Further complexity of the human *SOX* gene family revealed by the combined use of highly degenerate primers and nested PCR

Frederic Cremazy, Stephan Soullier, Philippe Berta*, Philippe Jay

Human Molecular Genetic Group, Institut de Génétique Humaine, UPR CNRS 1142, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France

Received 7 September 1998

Abstract SOX proteins contain a conserved HMG-related DNA-binding domain. They fulfil essential functions during the development of animals. Mutations in several *SOX* genes have been implicated in human diseases. We present here a new set of PCR primers designed to amplify a broad range of *SOX* HMG-box sequences. These primers facilitated the cloning of several new *SOX* HMG boxes from human genomic DNA, revealing unexpected complexity of the *SOX* gene family.

© 1998 Federation of European Biochemical Societies.

Key words: SOX; SRY-related; HMG; Nested polymerase chain reaction; Multigene family; Human

In mammals, male sex determination is initiated by the *SRY* gene located on the Y chromosome. The corresponding protein, SRY, contains a characteristic 79 amino acid long DNA-binding domain called the HMG box (high mobility group) because of its homology with two regions of the HMG1 and HMG2 proteins [1]. Numerous other genes have been identified that encode proteins containing an HMG domain related to that of SRY. They have been called *SOX* (*SRY* box-related) when their HMG domains shared at least 50% identical residues with that of SRY [2]. In animals, *SOX* genes appear to regulate essential aspects of development [3–8]. Several of them have also been implicated in human disease, such as *SRY* in sex reversal [9], *SOX9* in campomelic dysplasia [10] and *SOX10* in Waardenburg-Hirschsprung disease [11]. As a first step towards a better understanding of the importance of the *SOX* gene family in development and disease, we and others have started to establish a catalog of *SOX* genes by screening the human and mouse genomes for *SRY*-related sequences. So far, 22 different *SOX* genes have been identified in mammals using degenerate PCR and low-stringency hybridization-based approaches [12–20]. Comparison of these 22 HMG domains showed that they are clustered within several distinct phylogenetic sub-groups [14,21]. We recently carried out an extensive phylogenetic study of the HMG-domain protein family (Soullier and Berta, unpublished results). From this analysis, it became possible to design new, highly degenerate primers capable of amplifying a broad spectrum of *SOX* HMG sequences with high specificity. Here, we show that these new primers facilitate the cloning of a wide range of as yet undetected *SOX*-box sequences from human genomic DNA. Our results validate these PCR primers as a powerful new tool for the identification and cataloguing of *SOX* genes.

PCR primers were designed using a multiple alignment of HMG-domain sequences representative of the SRY/SOX protein family. Primers were chosen to specifically amplify *SRY*/*SOX* sequences and were highly degenerate: for each position, an inosine nucleotide was used as soon as more than two different bases were represented in the alignment. Two pairs of primers were used in order to allow low-stringency PCR reactions coupled with a nested PCR approach. The primers were shown to specifically amplify *SOX* sequences from human genomic DNA (not shown). Pairs of primers were tested in all possible combinations for the two consecutive PCR amplification steps. The best results were obtained with the primers P5-1+P3-1 and P5-2+P3-1 for the first and second amplification steps respectively (Fig. 1A,B). We subsequently amplified plasmid DNA containing the human *SRY* [1] or *SOX11* [22] cDNAs to test whether the primers would be able to amplify strongly divergent *SOX* sequences. These two sequences were chosen because SRY and SOX11 HMG domains are amongst the most divergent, sharing only 59% identical residues. Although the yields of PCR products were different, both templates produced clearly detectable products after the two rounds of amplification (Fig. 1C). Human genomic DNA was then amplified using the optimal conditions and the PCR product was cloned into pUC18 plasmid. DNA was purified from 87 clones and analyzed by Southern blot using a mixed probe containing the *SRY* and *SOX11* HMG boxes. All 48 clones that displayed a positive signal by Southern blotting were sequenced using the universal primers present in the vector. Importantly, all clones containing an insert were positive by hybridization, demonstrating the high specificity of the nested PCR approach used. The identity of each sequence was then determined using BLAST analysis [23]. Among 48 clones sequenced, 18 (38%) corresponded to known *SOX* sequences. These sequences represented essentially SOX12 and SOX4, raising the possibility of a biased amplification from genomic DNA. Comparison of the 30 unknown sequences showed that they represented six distinct new *SOX* HMG-box sequences. After translation of the nucleotide sequences, it appeared that one of the new peptide sequences was identical to the mouse Sox14 HMG domain (accession number: Z18963) and we therefore named the corresponding nucleotide sequence *SOX14*. Four other sequences showed no identity with any previously described *Sox* sequence and they were labelled *SOX25* to *SOX28*. The last of the new *SOX* HMG-box sequences contained a frameshift caused by a 2-bp deletion after 48 residues (as numbered in Fig. 2). This was named *SOX29* and is likely to be a pseudogene. The *SOX29* HMG-box nucleotide sequence is 92% identical to that of *SOX5*. Since a SOX5 pseudogene, containing no significant open reading frame, was already reported on chromosome 8q21.1 [24], it might be possible that *SOX29* actually is this *SOX5* pseudogene. It can also not be formally excluded that the frameshift observed in the *SOX29* sequence results from a PCR artefact. Fig. 2 shows a comparison of the

*Corresponding author. Fax: (33) 499.61.99.01.
E-mail: berta@igh.cnrs.fr

PII: S0014-5793(98)01294-0

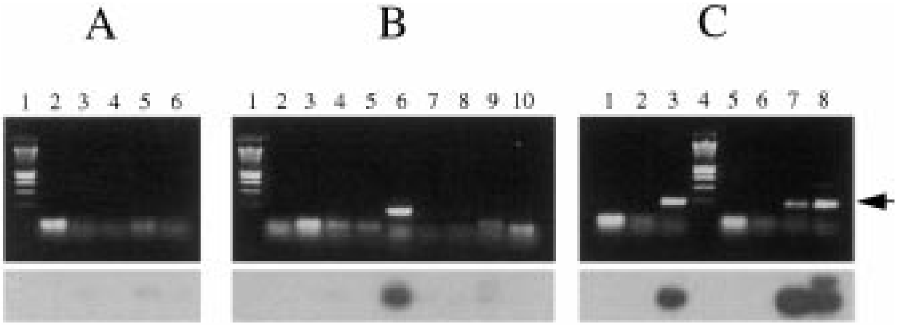

Fig. 1. Characterization of the primers used and amplification of *SOX* sequences from human genomic DNA. A and B: Determination of the best primer combination for the two PCR amplification steps. A: First amplification from human genomic DNA. 1 = molecular weight marker (Smart ladder, Eurogentec) 2 = H_2O; 3 = P5-1+P3-1; 4 = P5-1+P3-2; 5 = P5-2+P3-1; 6 = P5-2+P3-2. B: Second, nested amplification using the first PCR product as template. 1 = molecular weight marker (Smart ladder, Eurogentec); 2 = H_2O; 3 = H_2O from A2; 4 = P5-1+P3-1 from A3; 5 = P5-1+P3-2 from A3; 6 = P5-2+P3-1 from A3; 7 = P5-2+P3-2 from A3; 8 = P5-2+P3-2 from A4; 9 = P5-2+P3-2 from A5; 10 = P5-2+P3-2 from A6. Arrow shows the amplification product. A and B: Lower panels show autoradiograms from Southern hybridization using a mix of the *SRY* and *SOX11* HMG boxes as a probe. C: Amplification from diverging templates using the optimal primer combination. 1 = H_2O; 2 = SRY (first amplification); 3 = SOX11 (first amplification); 4 = molecular weight marker; 5 = H_2O; 6 = H_2O from 1; 7 = SRY (second amplification); 8 = SOX11 (second amplification). The double band seen in the autoradiography of lane 8 likely represents the products from both the first and second amplification steps. PCR cycling conditions were: 35 cycles, each with 1 min, 94°C; 1 min, 50°C; 30 s, 70°C in a 50 µl reaction mix containing 10 mM Tris-HCl pH 8.3; 1.5 mM $MgCl_2$; 50 mM KCl; 400 µM each dNTP; 3 µM each primer; 1.25 U Taq polymerase (Boehringer Mannheim). Approximately 1 ng and 500 ng of template were used for respectively plasmid and genomic DNA amplifications. Sequences of the primers are: P3-1 5′-GG(C,T)(C,T)(G,T)(A,G)TA(C,T)TT(A,G)TA(A,G)T(C,T)(G,C)GG-3′; P3-2 5′-T(T,C)IGG(A,G)T(A,G)-T(A,G)I(T,C)(T,C)(A,C)(T,G)I(T,C)AIGTG-3′; P5-1 5′-(A,G)T(G,C)(A,C)(A,G)(A,G)(A,C)G(G,C)CC(A,C)ATGAA(C,T)GC-3′; P5-2 5′-CC(A,C)ATGAA(C,T)GC(G,C)TT(C,T)AT(G,C)GT(G,C)TGG-3′.

SOX14, SOX25–28, SRY and SOX11 HMG-domain sequences. Besides mediating DNA binding, one of the proposed functions of the HMG domain is to target the SOX protein to the nucleus of cells. In SRY, the first 20 residues of the HMG motif constitute a bipartite nuclear localization signal [25]. The residues involved in this putative NLS and present in the amplified sequences of SOX14 and SOX25–28 are highly conserved among all these sequences (Fig. 2). Furthermore, the variations of the SRY NLS sequence found in SOX27 have already been described for example in mouse Sry [12].

The new SOX protein sequences were included in a phylogenetic tree representing all SOX HMG domains described so far in human and mouse (Fig. 3). SOX14 as well as SOX25–28 did not cluster together but were instead found scattered throughout the most abundant group of SOX proteins, originally described as group B [14]. We clearly demonstrate the capability of our PCR primers to amplify sequences from other groups, including *SRY*, group A; *SOX11*, group C; *SOX9*, group E (Fig. 1C and data not shown). Using the same primer set, we could also amplify *SOX4, 5, 7, 8, 9, 10, 17* and *18* from various cDNA sources (Wafaa Takash, unpublished results). In consequence, although there might be a bias leading to preferential amplification of group B sequences, it is likely that group B is much more important than previously expected and still contains numerous sequences to be identified. Sequences within group B cluster into several sub-groups. We propose to distinguish between sub-groups B1 (SOX12, 15, 16, 20, 26), B2 (SOX1, 2, 3, 14, 25, 28) and B3 (SOX27), sub-group B2 representing the original group B. All sequences within group B share at least 75% identical residues,

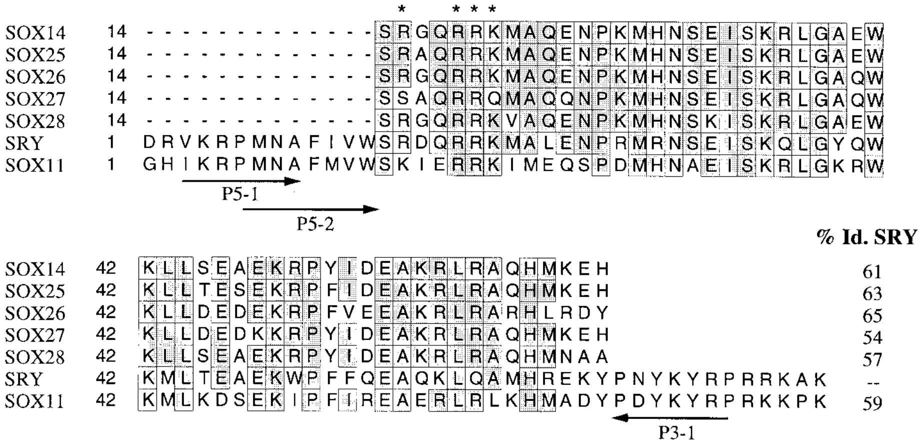
* * * *
SOX14 14 - - - - - - - - - - - - - - S R G Q R R K M A Q E N P K M H N S E I S K R L G A E W
SOX25 14 - - - - - - - - - - - - - - S R A Q R R K M A Q E N P K M H N S E I S K R L G A E W
SOX26 14 - - - - - - - - - - - - - - S R G Q R R K M A Q E N P K M H N S E I S K R L G A Q W
SOX27 14 - - - - - - - - - - - - - - S S A Q R R Q M A Q Q N P K M H N S E I S K R L G A Q W
SOX28 14 - - - - - - - - - - - - - - S R G Q R R K V A Q E N P K M H N S K I S K R L G A E W
SRY 1 D R V K R P M N A F I V W S R D Q R R K M A L E N P R M R N S E I S K Q L G Y Q W
SOX11 1 G H I K R P M N A F M V W S K I E R R K I M E Q S P D M H N A E I S K R L G K R W
P5-1
P5-2
% Id. SRY
SOX14 42 K L L S E A E K R P Y I D E A K R L R A Q H M K E H 61
SOX25 42 K L L T E S E K R P F I D E A K R L R A Q H M K E H 63
SOX26 42 K L L D E D E K R P F V E E A K R L R A R H L R D Y 65
SOX27 42 K L L D E D K K R P Y I D E A K R L R A Q H M K E H 54
SOX28 42 K L L S E A E K R P Y I D E A K R L R A Q H M N A A 57
SRY 42 K M L T E A E K W P F F Q E A Q K L Q A M H R E K Y P N Y K Y R P R R K A K --
SOX11 42 K M L K D S E K I P F I R E A E R L R L K H M A D Y P D Y K Y R P R K K P K 59
P3-1

Fig. 2. Multiple alignment of the SRY, SOX11, SOX14 and SOX25–28 HMG-domain peptide sequences obtained using the CLUSTALW software. Identical residues are shadowed. The positions of the primers used are indicated by arrows. Residues are numbered on the left side and the conservation (% identity of amino acids) of each new SOX sequence with the HMG domain of SRY is indicated at the end of each sequence. The residues included in the putative NLS (and present in the amplified sequences) are indicated by asterisks. The sequences of *SOX25, SOX26, SOX14, SOX27, SOX28 and SOX29* are registered in GenBank under the accession numbers AFO32449–AFO32454 respectively.

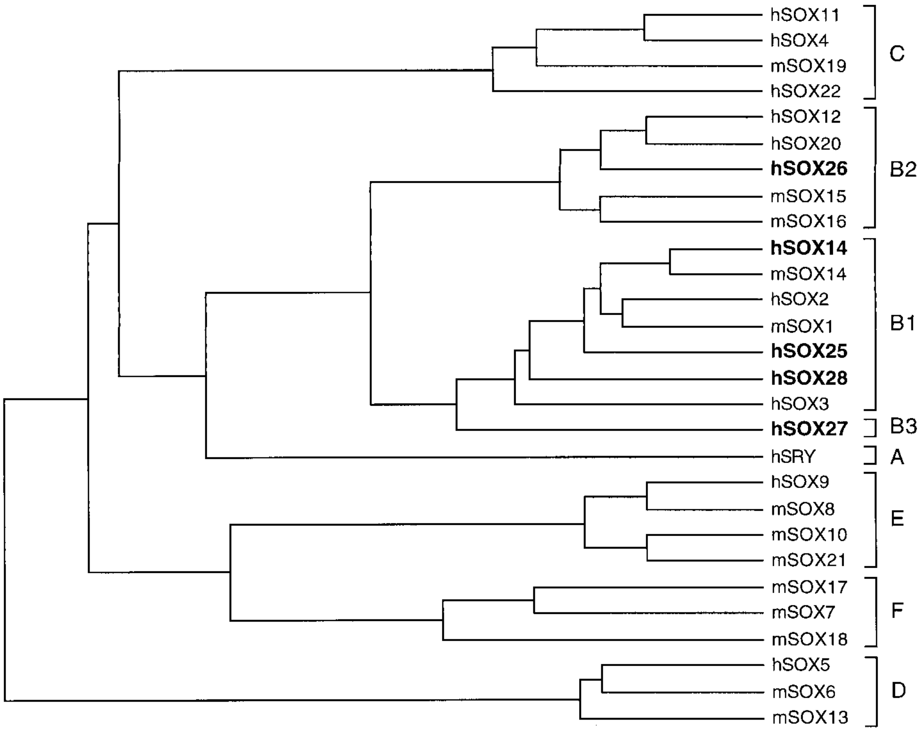

Fig. 3. Dendrogram of the multiple alignment of all known SOX HMG-domain sequences known in mammals, as well as of the human SOX14 and SOX25–28. Alignment was performed with the CLUSTALW software. Mouse sequences were used when the human orthologous sequences were not available. SOX23 [26] and SOX24 [27], cloned uniquely in rainbow trout, are not represented in the dendrogram. New sequences described in this report are highlighted in bold. Upgrade of the groups initially described by Wright et al. [14] is presented on the right of the dendrogram. Note that the initial group B is now split into B1, B2 and B3.

whereas in a given sub-group, identity between members rises to at least 85%.

The present study supports the hypothesis that mammalian genomes contain an unexpectedly large number of *SOX* genes and, possibly, pseudogenes. The identification of these genes is currently limited by the lack of PCR primers able to amplify divergent *SRY*-related sequences. Such strongly degenerate primers are presented here. It is shown that their use, in combination with nested PCR, allows specific amplification of a broad spectrum of *SOX* sequences. Interestingly, all the new peptide sequences presented here are only distantly related to SRY (50–60% identical residues). This suggests that the primers used in this study allow the amplification of a new pool of *SOX* genes, inaccessible to analysis with the previously published sets of primers. The generalization of the use of the primers presented here with genomic DNA as well as with cDNAs isolated from various tissues will probably allow a rapid increase in our knowledge of the complexity and diversity of the *SOX* gene family.

Acknowledgements: This work was supported by CNRS and partly by grants from the Ligue Nationale contre le Cancer and ARC (No. 9106) to P.B. We thank Dr. Nick Barker for corrections to the manuscript.

References

[1] Sinclair, A.H., Berta, P., Palmer, M.S., Hawkins, J.R., Griffiths, B.L., Smith, M.J., Foster, J.W., Frischauf, A.M., Lovell-Badge, R. and Goodfellow, P.N. (1990) Nature 346, 240–244.

[2] Pevny, L.H. and Lovell-Badge, R. (1997) Curr. Opin. Genet. Dev. 7, 338–344.

[3] Koopman, P., Gubbay, J., Vivian, N., Goodfellow, P.N. and Lovell-Badge, R. (1991) Nature 351, 117–121.

[4] Wright, E., Hargrave, M.R., Christiansen, J., Cooper, L., Kun, J., Evans, T., Gandadharan, U., Greenfield, A. and Koopman, P. (1995) Nature Genet. 9, 15–20.

[5] Schilham, M.W., Oosterwegel, M.A., Moerer, P., Ya, J., de Boer, P.A.J., van de Wetering, M., Verbeek, S., Lamers, W.H., Kruisbeek, A.M., Cumano, A. and Clevers, H. (1996) Nature 380, 711–714.

[6] Russell, S.R.H., Sanchez-Soriano, N., Wright, C.R. and Ashburner, M. (1996) Development 122, 3669–3676.

[7] Nambu, P.A. and Nambu, J.R. (1996) Development 122, 3467–3475.

[8] Southard-Smith, E.M., Kos, L. and Pavan, W.J. (1998) Nature Genet. 18, 60–64.

[9] Berta, P., Hawkins, J.R., Sinclair, A.H., Taylor, A., Griffiths, B.L., Goodfellow, P.N. and Fellous, M. (1990) Nature 348, 448–450.

[10] Wagner, T., Wirth, J., Meyer, J., Zabel, B., Held, M., Zimmer, J., Pasantes, J., Bricarelli, F.D., Keutel, J., Hustert, E., Wolf, U., Tommerup, N., Schempp, W. and Scherer, G. (1994) Cell 79, 1111–1120.

[11] Pingault, V., Bondurand, P., Kuhlbrodt, K., Goerich, D.E., Préhu, M.-O., Puliti, A., Herbarth, B., Hermans-Borgmeyer, I., Legius, E., Matthijs, G., Amiel, J., Lyonnet, S., Ceccherini, I., Romeo, G., Smith, J.C., Read, A.P., Wegner, M. and Goossens, M. (1998) Nature Genet. 18, 171–173.

[12] Gubbay, J., Collignon, J., Koopman, P., Capel, B., Economou, A., Münsterberg, A., Vivian, N., Goodfellow, P.N. and Lovell-Badge, R. (1990) Nature 346, 245–346.

[13] Denny, P., Swift, S., Brand, N., Dabhade, N., Barton, P. and Ashworth, A. (1992) Nucleic Acids Res. 20, 2887.

[14] Wright, E.M., Snopek, B. and Koopman, P. (1993) Nucleic Acids Res. 21, 744.

[15] Gozé, C., Poulat, F. and Berta, P. (1993) Nucleic Acids Res. 21, 2943.

[16] van de Wetering, M. and Clevers, H. (1993) Nucleic Acids Res. 21, 1669.

[17] Dunn, T.L., Mynett-Johnson, L., Wright, E.M., Hosking, B.M., Koopman, P.A. and Muscat, G.E.O. (1995) Gene 161, 223–225.

[18] Jay, P., Sahly, I., Gozé, C., Taviaux, S., Poulat, F., Couly, G., Abitbol, M. and Berta, P. (1997) Hum. Mol. Genet. 6, 1069–1077.

[19] Tani, M., Shindo-Okada, N., Hashimoto, Y., Shiroishi, T., Takenoshita, S., Nagamachi, Y. and Yokoda, J. (1997) Genomics 39, 30–37.

[20] Meyer, J., Wirth, J., Held, M., Schempp, W. and Scherer, G. (1996) Cytogenet. Cell Genet. 72, 246–249.

[21] Laudet, V., Stehelin, D. and Clevers, H. (1993) Nucleic Acids Res. 21, 2493–2501.

[22] Jay, P., Gozé, C., Marsollier, C., Taviaux, S., Hardelin, J.-P., Koopman, P. and Berta, P. (1995) Genomics 29, 541–545.

[23] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) J. Mol. Biol. 215, 403–410.

[24] Wunderle, V.M., Critcher, R., Ashworth, A. and Goodfellow, P.N. (1996) Genomics 36, 354–358.

[25] Poulat, F., Girard, F., Chevron, M.-P., Gozé, C., Rebillard, X., Calas, B., Lamb, N. and Berta, P. (1995) J. Cell Biol. 128, 737–748.

[26] Yamashita, A., Susuki, S., Fujitani, K., Kojima, M., Kanda, H., Ito, M., Takamatsu, N., Yamashita, S. and Shiba, T. (1998) Gene 209, 193–200.

[27] Kanda, H., Kojima, M., Miyamoto, N., Ito, M., Takamatsu, N., Yamashita, S. and Shiba, T. (1998) Gene 211, 251–257.