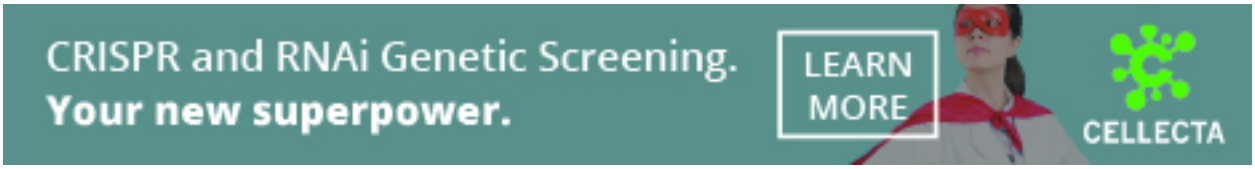
Restriction-site PCR: a direct method of unknown sequence retrieval adjacent to a known locus by using universal primers.

G Sarkar, R T Turner and M E Bolander

Genome Res. 1993 2: 318-322
Access the most recent version at doi:10.1101/gr.2.4.318



Restriction-site PCR: A Direct Method of Unknown Sequence Retrieval Adjacent to a Known Locus by Using Universal Primers

Gobinda Sarkar, Russell T. Turner, and Mark E. Bolander

Department of Orthopedic Research, Mayo Clinic, Rochester, Minnesota 55905

Fast acquisition of unknown nucleotide sequences around a known sequence has important implication in molecular biology, especially in genome mapping. We have developed a method, termed restriction site polymerase chain reaction (RS-PCR), that utilizes specially designed primers that recognize, anneal, and sustain PCR. These primers, termed restriction site oligonucleotides (oligonucleotide primers specific for a given restriction enzyme recognition sequence or RSOs), could be generated corresponding to any restriction enzyme irrespective of the length of the recognition site and used as PCR primers corresponding to the unknown region of a DNA segment. In this method a first round of PCR is carried out in different tubes with a set of RSOs and a primer specific to the known region. A second round of PCR is then performed on the products of the first PCR with the same RSOs and another specific primer internal to the first one. Subsequently, the products of the last round of PCR are transcribed with an appropriate RNA polymerase and sequenced with a reverse transcriptase with an end-labeled specific primer internal to the second specific PCR primer. To demonstrate the applicability of RS-PCR in retrieving unknown sequences around a known sequence, we have used a set of four RSOs and three specific primers representing the known sequence and have successfully obtained hitherto unknown factor IX sequences (12 of 12 times) from three species starting from genomic DNA. The sequences obtained indicate the presence of a conserved stretch of 20 nucleotides in the 3′ noncoding region of the factor IX gene. Besides being useful in retrieving adjacent unknown sequences, these universal primers used in RS-PCR could be useful in a low-cost genome mapping project because only a few should be necessary and the same primers could be used for mapping any genome.

PCR[1] has produced a tremendous impact on the advancement of molecular biology and genetics. PCR requires that a pair of primers be generated from known sequences. However, in many cases, sequence may be available only from one end of a DNA segment. Several methods have been reported to overcome this problem;[2–4] however, these methods are quite complex and may not be suitable for routine applications.

Inverse PCR is the first method that reported successful acquisition of unknown sequences from a known region.[2] The method employs a strategy in which several restriction enzymes are used to generate a suitable fragment encompassing the known region. The segment is then circularized by intramolecular ligation and used as a template for PCR with divergent primers created from the known region. The strategy of the method is unique; however, the requirement of multiple restriction enzyme digestions followed by multiple ligations (before PCR could be carried out) render it a slow and somewhat expensive approach. The method described by Lagerstrom et al.[4] also has a requirement of multiple restriction enzyme digestions and ligation of an engineered double-stranded primer before PCR. Although the restriction and ligation reactions are carried out simultaneously in this method, the requirement of extension reaction, immobilization of the extended product, two rounds of PCR, and purification of template prior to sequencing render it cumbersome and time consuming as well. The walking PCR[3] is another option for unknown sequence retrieval, but the requirement of oligomer–extension assay followed by identification and gel purification of the desired band prior to sequencing could compromise the applicability of the method severely.

Restriction site polymerase chain reaction (RS-PCR) is a direct method that rapidly retrieves unknown sequences adjoining a known sequence. RS-PCR requires primers that would recognize a given restriction enzyme recognition site, anneal to it, and promote DNA synthesis in a PCR. To retrieve an unknown sequence, PCR is carried out with a primer corresponding to the known sequence and a restriction site oligonucleotide (RSO) that would serve as a primer corresponding to the other end. A nested PCR[5] is carried out on a small aliquot of the first PCR using an internal primer from the known sequence and the same RSO. The product of this PCR is then directly sequenced by genomic amplification with transcript sequencing (GAWTS)[6] with another internal primer from the known sequence. This approach was used to generate a hitherto unknown 3′ end of the factor IX gene sequences from cow, pig, and rat (and dog as a control). Four RSOs correspond-

2:318–322©1993 by Cold Spring Harbor Laboratory ISSN 1054-9803/93 $3.00

ing to *Eco*RI, *Bam*HI, *Taq*I, and *Sau*3A restriction sites were used in combination with three primers from the known sequence. In every case (12 of 12), the desired sequences (known and unknown) could be obtained with each of the RSOs. The method promises to have application in many contexts, especially in genome mapping.

MATERIALS AND METHODS

Oligonucleotides were synthesized in an Applied Biosystems automated DNA synthesizer. PCR and direct sequencing was carried out essentially as described.[7] AmpliTaq was obtained from Perkin-Elmer Cetus.

RS-PCR

Two picomoles of the primers from the known sequence and 20 pmoles of the RSOs were used for a 20-μl PCR in all steps. The PCR mixture contained (in a total volume of 20 μl) 10 mM Tris-HCl (pH 8.3), 1.5 mM $MgCl_2$, 200 μM dNTPs, 200 ng of genomic DNA, two primers (one from the known sequence and an RSO), and 0.5 units of AmpliTaq. The PCR condition was 1 min at 94°C, 2 min at 50°C, and 3 min at 72° for 30 cycles, and another 10 min at 72°C after the last cycle. The nested PCR was performed the same way except that a second primer (from known end) internal to the first known primer was used and 1 μl of the product of the first PCR was used as template.

Direct Sequencing of the PCR Product

Three microliters of the product of the nested PCR was transcribed with the appropriate phage polymerase in a total volume of 20 μl. Two microliters of this transcription reaction was used as template for sequencing with an internal end-labeled primer.[7]

RESULTS

Strategy for the Design of the RSOs

A novel strategy was developed for synthesizing primers that would (1) recognize any restriction enzyme recognition sequence irrespective of length (4 or 6 base cutters); (2) remain stably annealed to the template during PCR; and (3) promote direct sequencing of the amplified product without purification. Restriction site sequences are ideal candidates for generating such oligonucleotides (or RSOs, see Fig. 1b) representing the unknown end of a target because (1) they are omnipresent across organisms; (2) they are highly repeated in a given organism; therefore, the chances of finding a site within the range of PCR is not an issue; and (3) usually, there are no repeat sequences in them. In practice, the useful RSOs should come from the 4- and 6-base cutters (longer recognition sequences may not be very useful for RS-PCR because the chances of finding one within the PCR range is relatively small). A primer based only on restriction recognition sequence is not useful for PCR. Therefore, an anchor sequence was attached 5′ to a restriction site sequence. A completely redundant sequence of 10-base length was used for this purpose. Finally, a phage promoter sequence was introduced 5′ to the anchor sequence to allow direct sequencing of the amplified product. Each RSO had the general structure T7–phage promoter sequence–NNNNNNNNN–restriction-site sequence (N = A, T, G, or C; Fig. 1a). In total, four such RSOs were generated corresponding to *Eco*RI, *Bam*HI, *Sau*3A, and *Taq*I, restriction enzyme sites.

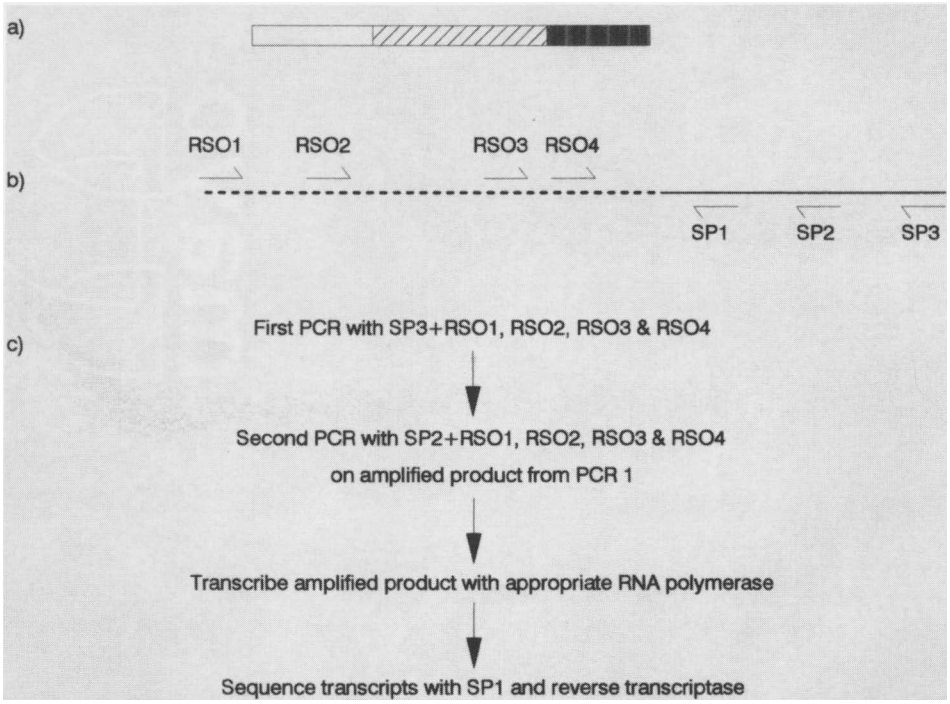


FIGURE 1 Schematic representation of the various components and strategy employed in RS-PCR (*a*) Structure of an RSO: the shaded box represents a restriction enzyme site, the hatched box is the anchor, and the open box refers to a phage promoter sequence. (*b*) RSO1–RSO4 are four specific RSOs; SP1–SP3 are specific primers. Solid lines and broken lines represent known and unknown sequences, respectively. (*c*) Schematic illustration of the method.

Design of Specific Primers from the Known Sequence

Specific primers (Fig. 1b) representing the known portion of the factor IX gene were designed from human sequences that are mostly conserved across species.[8] RS-PCR was used to acquire factor IX sequences from the 3′ end of the gene from cow, dog, pig, and rat genomic DNA (a much more complex starting material than any clone). This part of the gene was selected because sequence at the carboxyl end is mostly unknown in these species (except dog, which served as control); therefore, a sequencing gel obtained through RS-PCR should show a continuum of known and unknown sequences.

Strategy of the RS-PCR

Figure 1 illustrates the strategy, the components, and the steps that are involved in RS-PCR-mediated adjacent sequence retrieval. RSO1–RSO4 (arbitrarily corresponding to the *Bam*HI, *Eco*RI, *Sau*3A,

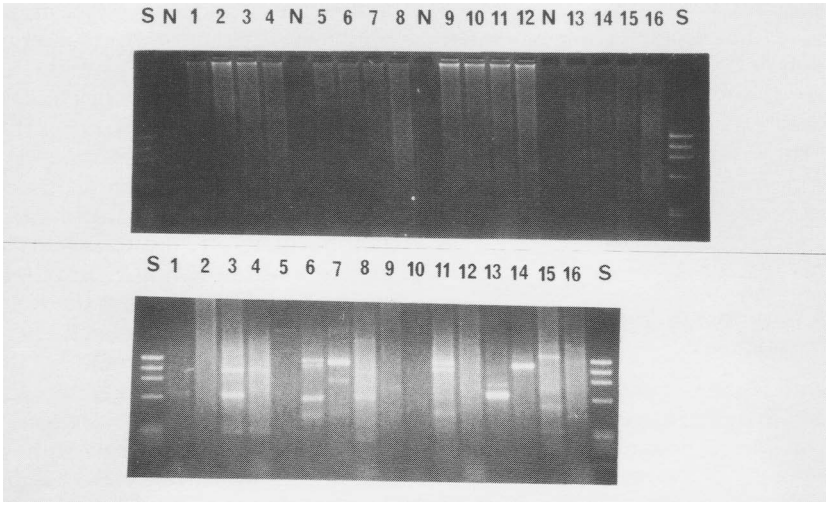


FIGURE 2 Agarose gel electrophoresis of the products after RS-PCR by genomic DNA from cow (lanes *1–4*), dog (lanes *5–8*), pig (lanes *9–12*), and rat (lanes *13–16*). (*Top*) First PCR with SP3 (5′-GCTTTAGTTCTTCAGTACCT-3′) and the RSOs corresponding to the *Eco*RI (lanes *1,5,9,13*), *Bam*HI (lanes *2,6,10,14*), *Sau*3A (lanes *3,7,11,15*), and *Taq*I (lanes *4,8,12,16*) restriction enzymes. (*Bottom*) Second PCR with SP2 (5′-ATCTATAACAACATGTTCTGTGC-3′) and the same RSOs on an aliquot of the first PCR from tubes 1–16. Lanes in the two panels are complementary, e.g., lane *1* (*bottom*) represents second PCR on amplified product from lane *1* of the *top* panel and used the same RSO (*ECO*RI). (Lane *S*) Size standards obtained by digesting *Phi*X174 DNA by *Hae*III; (lane *N*) no DNA added. The gel (1.5%) was stained with ethidium bromide.

and *Taq*I sites) are placed at random. SP1–SP3 are specific primers designed from the known portion of a segment of interest. The first PCR involves PCR with the outermost specific primer, SP3, and RSOs on genomic DNA. A small aliquot of the product of the first PCR is then used as template for a nested PCR with primer SP2 and the same RSOs. A small aliquot of the amplified DNA after nested PCR is transcribed with an appropriate RNA polymerase and directly sequenced by GAWTS[6] with an end-labeled specific primer, SP1. No purification is involved in any step of the method.

Acquisition of Unknown Adjacent Sequences by RS-PCR

The DNA sequence of the 3′ carboxyl end of the factor IX gene from cow, pig, and rat is unknown. Therefore, a modest goal was undertaken to retrieve these sequences starting from known sequences to demonstrate the feasibility of RS-PCR. Three specific primers (corresponding to SP1–SP3 in Fig. 1) were designed from human sequences that are conserved in these species. Dog was used as a control because the coding sequence (and beyond) is completely known.[9] The first PCR involved four reactions with the outermost primer SP3 and the four RSOs on genomic DNA. This step of PCR did not produce any visible band with ethidium bromide stain (Fig. 2, top) but produced a faint smear across a lane. An aliquot of this material was then used for a nested PCR with SP2 and the same RSOs. Bands were visible after this amplification (Fig. 2, bottom). Twelve of these amplified products (lanes 1–3, and 4 for cow; lanes 6–8 for dog; lanes 9, 11, and 12 for pig; and lanes 13–15 for rat; Fig. 2, bottom) were directly sequenced with SP1 by GAWTS, and all of them produced good quality sequences. Approximately 175 bp of sequence information was obtained with this primer in one 2.5-hr run. A representative sample of the sequencing reactions for each of these species is shown in Figure 3. The sequences display a continuum of known and hitherto unknown nucleotides. For each species, identical sequences were obtained with three different RSOs; for example, cow and pig sequences were obtained with *Sau*3A, *Eco*RI, and *Bam*HI RSOs, whereas dog and rat sequences were obtained with *Taq*I, *Eco*RI, and *Bam*HI RSOs. PCR products with *Taq*I RSO for cow and pig, as well as PCR products with *Sau*3A RSO for dog and rat, were not followed up for sequencing because distinct bands were not visible in these lanes, even after nested PCR. This is possible because cow and pig may not have a *Taq*I site at this locus within the range of PCR. Also, dog and rat may not have a *Sau*3 site at this locus. This is exactly why a set of RSOs should be used. It should also be noted that RSOs generated from 4- and 6-base cutting sequences should work with equal efficiency.

Figure 4 presents an alignment of the sequences obtained by RS-PCR for these species. Both known and the last eight amino acids (in bold) were unknown for

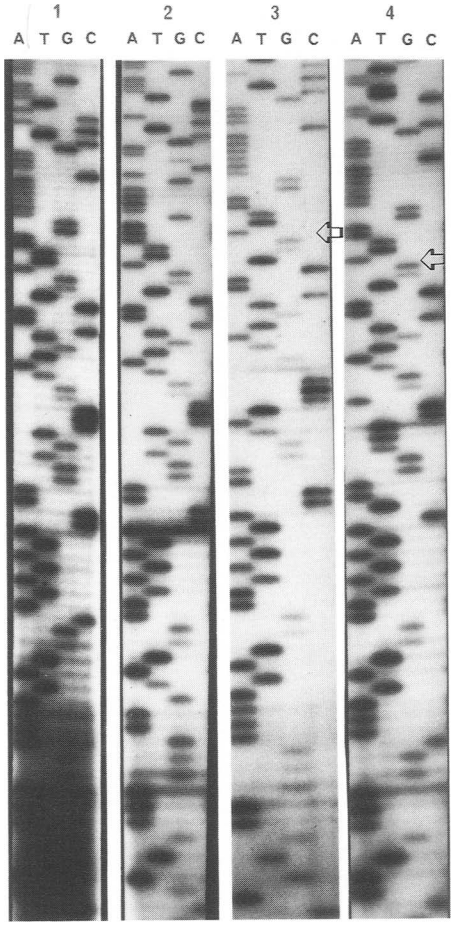


FIGURE 3 A sample of the sequences obtained by RS-PCR (Panel *1*) Cow; (panel *2*) dog; (panel *3*) pig; (panel *4*) rat. The sequencing was accomplished with GAWTS as described.[6] Briefly, 3 μl of the amplified products from the second PCR (Fig. 2, *bottom*) was directly transcribed with T7 RNA polymerase; 2 μl of the transcription reaction was then directly used for sequencing with reverse transcriptase and SP1 (5′-ATTATTAGCTGGGGTGA-3′). The dog cDNA sequence is known.[9] An arrow indicates a junction of known and unknown sequences. All of the cow sequence is novel.

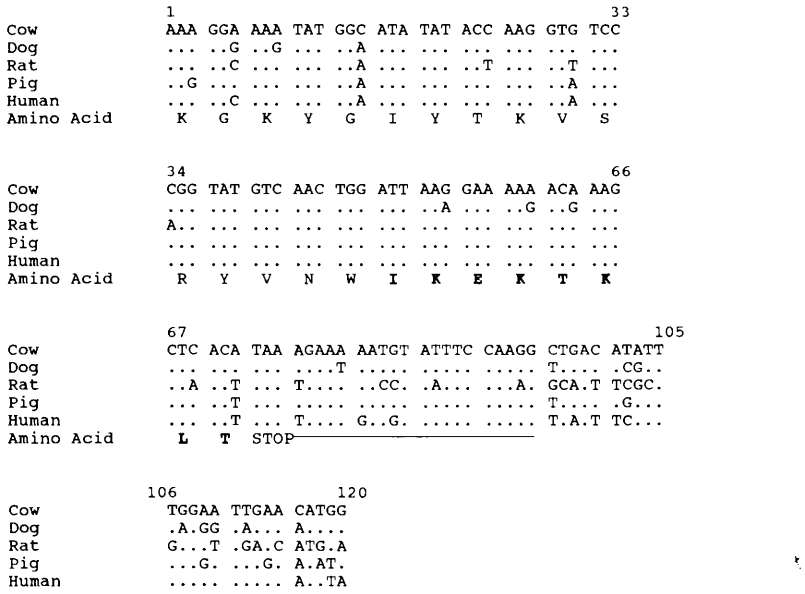


FIGURE 4 Alignment of sequences obtained by RS-PCR for factor IX gene in different species. The human sequence has been included for comparison. Single-letter amino acid codes have been used. Amino acids in bold letters were unknown for pig and rat, and the dog factor IX sequence is completely known. The complete amino acid sequence for cow is known; however, the nucleic acid sequences presented are novel. The 20 highly conserved nucleotides following the stop codon are underlined. A dot indicates identity to the uppermost sequence.

pig and rat[8] and all the nucleic acid sequence presented here for cow factor IX is novel, although the amino acid sequence is known.[10] The dog factor IX sequence obtained matches completely with the published sequence.[9] The human sequence[11] has been included for comparison. It is clear that all of these amino acids are completely conserved in these species. It is interesting to note that all of these species use the same codon for translation termination. In addition, it appears that 20 nucleotides after the stop codon are highly conserved across these species. Sequences beyond what has been presented here could have been obtained but not pursued considering the immediate goal of the investigation.

DISCUSSION

The data presented indicate strongly that RS-PCR can be used to retrieve adjacent unknown sequence from a known sequence. Also, to our knowledge, this is the first report on the successful use of short sequences as PCR primers. The method is simple because no purification step is involved. Repetitive sequences in the human genome have been used to generate generic primers for PCR, for example, Alu–PCR.[12] However, the usefulness of primers corresponding to Alu sequences are limited because they are not present in all organisms and their spacing in the genome is often beyond the range of optimal PCR. On the other hand, restriction site sequences are omnipresent across organisms and are repeated so frequently that in theory, there will always be a site (starting from any given site) within the range of PCR.

Although there are hundreds of different restriction enzyme sites, it may not be necessary to generate primers specific for all of the enzymes. Data presented here indicate that four RSOs were adequate (correct factor IX sequences were obtained with all of the RSOs used) to consistently generate desired sequences from multiple species. However, it is possible that additional RSOs will be required for GC-rich sequences.

Use of RSOs could theoretically present a problem: Because they are not unique sequences, they would not provide much specificity in the PCR. In our hands, this concern was not effective, because in RS-PCR, adequate specificity is generated from the known end by doing a nested PCR and using an internal primer in the sequencing reaction. Also, although one round of nested PCR has produced satisfactory results in this investigation, multiple rounds of nested PCR may be necessary in other situations.

The strategy employed in generating RSOs offers a new approach to primer design. For example, primers can now be generated corresponding to the TATA box or Shine–Dalgarno sequence,[13] or a consensus splice site. The stringency of these primers can be controlled if necessary, for example, an RSO with 5′-GGATCCT-3′ at the 3′ end should prime DNA synthesis only from the *Bam*HI sites that are followed by a T residue. Also, if desired, suitable restriction enzyme sites may be appended 5′ to an RSO if cloning of the amplified product is the goal. Moreover, the inclusion of the phage promoter is not necessary if other direct sequencing protocols are to be followed.

Mapping initiatives of the human genome and other genomes will continue to be of prime importance to scientists throughout the world. However, simple as well as inexpensive methods need to be formulated to complete these efforts within a reasonable time frame and budget. In this regard RS-PCR may be valuable in genome mapping because only a handful of universal primers should be necessary.

Because of its simplicity and robustness, RS-PCR should find application in many areas of research. The determination of integration sites of different viruses and phages into the host genome, the acquisition of upstream or downstream sequences corresponding to a gene, and the location(s) of a transgene are immediate examples among many.

In the course of demonstrating the feasibility of RS-PCR, we have defined hitherto unknown amino acids at the carboxyl terminus of rat and pig factor IX genes. In addition, we have detected a stretch of 20 nucleotides at the 3′ noncoding end of this gene that is highly conserved in the species examined. Conservation of nucleic acid sequences usually suggests evolutionary selection pressure. Whether this conserved stretch of nucleotides exert any role in the regulation of factor IX gene expression remains to be defined.

ACKNOWLEDGMENTS

We thank Dr. M. Subramaniam for a generous supply of the oligonucleotides and Dr. G. Sandhu for help with the fig-

ures. We also thank Drs. S. Kapelner and B.C. Kline for their support, and Ms. Carolyn Blankenship, Ms. Sandy Halloran, and Ms. Sara Erickson for typing the manuscript. G.S. was supported by a Mayo Thompson Fellowship.

REFERENCES

1. Mullis, K.B. and F.A. Faloona. 1987. Specific synthesis of DNA *in vitro* via a polymerase-catalyzed chain reaction. *Methods Enzymol.* **155:** 335–350.
2. Triglia, T., M.G. Peterson, and D.J. Kemp. 1988. A procedure for in vitro amplification of DNA segments that lie outside the boundaries of known sequence. *Nucleic Acids Res.* **16:** 8186.
3. Parker, J.D., P.S. Rabinovitch, and G.C. Burmer. 1991. Targeted gene walking polymerase chain reaction. *Nucleic Acids Res.* **19:** 3055–3060.
4. Lagerstrom, M., J. Parik, H. Malmgren, J. Stewart, U. Patterson, and U. Landegren. 1991. Capture PCR: Efficient amplification of DNA fragments adjacent to a known sequence in human and YAC DNA. *PCR Methods Applic.* **1:** 111–119.
5. Haqqi, T.M., G. Sarkar, C.S. David, and S.S. Sommer. 1988. Specific amplification with PCR of a refractory segment of genomic DNA. *Nucleic Acids Res.* **16:** 11844.
6. Stoflet, E.S., D.D. Koeberl, G. Sarkar, and S.S. Sommer. 1988. Genomic amplification with transcript sequencing. *Science* **239:** 491–494.
7. Sarkar, G. and S.S. Sommer. 1989. Access to a messenger RNA sequence or its protein product is not limited by tissue or species specification. *Science* **244:** 331–334.
8. Sarkar, G., D.D. Koeberl, and S.S. Sommer. 1990. Direct sequencing of the activation peptide and the catalytic domain of the factor IX gene in six species. *Genomics* **6:** 133–143.
9. Evans, J.P., H.H. Watzke, H.L. Ware, D.W. Stafford, and K.A. High. 1989. Molecular cloning of a cDNA encoding canine factor IX. *Blood* **74:** 207–212.
10. Katayama, K., L.H. Ericsson, D.L. Enfield, K.A. Walsh, H. Neurath, E.W. Davie, and K. Titani. 1979. Comparison of amino acid sequence of bovine coagulation factor IX (Christmas Factor) with that of other vitamin-k-dependent plasma proteins. *Proc. Natl. Acad. Sci.* **76:** 4990–4994.
11. Yoshitake, S., G.G. Schach, D.C. Foster, E.W. Davie, and K. Kurachi. 1985. Nucleotide sequence of the gene for human factor IX (antihemophilic factor B). *Biochemistry* **24:** 3736–3750.
12. Nelson, D.L., A. Ballabio, M.F. Victoria, M. Pieretti, R.D. Bies, R.A. Gibbs, J.A. Maley, A.C. Chinault, T.D. Webster, and C.T. Caskey. 1991. Alu-primed polymerase chain reaction for regional assignment of 110 yeast artificial chromosome clones from the human X chromosome: Identification of clones associated with a disease locus. *Proc. Natl. Acad. Sci.* **88:** 6157–6161.
13. Shine, J. and L. Dalgarno. 1974. The 3'-terminal sequence of *E. coli* 16S ribosomal RNA: Complementary to nonsense triplets and ribosome binding sites. *Proc. Natl. Acad. Sci.* **71:** 1342–1346.

Received December 14, 1992; accepted in revised form February 14, 1993.