© 2001 Oxford University Press

Isolation of segments of homologous genes with only one conserved amino acid region via PCR

Martin Laging, Berthold Fartmann and Wilfried Kramer*

Abteilung Molekulare Genetik und Präparative Molekularbiologie, Institut für Mikrobiologie und Genetik, Grisebachstraße 8, D-37077 Göttingen, Germany

Received September 18, 2000; Revised and Accepted November 20, 2000

ABSTRACT

We present a method which allows the isolation of fragments from genes coding for homologous proteins via PCR when only one block of conserved amino acids is available. Sets of degenerated primers are defined by reverse translation of the conserved amino acids such that each set contains not more than 128 different sequences. The second primer binding site is provided by a special cassette that is designed such that it does not allow binding of the second primer prior to being copied by DNA synthesis. The cassette is ligated to partially-digested chromosomal DNA. The second primer is biotinylated to allow elimination of PCR products carrying degenerated primers on both sides via streptavidin binding. Fragments obtained after amplification and enrichment are cloned and sequenced. The feasibility of this method was demonstrated in a model experiment, where degenerated primers were deduced from six conserved amino acids within the family of homologs to the *Escherichia coli* Vsr protein.

INTRODUCTION

Since the advent of recombinant DNA technology, cloning the gene of interest has become most often the crucial first step in the functional analysis of this gene, e.g. as a means to construct mutants, to get hold of the protein by overexpression or for other analyses too numerous to be listed here in detail. Unless the gene could be cloned by complementation of a mutant, a pivotal role in the cloning procedure can be ascribed to the availability of some sequence information, however limited it may be. A classical approach involves the purification of the protein, obtaining amino acid sequences from peptides generated by proteolytic digestion and reverse translation of the peptides. The derived DNA sequence, which is bound to be ambiguous due to the degeneracy of the genetic code, can then be employed for the construction of probes to screen a gene library.

The ever increasing amount of available sequence information and in particular the number of completely sequenced genomes has vastly expanded the families of homologous genes in the last few decades. Isolation of a corresponding homologous gene from a given organism could in some cases, if conservation was strong, be achieved using a gene already available as a probe to screen a library by low stringency hybridization. The invention of PCR, however, has opened up the possibility to isolate gene fragments from genes containing just two (or more) blocks of conserved amino acids with little or no homology interspersed between these blocks. The amino acid sequence of a conserved region is reverse translated and a mixture of oligonucleotides is synthesized representing all possible DNA sequences coding for that particular amino acid sequence. Two such degenerate primer mixtures derived from appropriately spaced conserved blocks are employed in a PCR reaction. The PCR products are then, usually after enrichment for the expected fragment length, cloned and sequenced. Occasionally, however, one might be interested in cloning a gene from a family containing just one block of contiguous conserved amino acids sufficient in length or one might want to explore the diversity of genes in an organism containing a particular sequence motif of known functional significance. In these cases, the scheme outlined above will not suffice. The method described here and shown in Figure 1 will combine several improvements to circumvent this limitation.

MATERIALS AND METHODS

Oligonucleotides

Unmodified oligonucleotides were purchased from NAPS (Göttingen, Germany) and biotinylated and IRD800 labeled oligonucleotides from MWG (Ebersberg, Germany). For sequences see Figure 2b and c.

Enzymes and chemicals

Tfl DNA polymerase was purchased from Promega, tetramethylammoniumchloride was from Sigma and dNTPs from Boehringer Mannheim. Restriction enzymes were from New England Biolabs or MBI Fermentas and T4-DNA-Ligase and polynucleotide kinase from MBI Fermentas. The NucleotraPCR® Kit was purchased from Macherey & Nagel and Streptavidin MagneSphere particles from Promega. The GFX Micro Plasmid kit and the Thermo Sequenase cycle sequencing kit were from Amersham Pharmacia. For all enzymatic reactions the buffers supplied by the vendors were used.

*To whom correspondence should be addressed. Tel: +49 551 399653; Fax: +49 551 393805; Email: wkramer@uni-molgen.gwdg.de
Present address:
Berthold Fartmann, MWG Biotech AG, Anzinger Straße 7a, D-85560 Ebersberg, Germany

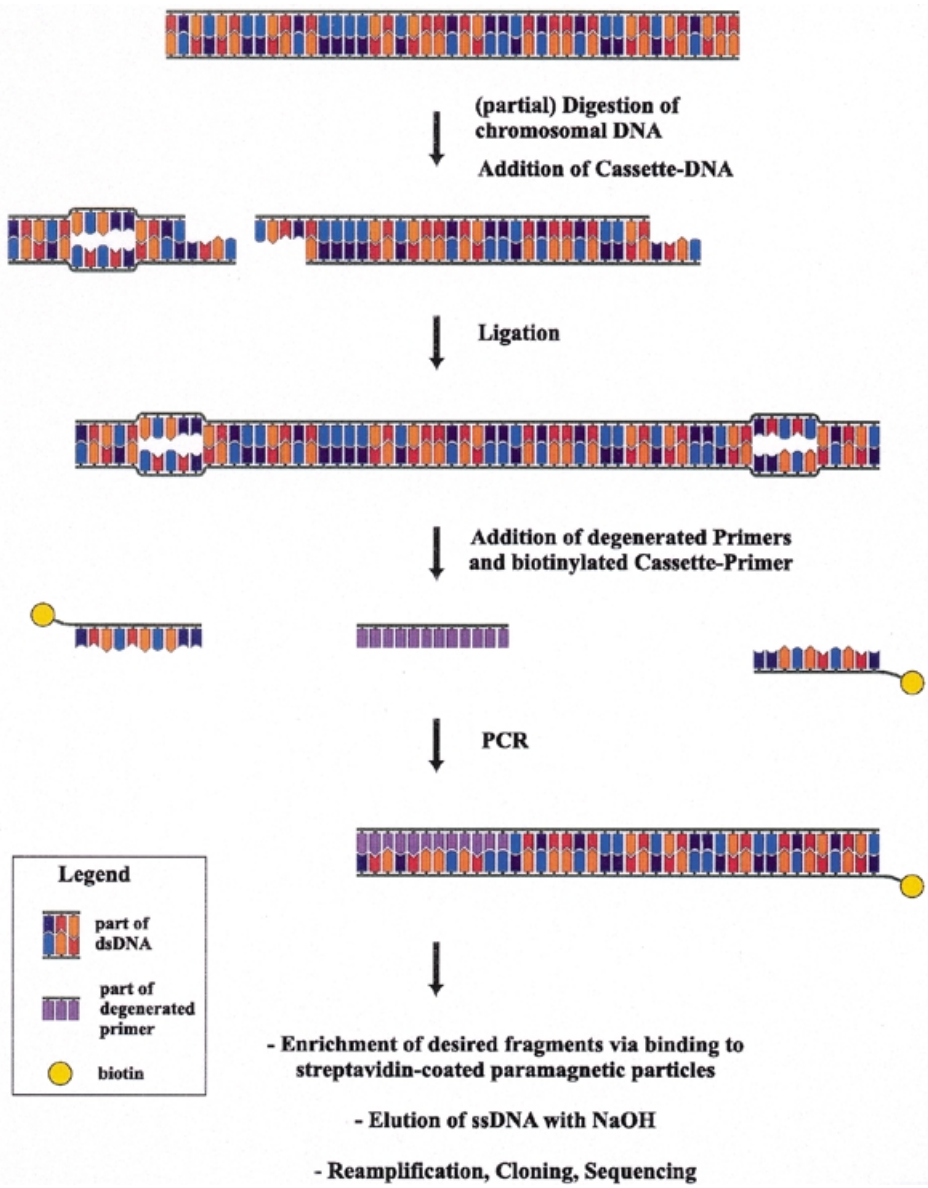

Figure 1. Schematic representation of the method described in the article. Partial digestion and subsequent ligation with cassette DNA at both ends produces template DNA, which is then used for PCR amplification with different degenerated primer sets and a biotinylated cassette primer. Desired products as shown at the bottom of the figure are enriched via binding to streptavidin, reamplified in a second PCR, cloned and sequenced.

Preparation of template DNA

Chromosomal DNA from *Escherichia coli* was prepared from strain DH5α (genotype: [*endA1 hsdR17* ($r_k^-m_k^-$), *supE44*, *thi1*, *recA1*, *gyrA(Nal^r)*, *relA1*, Δ(*lacZYA-argF)*U169, Φ*80lacZ*Δ*M15*]) (1). Cells from 1.5 ml overnight culture were collected by centrifugation and resuspended in 570 µl TE buffer (10 mM Tris–HCl, 1 mM EDTA, pH 8) containing 1 mg/ml Proteinase K. Thirty microliters 10% (w/v) SDS were added and the mixture was incubated at 37°C for 1 h. One hundred microliters of 5 M NaCl and ethidium bromide to a final concentration of 1 mg/ml were added and the mixture was extracted with phenol/chloroform (1:1 v/v). The aqueous phase was re-extracted twice with phenol/chloroform (1:1 v/v). RNase A was added to the aqueous phase to a final concentration of 1 mg/ml and the mixture was incubated for 1 h at room temperature. DNA was precipitated with 0.6 vol isopropanol, washed with 75% ethanol and redissolved in 100 µl TE buffer at 4°C overnight. The chromosomal DNA was partially digested with *Bsp*143I to an average length of 2 kb, extracted with phenol/chloroform, precipitated with ethanol and resuspended in TE buffer. The two cassette oligonucleotides (for sequences see Fig. 2c) were mixed (2 µg each, 360 pmol in total) and phosphorylated with 2 U of T4 polynucleotide kinase in a total volume of 100 µl. For annealing, the oligonucleotide mixture was heated to 80°C for 2 min in an incubation block, which was allowed to cool to room temperature. Two micrograms of partially-digested

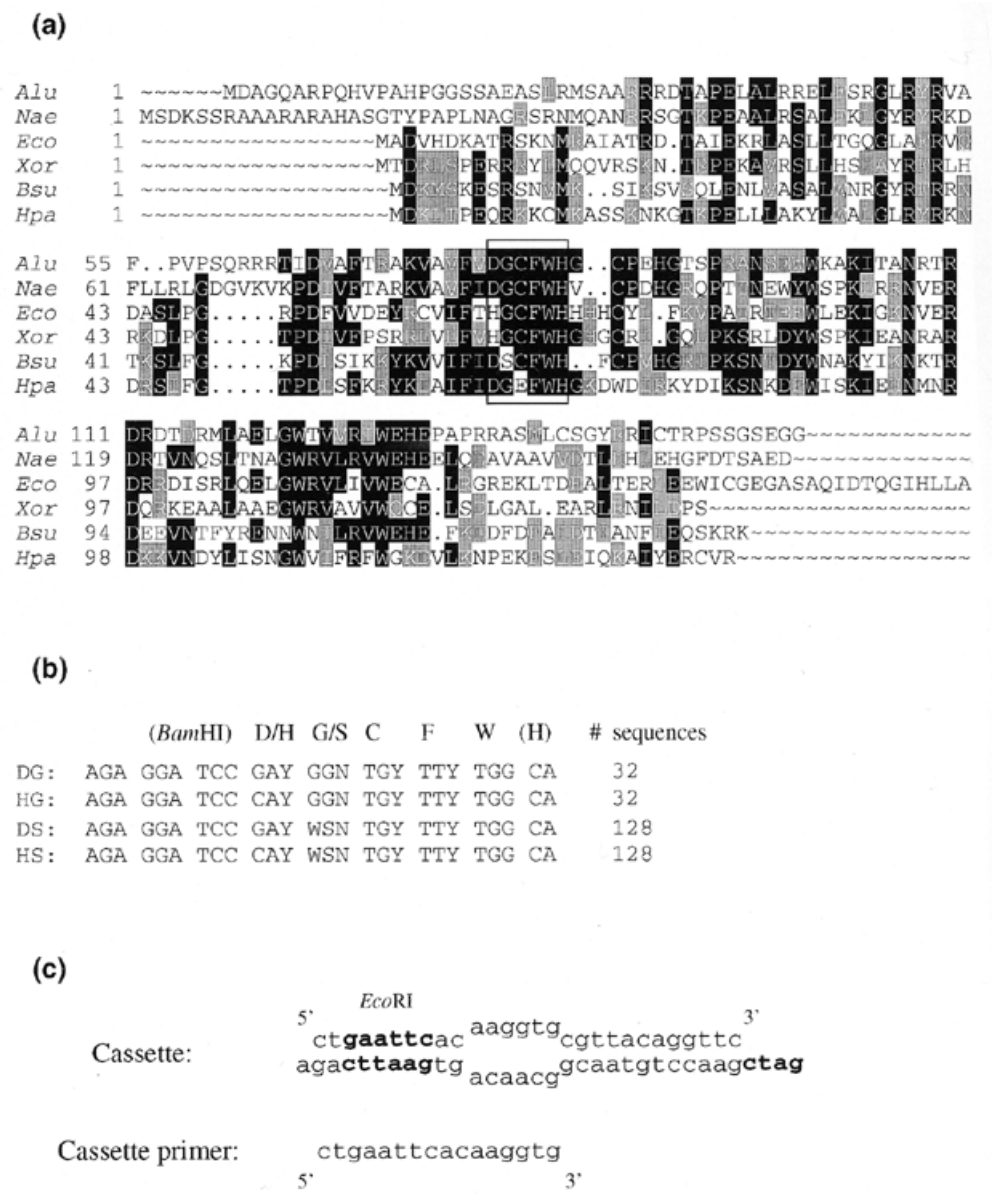

Figure 2. (**a**) Alignment of Vsr homologs (2–7) from *Arthrobacter luteus* (*Alu*), *Nocardia aerocolonigenes* (*Nae*; U09581), *E.coli* (*Eco;* D90835), *Xanthomonas oryzae* (*Xor*; U06424), *Bacillus subtilis* (*Bsu*) and *Haemophilus parainfluenzae* (*Hpa*; L17342). The region chosen for definition of degenerated primer mixes is boxed. Where available, GenBank accession numbers for the nucleotide sequences are given in parentheses. (**b**) Sets of primer sequences coding for the amino acids D/H, G/S, C, F, W and the first two bases of the H codon (W: A or T; Y: C or T). The total number of different sequences within a given primer set is indicated. (**c**) Cassette construct consisting of two oligonucleotides with a central non-pairing bubble. The cassette has a sticky end for ligation to GATC overhangs and an *Eco*RI restriction site. The cassette primer is complementary to the upper strand and its 3′-end covers the bubble region in a way that elongation is possible only after a first round of DNA synthesis with the upper strand as template.

chromosomal DNA were mixed with 200 ng annealed cassette DNA and ligated for 2 h at room temperature with 2 U T4 DNA ligase in a total volume of 120 µl. One microliter of this ligation mixture per reaction was used as template in the subsequent PCRs.

Fragment amplification by PCR

The PCR was carried out in a total volume of 50 µl. The reaction contained 20 pmol of the 5′-biotinylated cassette-primer (see Fig. 2c for sequence), 100 pmol of the respective degenerated primer mix (see Fig. 2b for sequences), 1 µl template DNA (see above), 20 mM Tris–acetate pH 9.0, 10 mM ammonium sulfate, 75 mM potassium acetate, 0.05% Tween-20, 60 mM tetramethylammonium chloride, 1.5 mM $MgCl_2$ and 1 U *Tfl* DNA polymerase. For each degenerated primer mix, eight reactions were set up. PCR was run in an Eppendorf Mastercycler® gradient. Denaturation was for 30 s at 92°C and annealing for 45 s. For the first 10 cycles, the annealing temperature gradient was in eight uniform steps between 60.4 and 68.9°C with a decrement of 0.5°C per cycle, followed by

30 cycles with an annealing temperature gradient in eight steps between 55.5 and 64.0°C. Elongation was for 1 min at 67°C.

Enrichment

Ten microliters of each PCR were treated with the NucleotraPCR® Kit according to the supplier's instructions and the DNA was eluted in 50 µl TE buffer. Forty microliters of the eluate were mixed with 50 µl of Streptavidin MagneSphere particle suspension and 40 µl 2× BW buffer (10 mM Tris–HCl pH 7.5, 1 mM EDTA, 2 M NaCl) and incubated at room temperature on a rocking platform for 15 min. After magnetic collection and washing four times with 100 µl 1× BW buffer, samples were placed on ice and the bound DNA was denatured by suspension of the particles in 40 µl ice-cold 0.1 M NaOH for 2 min. The supernatant was neutralized (addition of 40 µl ice-cold 0.1 M HCl and 5 µl of 1 M Tris–HCl pH 7.5) and used as template (1 µl per reaction) for an additional amplification step. PCR conditions were similar to those described above, but with a fixed annealing temperature of 50°C and 25 rounds of amplification.

Cloning and identification of fragments by sequencing

Reamplification products were purified with the NucleotraPCR® Kit, digested with *Bam*HI and *Eco*RI and cloned into pBluescript II SK(–) Vector (Stratagene). Clones carrying inserts were identified in a PCR-screen with M13 universal and reverse primers flanking the multiple cloning site. DNA was prepared with the GFX Micro Plasmid kit and sequenced with a Thermo Sequenase cycle sequencing kit on a LiCor 4000L sequencer using an IRD800 labeled M13 reverse primer. The sequences obtained were compared with the *vsr*-sequence (2) using the GCG program package.

RESULTS

As a model system for improving the method for isolating segments of genes coding for proteins with just one conserved region sufficiently long for definition of PCR primers, we chose the family of homologs to the *vsr* gene from *E.coli* (2–7). Vsr is a sequence- and strand-specific mismatch endonuclease (8). The *E.coli* protein cleaves DNA 5′ to a thymine mispaired with guanine within the sequence context CTA/TGN or NTA/TGG (the mispaired thymine is underlined). These mismatches frequently occur in *E.coli* due to hydrolytic deamination of 5-methylcytosine within the sequence context CCA/TGG. The inner cytosines (underlined) in this sequence context are modified to 5-methylcytosine by the Dcm methyltransferase. A sequence alignment of the deduced amino acid sequences for *vsr* and several homologous genes is shown in Figure 2a. As can be seen, there is only one suitable region (amino acids 64–69 in the *E.coli* gene) for the definition of degenerate primers. To keep the degeneracy of the primer mixtures low, four sets of primer mixes were defined as shown in Figure 2b. None contained more than 128 different primer sequences.

Having only one conserved region, it was necessary to introduce a primer binding site for a second PCR primer. To this end, chromosomal DNA was partially digested with *Bsp*143I (an isoschizomer of *Sau*3AI) and a cassette as shown in Figure 2c providing a primer binding site was ligated to the chromosomal DNA fragments. By this procedure, both ends of a given chromosomal fragment might acquire a cassette and thus be amplified with the cassette PCR primer irrespective of priming by the degenerated primers. Therefore, the cassette was designed according to the principle of the vectorette (9). It contains a single-stranded overhang at one side that matches the GATC overhangs created by partial digestion of chromosomal DNA. The 5′-ends were phosphorylated to allow ligation of both strands of the cassette. Although ligation is necessary only for the upper strand, which could already be achieved with an unphosphorylated cassette, ligation of the lower strand prevents liberation of this oligonucleotide during denaturation, which might interfere with the PCR. Two double-stranded regions are flanking the central part, which is non-complementary and thus is expected to form a bubble. The cassette primer used for PCR was synthesized such that the 3′-end of this primer is identical in sequence with the upper strand of the bubble. Thus, the cassette primer can prime efficiently only if the upper strand has been copied by DNA synthesis, and thus creating a strand with a sequence complementary to the cassette primer (Fig. 1). This effectively precludes the amplification of DNA fragments via the cassette primer alone without prior DNA synthesis primed by the degenerated primers. This, however, does not eliminate the background of fragments created by priming of the degenerated primers on both sides, which is quite substantial as can be seen in Figure 3a and b, when only the degenerated primers were added. To circumvent this problem, the cassette primer carried a biotin group at the 5′-end (Fig. 1), which allowed the enrichment of the desired products via binding to immobilized streptavidin, which was introduced in a preceding publication for genomic walking, performing radioactive sequencing reactions directly on the single-stranded DNA attached to magnetic beads via the biotinylated specific primer (9).

Whereas in methods employing degenerate primer sets for two conserved regions often a quite sound guess to the length of the expected fragment can be made, if in the alignment the region between the conserved regions contains no or few gaps, this is not possible if only one conserved region is available due to the unpredictable location of the restriction sites. Since the sequence of the *vsr* gene is known, fragments of 257 and 284 bp were to be expected, as this is the distance of two *Bsp*143I sites close to the binding site of the degenerated primers. The knowledge of expected fragment lengths greatly facilitated the interpretation of the data. Template DNA from chromosomal DNA of *E.coli* was prepared as described above and subjected to a touchdown PCR as described in Materials and Methods with the primer sets shown in Figure 2b. As can be seen in Figure 3a, the different primer sets yielded different banding patterns. In comparison with the mixture of all four sets (Fig. 3a, lane 9), the different bands are more pronounced in the reaction using only one set. This demonstrates that lower degeneracy of the primer mixture results in a more efficient amplification of particular bands. A more detailed analysis with the HG primer set containing the matching *vsr* primer is shown in Figure 3b. As can be seen in lanes 1, 4, 7 and 10, the expected fragments sizes of 257/284 bp were obtained in the first round at lower annealing temperatures. (The differences in corresponding lanes in Fig. 3a and b can be explained by the use of different DNA template preparations, which might have differed in the extent of partial digestion and efficiency of cassette ligation.) To enrich PCR products containing the cassette primer, the PCR reactions were treated with streptavidin

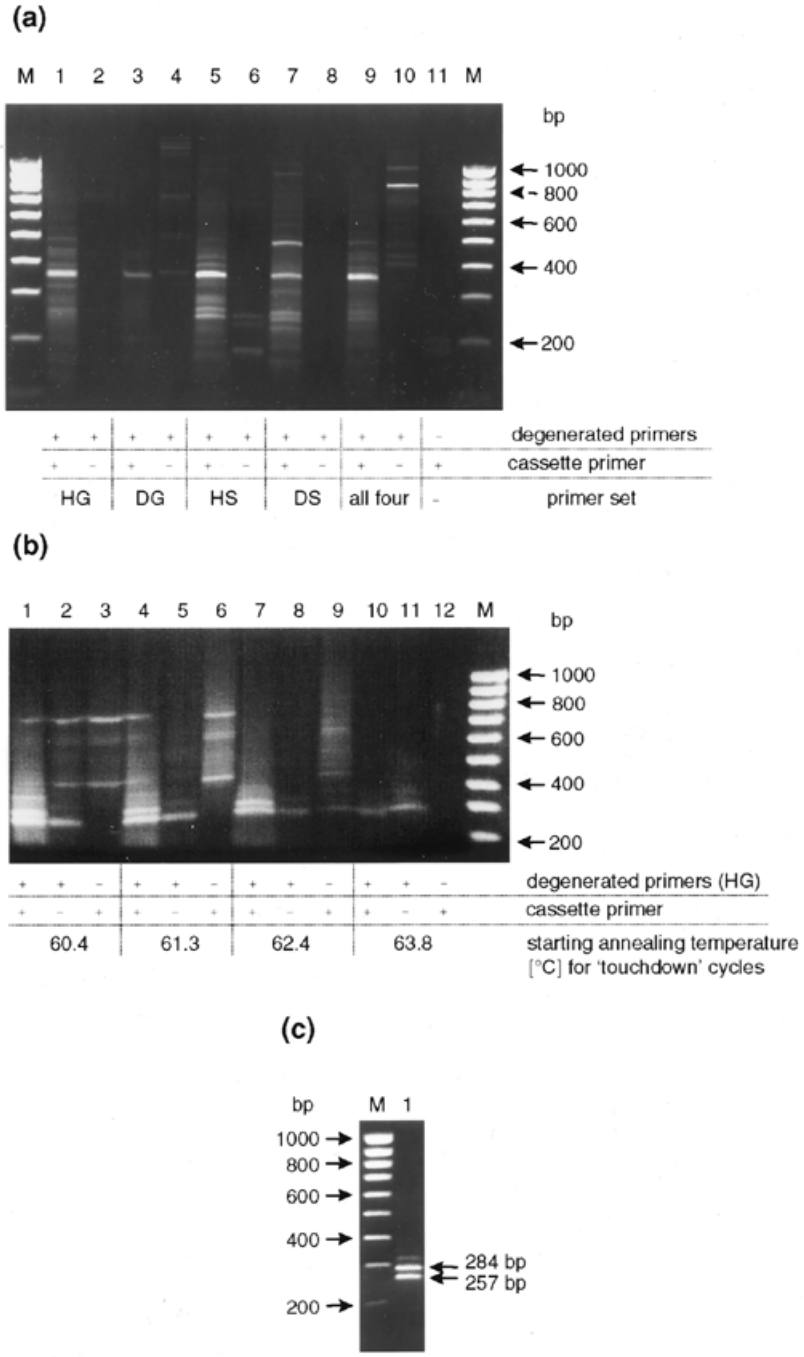

Figure 3. (**a**) PCR products generated by amplification with the four different sets of degenerated primers and a mixture of these. The starting annealing temperature of the touchdown PCR was 60°C. (**b**) PCR products after the first round of amplification with the HG primer set with different annealing temperatures. Primer combinations and starting annealing temperatures of the touchdown cycles are indicated. (**c**) PCR products after reamplification using enriched products from lane 1 in (b) as template. The bands corresponding to the expected fragments of 257 and 284 bp are marked by arrows.

coated MagneSphere beads and eluted under alkaline conditions to eliminate products resulting from amplification with degenerate primers on both sides. The enriched biotinylated fragments were subjected to a reamplification step. As can be seen in Figure 3c, this reamplification yielded only very few bands, demonstrating the efficiency of the enrichment. PCR products from the reamplification step (Fig. 3c) were cloned into pBluescript II SK(–), clones with insert identified by PCR screening and then the insert DNA of six randomly chosen clones was sequenced. Two of the six sequences could be unambiguously assigned to the *vsr* gene (corresponding to the 257 bp PCR product in Fig. 3c). The quality of the third sequence was somewhat poor, but was also most likely derived from the *vsr* gene. Sequence 4 represented bases 86 870–86 955 and sequence

5 bases 1 087 857–1 087 921 of the *E.coli* genome. The chromosomal sequence that can be inferred to be bound by the degenerated primer is **CACGGTTTC**T**TC**T**GGCA** for sequence 4 and AGTC**G**GGT**T**GCC**TGGCA** for sequence 5 (positions matching the primer sequence are in bold). Thus, as is evident from sequence 5, already five matching nucleotides at the 3′-end of the primer are sufficient in cases to yield a PCR product, supporting the necessity for several annealing temperatures to be tested to minimize the frequency of such mispriming events. The origin of the sixth sequence could not be assigned. Although a short identical stretch of 20 bp was found in *E.coli* DNA, similar identities were also found in other organism, e.g. mouse and *Arabidopsis*. Thus, this fragment might have originated from trace amounts of contaminating DNA.

Taken together, these results demonstrate the feasibility of the method to efficiently identify homologous genes carrying just one conserved region. With this method, we were also able to identify a hitherto unknown Vsr homolog from *Bacillus stearothermophilus* (unpublished data).

DISCUSSION

The method presented in this study combines three features, which together greatly facilitate the isolation of fragments of genes with just one conserved region. (i) The splitting of degenerate primers into different sets alleviates one of the major problems in working with such mixtures. If the mixture contains a large number of different sequences, the concentration of the correctly matching primer is very low, leading to low product yields. Choosing annealing conditions that allow primers with one or two mismatches to prime reduces this problem to some extent, but the trade-off is a reduced specificity resulting in more background. The technique of using different primer sets with a degeneracy between 64- and 256-fold from two conserved regions was successfully employed for isolating fragments of *mutS*-homologous genes from a variety of sources, e.g. from the *msh2* gene of *Schizoaccharomyces pombe* (10), from several *Zea mays* homologs (R.Kunze, B.Fartmann and W.Kramer, unpublished results) and from MutS II-homologs (11) of *Thermus aquaticus* and *Clostridium acetobutylicum* (B.Fartmann and W.Kramer, unpublished results). (ii) Application of the vectorette principle (9) allows the provision of a second primer binding site by ligation of an oligonucleotide cassette to the ends of chromosomal DNA fragments without the problem of amplification of any chromosomal DNA fragment with a cassette on either end via the second primer, which would create a vast background of unspecific fragments. (iii) The use of biotinylated cassette primers allowing the enrichment of PCR products containing the cassette primers effectively reduces the background of PCR products generated by amplification with degenerated primers from both sides. Since 50% of the clones obtained did contain a *vsr* fragment, it can be concluded that this PCR method should be an effective means for isolating members of gene families coding for proteins with just one conserved region.

ACKNOWLEDGEMENTS

This work was supported through a grant of the Deutsche Forschungsgemeinschaft. M.L. is supported by the Graduiertenkolleg ‘Chemische Aktivitäten von Mikroorganismen’.

REFERENCES

1. Hanahan,D. (1983) Studies on transformation of *Escherichia coli* with plasmids. *J. Mol. Biol.*, **166**, 557–580.
2. Hanck,T., Gerwin,N. and Fritz,H.-J. (1989) Nucleotide sequence of the *dcm* locus of *Escherichia coli* K12. *Nucleic Acids Res.*, **17**, 5844.
3. Zhang,B., Tao,T., Wilson,G.G. and Blumenthal,R.M. (1993) The M.AluI DNA-(cytosine C5)-methyltransferase has an unusually large, partially dispensable, variable region. *Nucleic Acids Res.*, **21**, 905–911.
4. Taron,C.H., Van Cott,E.M., Wilson,G.G., Moran,L.S., Slatko,B.E., Hornstra,L.J., Benner,J.S., Kucera,R.B. and Guthrie,E.P. (1995) Cloning and expression of the NaeI restriction endonuclease-encoding gene and sequence analysis of the NaeI restriction-modification system. *Gene*, **155**, 19–25.
5. Choi,S. and Leach,J.E. (1994) Identification of the XorII methyltransferase gene and a *vsr* homolog from *Xanthomonas oryzae* pv. *oryzae*. *Mol. Gen. Genet.*, **244**, 383–390.
6. Kiss,A., Posfai,G., Keller,C.C., Venetianer,P. and Roberts,R.J. (1985) Nucleotide sequence of the BsuRI restriction-modification system. *Nucleic Acids Res.*, **13**, 6403–6421.
7. Kulakauskas,S., Barsomian,J.M., Lubys,A., Roberts,R.J. and Wilson,G.G. (1994) Organization and sequence of the HpaII restriction-modification system and adjacent genes. *Gene*, **142**, 9–15.
8. Hennecke,F., Kolmar,H., Bründl,K. and Fritz,H.-J. (1991) The *vsr* gene product of *E. coli* K-12 is a strand- and sequence-specific DNA mismatch endonuclease. *Nature*, **353**, 776–778.
9. Arnold,C. and Hodgson,I.J. (1991) Vectorette PCR: a novel approach to genomic walking. *PCR Methods Appl.*, **1**, 39–42.
10. Rudolph,C., Kunz,C., Parisi,S., Lehmann,E., Hartsuiker,E., Fartmann,B., Kramer,W., Kohli,J. and Fleck,O. (1999) The *msh2* gene of *Schizosaccharomyces pombe* is involved in mismatch repair, mating-type switching, and meiotic chromosome organization. *Mol. Cell. Biol.*, **19**, 241–250.
11. Eisen,J.A. (1998) A phylogenomic study of the MutS family of proteins. *Nucleic Acids Res.*, **26**, 4291–4300.