged at 12,000×g for 15 min. The pellets were washed with 70% ethanol, air dried and dissolved in RNase free water.

Amplification Protocol

The combined reverse transcription/polymerase chain reaction was performed as previously described (11). The polymerase chain reaction was run for 30 cycles using 92°C for 2 min, 50°C for 2 min, and 72°C for 3 min. Following amplification, samples were prepared for electrophoresis by chloroform extraction of the mineral oil. With the exception of the retinoblastoma susceptibility gene amplification product, Southern analysis of the amplification reactions was performed with polynucleotide kinase-labelled oligonucleotide probes as previously described (11), using the hybridization and blot washing conditions described by Saiki et al. (12). The Rb product was analyzed following Southern blotting as described by hybridization with a Rb cDNA clone labelled using the random primer method (13). The blot was washed as described (9).

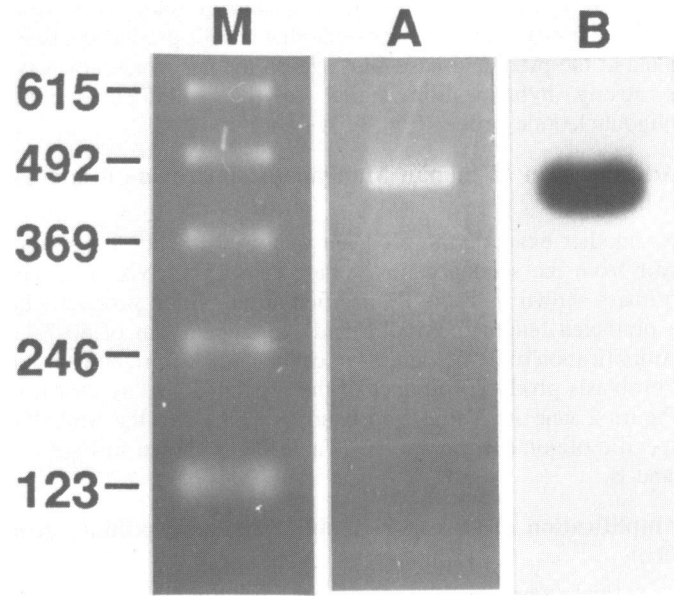

Figure 1. Preproendothelin amplification product. The marker lane shows the bottom 4 bands of the 123 bp ladder. Photograph of amplified product lying close to predicted size of 442 bp is shown in lane A. An autoradiogram of the PCR product from lane A produced by hybridization with ASN513 is shown in lane B.

RESULTS

To test the output of the computer program, primer sets for 3 human mRNAs were used in a sequential RT/PCR system. The RT/PCR system employed was designed to limit the introduction of artifacts into the system. First, all RNA samples regardless of the method of preparation were precipitated with LiCl prior to use. This was done to minimize the contamination of genomic DNA and eliminate traces of cesium from samples extracted by the guanidinium isothiocyanate method (11). This additional step greatly improved the reliability of the RT/PCR system. Second, in performing the RT step, the antisense primer rather than oligodeoxythymidylate was always used to prime the RT reaction. This prevents the relative position of the amplified region within the mRNA/cDNA from having an effect on first strand synthesis. Third, a *Taq* replacement buffer chosen to compensate for levels of $MgCl_2$ and KCl from the 25μl of 1× RT buffer was used. These standardized methods have led to more reproducible RNA phenotyping.

TABLE II. Oligonucleotide sequences used for amplification of target sequences from human MnSOD and Rb mRNA.

Primer or probe	Sequence	Position
MnSoD Sense	5'-GAGATGTTACAGCCCAGATAGC	293–314
MnSoD Probe	5'-GGTTTCAATAAGCAACGGGGAC	517–538
MnSoD antisense	5'-AATCCCCAGCAGTGGAATAAGG	612–591
Rb Sense	5'-GGTCTAACACTGGCATGTTCAAAGC	3817–3841
Rb antisense	5'-CTAGCTGAAGCTACCTTAAATATCC	4071–4046

Listed above are the sequences of the oligonucleotides used for amplification and detection. Note that this listing differs from Table I where the antisense primers have not been reversed. Sequences are as numbered in Beck *et al.*(16) for MnSOD and Lee *et al.* (17) for Rb.

Amplification of human preproendothelin

Using this primer pair search algorithm, we analyzed the sequence of a cDNA clone encoding the precursor of human endothelin (14,15). The algorithm was applied under strictness conditions of allowing primer GC percentages to range from 47% to 53% and allowing a Tm between 76.0°C and 82.5°C. From among the pairs listed in Table I, we picked the sense and antisense primers (underlined). In addition we chose one of the antisense primers picked by the program at a position between these as an oligonucleotide probe for the amplified product. The amplified product had a predicted size of 442 base pairs (bp) and a predicted Tm of 79.5°C. Use of these primers in the sequential RT-PCR reaction with 2.0 micrograms of total RNA from cultured human umbilical vein endothelial cells produced a sharp band at the predicted base pair size on the gel which also gave a strong hybridization signal to the endothelin-specific oligonucleotide probe (Fig. 1, A+B).

Amplification of human Manganous superoxide dismutase (MnSoD)

As another example the program was used to choose a primer pair from the sequence for human MnSoD mRNA (16). The primers shown in Table II specified an amplified product with a predicted length of 319 bp and a predicted Tm of 80.7°C. Amplification of 2 micrograms of cellular RNA from human skin fibroblasts produced product of the predicted size as shown in Figure 2 lane A. The 319bp band hybridized to the MnSOD-specific oligonucleotide probe (Table II) as shown in Figure 2 lane B.

Amplification of human retinoblastoma susceptibility gene (Rb)

As a final example, the program was used to choose a primer pair from the sequence for the cDNA of human Rb gene (17). The amplified product had a predicted length of 254 base pairs and a predicted Tm of 74.8°C. In order to increase primer specificity, a total of 5 extra bases at the 5' end of each program-selected primer were added from the Rb cDNA sequence. These primers produced an amplified product of the predicted size which hybridized to an Rb cDNA probe (Fig 3A,B).

DISCUSSION

Polymerase chain reaction methods have a wide and growing range of applications in research, including detection of mRNA transcripts and identification of transfected sequences in tissue culture and embryonic stem cells. All of these applications have in common the need to choose primers at particular sites in a sequence. As the frequency of PCR use and the number of primer sets sought for various sequences grows, the process of examining sequences by eye to pick sense and antisense oligomers that satisfy even the most rudimentary rules becomes quite time consuming.

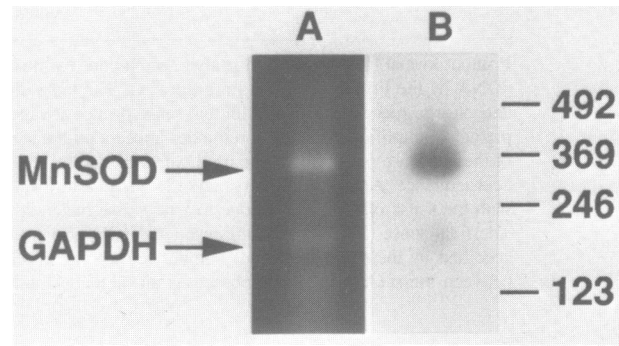

Figure 2. MnSOD amplification product. The photograph of the MnSOD amplification product and autoradiogram of the Southern blot are shown in lanes A and B respectively. The GAPDH (glyceraldehyde-3-phosphate dehydrogenase) product was present as a consequence of direct co-amplification of these two gene products.

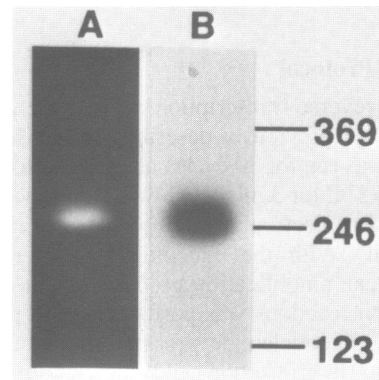

Figure 3. Rb amplification product. The 250 bp amplification product is shown in lane A. An autoradiogram of a Southern blot of the PCR product hybridized with a cDNA probe is shown in lane B.

The computer program described here has been used to pick primers for over 10 gene products. Experimental testing has shown that all the amplification products specified by these primers are of the predicted size and also hybridize with the appropriate cDNA or internal oligonucleotide probe.

As cited above (4,6,7) various sets of rules hitherto proposed for primer selection have been meant to be applied by simple visual inspection of printed sequences. Most of these criteria have focused on guaranteeing the uniqueness of primer-target

hybridization by avoiding obvious problems such as long runs of individual bases and regions of self-complementarily. To these criteria we have added a rule to raise primer specificity and avidity. We have also added the optional use of amplified region Tm to selection of products to help ensure the complete denaturation of the product during the PCR cycle. An additional benefit of using primer sets with closely matched Tm values is that it helps allow coamplification of more than one primer pair to be a valid means of comparing original signal levels among different gene products.

We have, however, made the choice of amplified segment size and the use of the melting temperature of the amplified segment features which the program simply offers as optional or useful information rather than using them to eliminate some primer choices. This was done because the need to position primers in particular ways such as creating nested primer sets for probe purification or positioning for later sequencing of an amplified segment may sometimes take priority over efficiency of amplification based on melting temperature. By also allowing searches to be done with varying degrees of strictness in selecting G+C percentage of primers, the program provides flexibility in choosing primers tailored to particular nucleotide positions and choices of times and annealing temperatures of amplification.

This program is available for distribution for fifty dollars by writing Dr. Carl W. Dieffenbach, C/O Henry M. Jackson Foundation, 11426 Rockville Pike, Suite 400, Rockville, MD 20852-3007.

ACKNOWLEDGEMENTS

The authors are grateful to Dr. Suzanne Eskin of the Department of Surgery, Baylor University, Houston, Texas, and Mr. Scott Diamond of Rice University, Houston, Texas for supplying total cellular RNA from human umbilical vein endothelial cell cultures used for PCR of endothelin cDNA. This work was supported by DOD Contract 86MM6511 DCRN A733 to CWD and NIH NHLBI/DVTB grant RO1 HL40680 to JS.

REFERENCES

1. Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B., and Erlich, H.A. (1988) *Science* **239**, 487–491.
2. Oste, C. (1988) *BioTechniques* **6**, 162–167.
3. Doherty, P.J., Huesca-Contreras, M., Dosch, H.M., Pan, S. (1989) *Analytical Biochemistry* **177**, 7–10.
4. Rappolee, D.A., Wang, A., Mark, D., and Werb, Z. (1989) *Journal of Cellular Biochemistry* **39**, 1–11.
5. Rappolee, D.A., Mark, D., Banda, M.J., and Werb, Z. (1988) *Science* **241**, 708–712.
6. Watson, R. (1989) *Amplifications* **1:2**, 5–6.
7. Willams, J.F. (1989) *BioTechniques* **7**,762–768.
8. McConaughy, B.L.,Laird, C.L., McCarthy, B.J. (1969) *Biochemistry* **8**,3289–3295.
9. Maniatis, T., Fritsch, E.F., and Sambrook, J. (1982) *Molecular Cloning: a laboratory manual*. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
10. Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Smith, J.A., Seidman, J.G., and Struhl, K. (1987) *Current Protocols in Molecular Biology*. Greene Publishing Associates and Wiley-Interscience, New York,N.Y.
11. Jacobsen, H., Mestan, J., Mittnacht, S., and Dieffenbach, C.W. (1989) *Molecular and Cellular Biology* **9**, 3037–3042.
12. Saiki, R.K., Bugawan, T.L., Horn, G.T., Mullis,K.B., and Erlich, H.A. (1986) *Nature* **324**, 163–166.
13. Feinberg, A.P., and Vogelstein, B. (1983) *Anal. Biochem.* **132**, 6–13.
14. Itoh, Y., Yanagisawa, M., Ohkubo, S., Kimura, C., Kosaka, T., Inoue, A., Ishida, N., Mitsui, Y., Onda, H., Fujino, M., Masaki, T. (1988) *FEBS Lett.* **231**,440–444.**332**,411–415.
15. Yanigasawa, M., Kurihara, H., Kimura, S., Tomobe, Y., Kobayashi, M., Mitsui, Y, Yazaki, Y., Goto, K., Masaki, T. (1988) *Nature*
16. Beck, Y., Oren, R., Amit, B., Levanon, A., Gorecki, M., and Hartman, J.R. (1987) *Nuc. Acids Res.* **15**, 9076.
17. Lee, W.-H., Bookstein, R., Hong, F., Young, L.-J., Shew, J.-Y., and Lee, E. Y.-H. P. (1987) *Science* **235**, 1394–1399.