RESEARCH ARTICLE

Remarkable evolutionary conservation of SOX14 orthologues

JELENA POPOVIC and MILENA STEVANOVIC*

Institute of Molecular Genetics and Genetic Engineering, Vojvode Stepe 444a, P.O. Box 23, 11010 Belgrade, Serbia

Abstract

SOX proteins constitute a large family of diverse, well-conserved transcription factors present in vertebrates and invertebrates, and also implicated in control of many developmental processes. Our objectives have been to identify *Sox14* gene of goat (*Capra hircus*), cow (*Bos taurus*) and rat (*Rattus norvegicus*), and to perform comparative analyses and mapping of SOX14 orthologues from numerous vertebrate species. PCR based approach was used to identify *Sox14* of goat, cow and rat, while nucleotide and amino acid sequence alignments and mapping were performed using information currently available in public database. Comparative sequence analysis revealed remarkable identity among *Sox14* orthologues and helped us to identify highly conserved motifs that represent molecular signatures of SOX14 protein that might have structural or functional significance. Further, we determined chromosomal locations of numerous predicted group B *Sox* genes and their neighbouring genes using currently available genome database. In conclusion, our study has not only supported the proposed model of group B *Sox* genes evolution in chicken and mammals, but has also revealed that additional evolutionary events split *Sox* B genes into different chromosomes in some mammals. Mapping data presented in this study could help in refining the understanding of the evolution of group B *Sox* genes in vertebrates.

[Popovic J. and Stevanovic M. 2009 Remarkable evolutionary conservation of SOX14 orthologues. *J. Genet.* **88**, 15–24]

Introduction

SOX proteins belong to the HMG-box superfamily of DNA-binding proteins that display properties of both classical transcriptional factors and architectural components of chromatin (Pevny and Lovell-Badge 1997). SOX transcription factors show diverse tissue-specific expression patterns during early development and they have been implicated in cell-fate decisions in numerous developmental processes (Pevny and Lovell-Badge 1997; Wegner 1999). Mutations in *SOX* genes often result in developmental defects and congenital diseases in humans (Wegner 1999). *Sox* genes have been identified in a broad range of animal taxa, including birds, reptiles, amphibians, fish, insects and nematodes, with at least 30 members recognized in mammals (Pevny and Lovell-Badge 1997; Wegner 1999; Bowles *et al*. 2000).

Based on HMG box homology and intron–exon structure, *SOX/Sox* genes are divided into 10 distinct groups designated from A to J (Bowles *et al*. 2000). The B group of SOX/*Sox* genes is of particular interest, since the members of this group play a major role in neural development. The five intron-less group B *Sox* genes (*Sox1, Sox2, Sox3, Sox14* and *Sox21*) participate in the earliest events of central nervous system (CNS) differentiation in *Drosophila, Xenopus*, chick and mouse (Collignon *et al*. 1996; Uchikawa *et al*. 1999; Hargrave *et al*. 2000a; Kishi *et al*. 2000; McKimmie *et al*. 2005). Based on the sequence analysis and functional studies in vertebrates, the group B *Sox* genes can be further subdivided into sub-group B1 comprising activators (*Sox1, Sox2* and *Sox3*) and sub-group B2 consisting of repressors (*Sox14* and *Sox21*) (Uchikawa *et al*. 1999). Members of sub-group B1 show functional similarity, in particular, in the regulation of the neuronal phenotype (Collignon *et al*. 1996). Sub-group B2 genes (*Sox14* and *Sox21*) are also expressed in the CNS and have been postulated to have a role in the specification of a particular subset of neurons, rather than neuronal development in general (Hargrave *et al*. 2000b)

SOX14/Sox14 gene has been identified in many vertebrate species, including human, mouse, platypus, chicken and fish. *Sox14* expression is restricted to a limited population of neurons in the developing brain and spinal cord of mouse and

*For correspondence. E-mail: stevanov@eunet.rs.

Keywords. comparative analysis; comparative mapping; molecular evolution; *Sox* genes; SOX14 orthologues.

chick embryos and it marks a subset of interneurons at a defined dorsoventral position in the spinal cord (Hargrave *et al*. 2000b). Chicken *Sox14* is also expressed exclusively in the apical-ectodermal ridge of limb-buds, and it has been speculated that the genetic mechanism which defines the dorso–ventral border of the limb activates *Sox14* (Uchikawa *et al*. 1999). The *SOX14/Sox14* gene in human, chicken and mouse have been cloned and characterized, and were found to be substantially conserved (Arsic *et al*. 1998; Hargrave *et al*. 2000a; Kuroiwa *et al*. 2002). Moreover, recently we have shown that mammalian orthologues of the human *SOX14* gene show high sequence identity in their promoter regions (Kovacevic-Grujicic *et al*. 2008).

Based on the comparative mapping of group B SOX/*Sox* genes in chicken, platypus and human, a model has been proposed, suggesting that *Sox*B genes have arisen by duplications, rearrangement and divergence from an ancestral *Sox*B gene, which probably arose initially by retrotransposition from an intron containing *Sox* gene (Kirby *et al*. 2002).

The rapid accumulation of data on the sequences of various vertebrate genomes provides a valuable tool to address issues of gene evolution and orthology. Accordingly, we have identified *Sox14* gene of goat, cow and rat for the first time and performed comparative analyses of SOX14 orthologues from numerous vertebrate species whose sequences are currently available in the public databases. Our aim was to identify conserved motif(s) that might help in defining protein domain(s) important for SOX14 function. To get better insight into genomic organization and evolution of group B *Sox* genes, we have also performed comparative mapping of these genes across various vertebrate species. The mapping data presented might help in improving the understanding of the evolution of *Sox*B genes in vertebrates.

Materials and methods

Sox14 gene of goat (*Capra hircus*), cow (*Bos taurus*) and rat (*Rattus norvegicus*) were amplified by PCR using primers specific for human *SOX14* gene. Forward primer: 5′-ATGTCCAAACCTTCAGACCAC-3′ corresponds to the sequence encoding MSKPSDH motif at the N-terminus, while reverse: 5′-ACATGGCCGTAGCGTGGGCTG-3′ primer corresponds to the sequence encoding SAHATAM motif at the C-terminus of SOX14 protein. Genomic DNA was used as a template in PCR amplification. Goat DNA was isolated from 5 ml of blood incubated with 45 ml of lysis buffer (0.32 M sucrose, 10 mM Tris HCl, pH 7.5, 5 mM $MgCl_2$ and 1x Triton X-100). After 15 min incubation on ice and 15 min centrifugation (3000 rpm at +4°C), the pellet was rinsed twice in 10 ml of fisio buffer (0.075 M NaCl and 0.025 M EDTA). The pellet was dissolved in 1x TE buffer with 20% SDS and 10 μg/ml proteinase K and incubated on 65°C for 1 h. This was followed by phenol-chloroform and chloroform extraction. After precipitation with ethanol, the pallet was rinsed in 70% ethanol, dried and dissolved in TE. Bovine DNA from the whole blood was isolated using CTAB protocol (Del Sal *et al*. 1989) and rat DNA was isolated from frozen liver tissue (50 mg) according to Herrmann *et al*. (1986). The PCR reactions contained 2.8 pmol of each primer, 0.1 μg of genomic DNA, 400 μM dNTPs, 1.5 mM $MgCl_2$ and 2.5 U of Vent Taq polymerase (New England Biolabs, Hitchin, UK) in a 25 μl reaction mix. PCR amplifications were performed for 35 successive cycles of 98°C for 1 min and 71°C for 3 min. The PCR products, 721 bp in length, were eluated from the gel and sequenced. The nucleotide sequence data reported here have been submitted to GenBank and assigned the accession numbers: EU853675 (*Capra hircus*), EU853676 (*Bos taurus*) and EU853677 (*Rattus norvegicus*).

Databases searches, nucleotide and amino acid sequence alignments, multiple sequence analyses of GC content, as well as amino acid substitutions were performed using the National Center for Biotechnology Information (NCBI), ClustalW software (EMBL-EBI) and GeneDoc software version 2.6.003 (Nicholas Karl and Nicholas H. B. Jr 1997 Gene Doc: a tool for editing and annotating multiple sequence alignments. Distributed by authors). In this study we included *SOX14/Sox14* genes of human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), macaca (*Macaca mulatta*), dog (*Canis familiaris*), horse (*Equus cabalus*), mouse (*Mus musculus*), opposum (*Monodelphis domestica*), platypus (*Ornithorinchus anatinus*), chicken (*Gallus gallus*), frog (*Xenopus tropicalis*), zebrafish (*Danio rerio*), blue tilapia (*Oreochromis aureus*) and fugu fish (*Takifugu rubripes*), their accession numbers are provided in table 1.

The chromosomal positions of group B *Sox* genes as well as positions of their neighbouring genes were determined using Genomic Database Map Viewer (NCBI, http://www.ncbi.nlm.nih.gov/genomes) and Sanger Institute Ensembl database (http://www.ensembl.org).

To determine whether group B *Sox* genes are encompassed within syntenic regions in mammals, we have determined the chromosomal positions of additional 16 genes that are selected. based on their locations in proximity to group B *Sox* genes in human genome. The analysis included the following genes: *BCL6*, B-cell CLL/lymphoma 6 (zinc finger protein 51); *B3GALNT1*, beta-1,3-N-acetylgalactosaminyltransferase 1 (globoside blood group); *TNK2*, tyrosine kinase, non-receptor 2; *RASA2*, RAS p21 protein activator 2; *CLDN18*, claudin 18, RAB6B member RAS oncogene family; *CEP70*, centrosomal protein 70 kDa; *TFDP2*, transcription factor Dp-2 (E2F dimerization partner 2); *COL4A1*, collagen, type IV, alpha 1; *ATP11A*: ATPase, class VI, type 11A; *F7*, coagulation factor VII (serum prothrombin conversion accelerator); *LAMP1*, lysosomal-associated membrane protein 1; *TGDS*, TDP-glucose 4,6-dehydratase; *GPR180*, G protein-coupled receptor 180; *DZIP1*, DAZ interacting protein 1; *OXGR1*, oxoglutarate (alpha-ketoglutarate) receptor 1. Chromosomal positions are determined for those mammalian species whose

Table 1. Map viewer release and accession numbers used in this study.

Species	Build	Map viewer release	*SOX14* accession number
Homo sapiens (human)	36.2	14 September 2006	AJ006230
Pan troglodytes (chimpanzee)	2.1	5 October 2006	XM_526317
Macaca mulatta (rhesus macaque)	1.1	23 June 2006	XM_001114754
Canis familiaris (dog)	2.1	8 September 2005	XM_542802
Equus cabalus (domestic horse)	1.1	11 July 2007	NW_001799668
Bos taurus (cow)	3.1	3 January 2007	XM_580751
	This report		EU853676
Capra hircus (goat)	This report		EU853675
Mus musculus (mouse)	37.1	5 July 2007	AF193435
Rattus norvegicus (rat)	RGSC v 3.4	6 July 2006	NM_001106850
	This report		EU853677
Monodelphis domesticus (opossum)	MonDom5	8 March 2007	XM_001365234
Ornithorhynchus anatinus (platypus)	1.1	11 July 2007	AY112710
Gallus gallus (chicken)	2.1	30 November 2006	AF193760
Xenopus tropicalis (african clawed frog)	BLAST search		BC135637
Danio rerio (zebrafish)	Z v 6	27 February 2007	NM_001037680
Oreochromis aureus (blue tilapia)	BLAST search		EF431925
Takifugu rubripes, Sox14a	BLAST search		AY277955

mapping information is currently available in the public database.

Results and discussion

Comparative nucleotide sequence analyses of SOX14 orthologues

The entire coding sequences of goat, cow and rat *Sox14* gene were obtained by PCR reaction using primers specific for human *SOX14* gene. Here we present a first report of the nucleotide and protein sequences of *Sox14* gene in goat (EU853675). Obtained bovine and rat *Sox14* nucleotide sequences were compared to the predicted sequences already present in the database. Both bovine and rat *Sox14* nucleotide sequences determined in this report display two silent substitutions comparing to corresponding sequences in the database. In bovine *Sox14* sequence, A versus C and T versus C were determined at positions 58 and 711, respectively (see figure 1 in electronic supplementary material at http://www.ias.ac.in/jgenet/). In rat *Sox14* sequence, A versus G at positions 568 and 678 were determined (see figure 2 in electronic supplementary material). All positions are given in relation to the A in the first ATG codon. Human, mouse, chicken, fugu fish and platypus *Sox14* genes have been previously cloned and characterized (Arsic *et al*. 1998; Hargrave *et al*. 2000a; Kirby *et al*. 2002; Koopman *et al*. 2004). Platypus *Sox14* gene was cloned by PCR using oligonucleotides corresponding to 5′ and 3′ ends of the coding sequence giving the SOX14 protein missing five and nine amino acids from N-terminus and C-terminus, respectively (Kirby *et al*. 2002).

Here we have carried out comparative analyses of *SOX14* orthologues from various vertebrate species whose sequences are currently available in the public databases as well as sequences first identified in this report (table 1). We have compared nucleotide sequences of the human and other vertebrate *Sox14* genes (see figure 3 in electronic supplementary material). This comparison revealed a high level of sequence identity ranking between 76% (*Xenopus*) and 99% (chimpanzee) (table 2). Multiple sequence analyses of GC content of vertebrate *Sox14* genes demonstrated that this gene is highly GC rich (GC content varies between 55%–66%). The GC content of the human *SOX14* gene in particular is 64.17% (table 2) that is considerably higher comparing to average GC content of the human genome that is approximately 40% (Lander *et al*. 2001).

GC content in coding and noncoding DNA, codon usage, and other DNA composition parameters are useful tools to investigate the mutational and selective forces shaping the DNA sequences. The nucleotide at the silent third codon position may change more freely than those on the other codon positions due to lower selective constraints, so that no changes takes place in the specified amino acid (Corrochano and Ruiz-Albert 2004). Thus, the silent changes in third codon position, and the GC_3 value in particular (the GC level of the third codon position) provide an additional tool for studying the evolution of protein-coding regions of genes. Analyses of silent third codon position showed that GC_3 value of human *SOX14* gene is 85% (the average GC_3 value of human genes is approximately 61% (Clay *et al*. 1996), while GC_3 value of other species varies between 61% and 87%). This GC-abundance and the constraint of nucleotide replacement by C or G, may explain the high level of homology of *Sox14* sequences, as it was previously shown for *Sox4* gene (Ganesh and Raman 1997).

Vertebrate genomes are mosaics of isochores, long DNA segments (~300 kb), that are compositionally homogeneous and characterized by different molar ratio of guanine and

Table 2. Sequence analysis of *SOX14* orthologues.

Species	A (%)	C (%)	G (%)	T (%)	GC content (%)	GC_3 value (%)	Sequence identity with human *SOX14* (%)
Homo	20.56	38.06	26.11	15.28	64.17	85.0	–
Pan	20.56	38.47	26.11	14.86	64.58	86.2	99
Macaca	20.56	38.19	26.25	15.00	64.44	85.0	98
Canis	21.53	38.06	25.97	14.44	64.03	85.0	96
Equus	20.83	38.47	26.39	14.31	64.86	87.0	96
Capra	20.56	38.19	26.53	14.72	64.72	85.8	96
Bos	20.56	38.06	26.81	14.58	64.87	87.1	96
Mus	20.99	37.02	26.10	15.88	63.12	81.6	95
Rattus	21.25	37.08	25.97	15.69	63.05	82.0	95
Monodelphis	23.33	34.17	24.17	18.33	58.34	72.0	84
Ornithorinchus	19.05	38.70	28.21	14.03	66.91*	92.0*	88
Gallus	21.13	36.86	26.16	15.85	63.02	82.5	84
Xenopus	27.74	30.10	22.19	19.97	52.29	61.1	76
Danio	26.52	31.73	23.84	17.91	55.57	66.8	77
Oreochromis	24.93	31.79	24.51	18.63	56.30	70.5	79
Takifugu	24.56	31.95	23.49	17.72	55.44	72.2	78

*Results obtained based on incomplete sequence data.

cytosine in DNA (GC content) (Bernardi 2000). The human genome, for instance, was claimed to consist of five distinct isochore families: L1, L2, H1, H2 and H3, with GC contents of <37%, 37%–42%, 42%–47%, 47%–52% and >52%, respectively (Cohen *et al*. 2005). The GC content of a gene is highly correlated to the GC content of the region of the genome in which gene is located (Bernardi *et al*. 1985; Clay *et al*. 1996). Comparison of GC content and GC_3 values in different vertebrate species has demonstrated that *Sox14* gene has high GC content as well as GC_3 values, fitting in the H3 type of isochores. It has been demonstrated previously that isochore patterns are remarkably different in cold-blooded and warm-blooded vertebrates (Bernardi 2000). The difference in GC level between orthologous sequences showed that *SOX14/Sox14* sequences in amniotes are GC-richer compared to those from anamniotes (table 2). This is in agreement with the difference in GC levels of the respective genomes and isochore evolution theories (Bernardi 2000).

It is interesting to point out that GC content of a multigene segment of a chromosome could be a rough measure of its protein-producing activity (Kudla *et al*. 2006). Thus, mammalian genes with a greater proportion of G's or C's in third position show higher levels of expression comparing with the genes containing A's or U's in third position (Kudla *et al*. 2006). The difference is neither in translation of RNA into protein nor in RNA stability, but in production, or in transcription, of RNA from the DNA template (Kudla *et al*. 2006). Accordingly, GC and GC_3 values presented in this study might indicate that *Sox14* gene represents a highly transcribed gene, since silent-site GC content correlates with gene expression efficiency in mammalian cells (Kudla *et al*. 2006).

Comparative protein sequence analyses of SOX14 orthologues

Figure 1 shows multiple alignments of SOX14 proteins from different species obtained by ClustalW (Genetics Computer Group, EB1, Cambridge, UK). SOX14 protein consists of 240 amino acids in amniotes, whereas anamniotes, frog and fish, have 1–2 fewer residues, respectively. SOX14 protein contains an N-terminal HMG box (79 amino acids) and a C-terminus important for its repressor activity, consisting predominantly of noncharged residues (Arsic *et al*. 1998; Uchikawa *et al*. 1999; Hargrave *et al*. 2000a). Although many SOX proteins contain distinct domains involved in transcriptional regulation or protein interactions (Wegner 1999), SOX14 has no significant homology to any of these regions (Hargrave *et al*. 2000a). Comparative analysis on SOX14 orthologues revealed striking conservation: human SOX14 showed 100% total amino-acid identity with chimpanzee, rhesus macaque, dog, horse, goat and cow; 99% with mouse and rat; 96% with chicken; 94% with opossum and platypus; 91% with zebra fish; 90% with blue tilapia and fugu fish; and 89% identity with frog. In all examined species SOX14 proteins (except platypus whose sequence is not complete) begins at N-terminus with same block of four amino acids (MSKP) designated as N block (figure 1), that is specific for sub-group B2 SOX proteins (SOX14 and SOX21). The next amino acid (at position 5) preceding the HMG box differs between amniotes (S/T) and anamniotes (A/V) (figure 1). Within HMG box there are two conserved amino acid substitutions (E/D and A/S at positions 29 at 52,

respectively) found in blue tilapia and fugu fish, while only one of them (A/S, at position 52) is found in zebrafish. Nuclear localization signal (amino acid positions: 66–83) (Arsic *et al*. 1998) is identical in all examined species. Previously identified highly conserved group B homology region (Uchikawa *et al*. 1999) is also found to be preserved in all species that were included in this study (figure 1).

Comparison of the protein sequences of SOX14 orthologues in vertebrates revealed highly conserved motifs. We have identified two conserved stretches of amino acids designated as block A (at positions: 93–100) and block B (figure 1; at positions: 123–131), that show eight and nine consecutive identical residues, respectively. The block designated as C (figure 1; at positions: 143–152) encompasses the stretch of 10 residues that are highly conserved, with only two amino-acid substitutions, one of them being conserved. The block designated as D encompasses amino acids 164–240 and shows high conservation in all examined species, with a single nonconserved amino acid substitution found in frog only. The conserved block A is followed by a stretch of 22 amino acids that contains numerous amino acid changes (figure 1; positions: 101–122). Although this region is conserved in placental mammals (eutherians), nonconserved, semiconserved and conserved amino-acid substitutions are detected in metatherians and lower vertebrates. An insertion of two alanines is observed within this region in all amniotes (at position 107), while one threonine is inserted at the same position in frog. The conserved motif B is followed by region of 11 amino acids (figure 1; positions: 132–142) that is conserved in placental mammals, while conserved and one nonconserved substitutions are detected in other vertebrates. Conserved, semiconserved and nonconserved amino acid substitutions are detected within the stretch of 11 amino acids (figure 1; positions: 153–163) that separates the regions designated as C and D. The amino acid substitutions revealed by comparative analysis might be important for fine modulation of SOX14 activity in different species.

Comparative analysis of orhologues from different vertebrates can be used to identify specific domains that are important for protein function. To find motif(s) specific

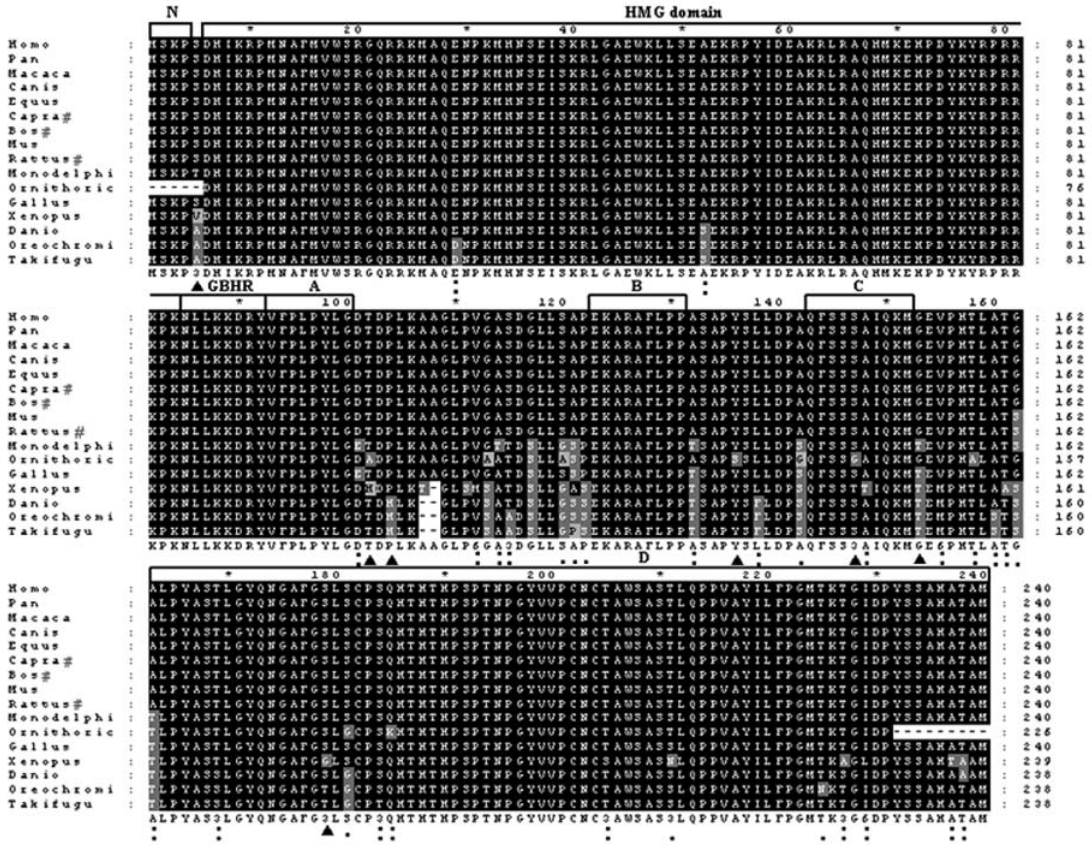

Figure 1. Amino acid alignment showing the high degree of sequence identity between SOX14 vertebrate orthologues. Identical amino acid residues are shaded black, while similar amino acid residues are shaded grey. Amino acid substitutions: (:) conserved; (.) semi-conserved; (▲) nonconserved; # deduced protein sequence of goat, cow and rat first identified in this report; N, N-terminus of SOX14 specific for B2 sub-group of SOX proteins; HMG domain, high mobility group binding domain; GBHR, group B homology region; A, B, C and D, highly conserved regions of SOX14 protein.

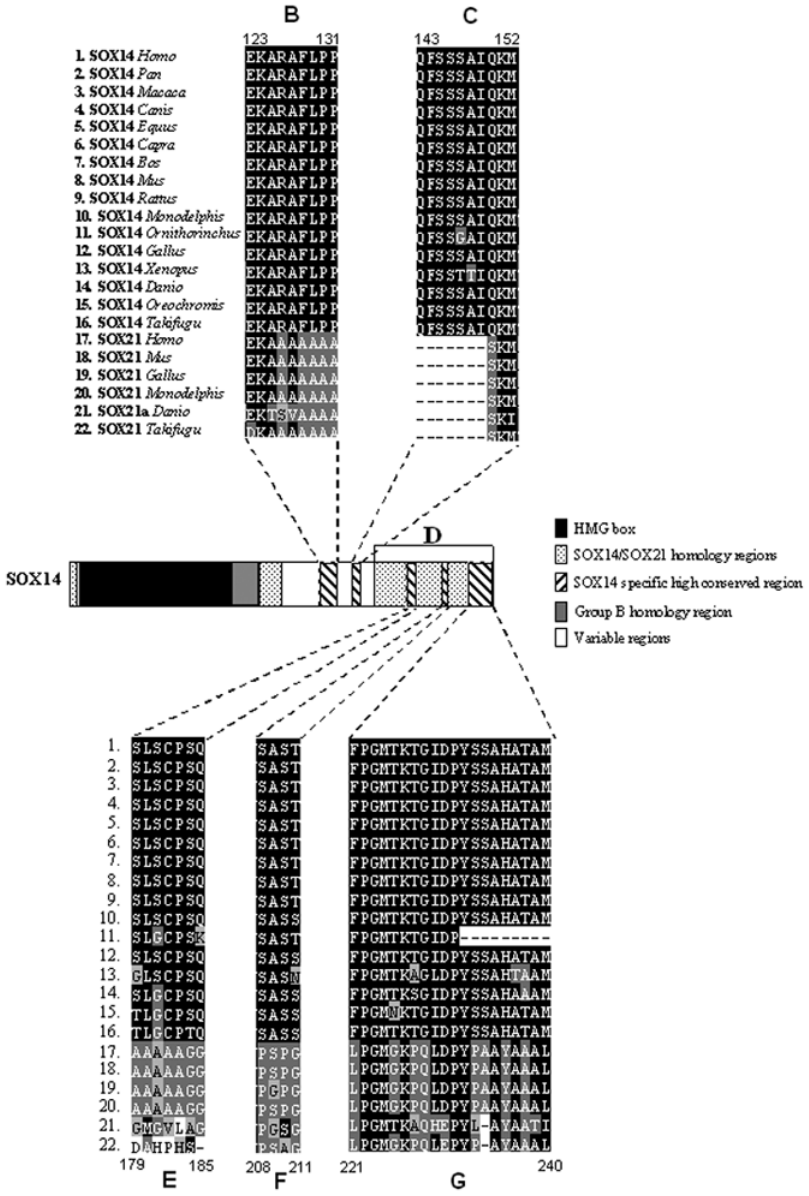

Figure 2. Schematic representation and domains organization of SOX14 protein and alignment with SOX21. The alignments of selected regions are presented, highlighting conservation within SOX14 orthologues. B, C, E, F and G, SOX14-specific regions. SOX21 accession numbers: *Homo sapiens* AAC95381, *Takifugu rubripes* AAQ18500, *Gallus gallus* BAA77266, *Mus musculus* AAN60055, *Danio rerio* NP_571361, *Monodelphis domestica* XP_001366503, 1. SOX14 *Homo sapiens*; 2, SOX14 *Pan troglodytes*; 3, SOX14 *Macaca mulatta*; 4, SOX14 *Canis familiaris*; 5, SOX14 *Equus cabalus*; 6, SOX14 *Capra hircus* 7, SOX14 *Bos taurus*; 8, SOX14 *Mus musculus*; 9, SOX14 *Rattus norvegicus*; 10, SOX14 *Monodelphis domestica*; 11, SOX14 *Ornithorinchus anatinus*; 12, SOX14 *Gallus gallus*; 13, SOX14 *Xenopus tropicalis*, 14, SOX14 *Danio rerio*, 15, SOX14 *Oreochromis aureus*; 16, SOX14 *Takifugu rubripes*; 17, SOX21 *Homo sapiens*; 18, SOX21 *Mus musculus*; 19, SOX21 *Gallus gallus*; 20, SOX21 *Monodelphis domestica*; 21, SOX21 *Danio rerio*; 22, SOX21 *Takifugu rubripes*.

for SOX14 protein, we compared this protein with its closest relative SOX21 and its orthologues from human, mouse, chicken, opossum, zebrafish and fugu fish (see figure 4 in electronic supplementary material). The alignments of selected regions are presented in figure 2. The comparative analysis of SOX14 and SOX21 orthologues confirmed the

conservation of previously identified regions with shared homology between SOX14 and SOX21 proteins, including N-terminal motif designated as block N, HMG box, group B homology region (Uchikawa *et al*. 1999), region designated in this study as block A, as well as short segments within the block D. This comparison also revealed that five sequence blocks designated in this study as B and C, as well as E, F and G (within region D) are specific for SOX14 protein only, and not for its closest relative SOX21 (figure 2) or other SOX proteins (data not shown). It is possible that these well conserved and SOX14 specific motifs define protein domains that might have structural or functional significance.

Recently, a comparative genomics study on SOX2 vertebrate orthologues showed total amino-acid identity ranking from 88.1% to 98.4% (Katoh and Katoh 2005). Additionally, other members of B group *Sox* genes (*Sox1*, *Sox3* and *Sox21*) also showed high level of protein conservation among orthologues (data not shown). However, comparative analyses of SOX14 orthologues presented in this study revealed amino-acid identity ranking between 89% and 100% indicating that SOX14 is the most conserved protein among B group *Sox* genes. High conservation of SOX14 protein in all examined species implies that this protein has been under strong evolution pressure during which it has retained its functional properties. To date, no phenotypes have been described associated with the mutation in *Sox14/SOX14* genes. The evolutional conservation and lack of mutated phenotype suggest that *Sox14* gene might be essential for development and that loss of its function might be lethal.

Evolution of group B Sox genes

Recently, Kirby *et al*. (2002) proposed a model for the evolution of group B *Sox* gene, suggesting that *Sox* B genes have arisen by duplication, rearrangement and divergence from an ancestral *Sox* B gene, which probably arose initially by retrotransposition from an intron-containing ancestral *Sox* gene (figure 3). The ancestral *Sox* B gene was then duplicated,

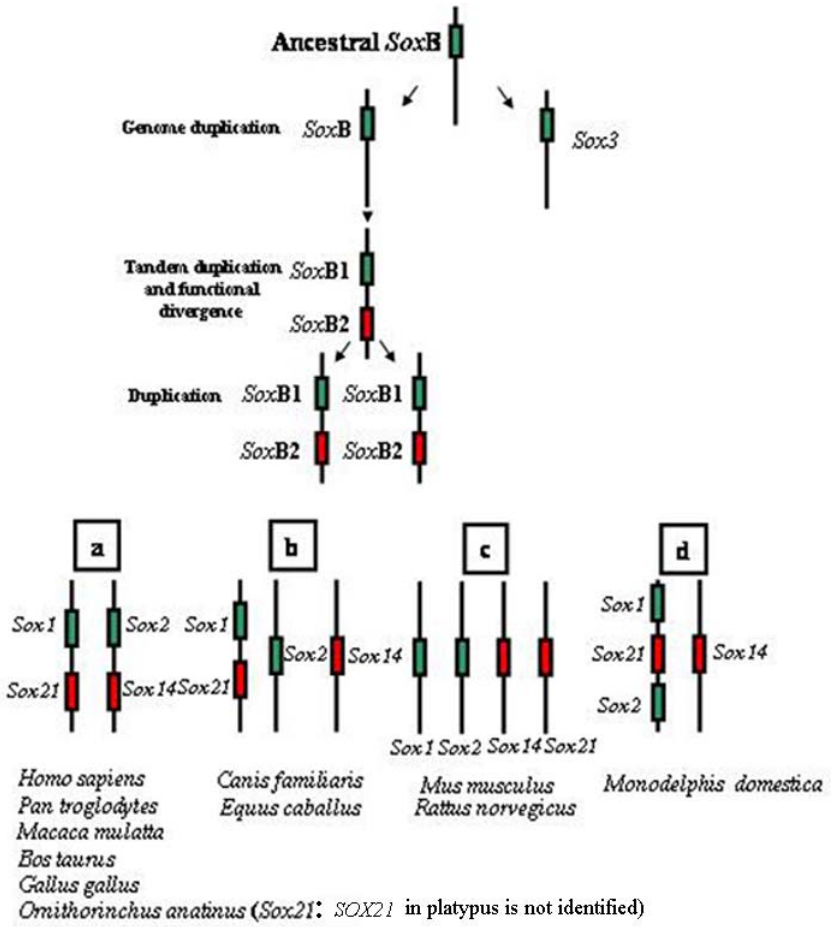

Figure 3. Schematic illustration of chromosomal locations of group B *Sox* genes in vertebrate species: proposed model of group B *Sox* genes evolution according to Kirby *et al*. (2002), and chromosomal positions according to literature data (human, mouse, platypus and chicken) and data presented in this study (chimpanzee, rhesus macaque, cow, dog, opossum, horse and rat) (a, b, c and d, respectively).

either by an unequal crossover event or as part of a whole genome duplication, to give rise to two *Sox* B genes that diverged to produce two distinct lineages on different chromosomes (figure 3). One, located on the pair of autosomes that subsequently differentiated into mammalian sex chromosomes, evolved into *Sox3*. The other *Sox* B gene copy experienced a tandem duplication creating two *Sox* B genes in close proximity, thats ubsequently diverged from one another giving rise to a pair of *Sox* B1 and *Sox* B2, one being an activator and other a repressor (figure 3). In agreement with this model, *SOX3/Sox3* is located on the mammalian X chromosome, while the remaining four autosomal *SOX* B genes in humans are arranged in two pairs (*SOX1-SOX21* and *SOX2-SOX14*) each comprising one *SOX* B1 activator and one *SOX* B2 repressor on the same chromosome. The proposed model is based on comparative mapping of group B *SOX* genes in chicken, platypus and human. *SOX2* and *SOX14* colocalize to chromosome 3q in humans, 1q in platypus and 9 in chicken (Arsic *et al*. 1998; Kirby *et al*. 2002; Kuroiwa *et al*. 2002; Stevanovic *et al*. 1994). On the other hand human *SOX1* and *SOX21* map together on chromosome 13q (Malas *et al*. 1999) and *Sox1-Sox21* map together in the chicken on chromosome one (Kuroiwa *et al*. 2002). The conserved chromosomal organization of group B members into two pairs *Sox1-Sox21* and *Sox2-Sox14* in chicken and human suggests that a common ancestor of birds and mammals had already undergone both *Sox* B duplications (Kamachi *et al*. 1998; Uchikawa *et al*. 1999).

To get better insight into genomic organization and evolution of group B *Sox* genes, we have performed comparative mapping of these genes across various vertebrate species. The localizations of *Sox* genes are presented in table 3. Using NCBI Map Viewer, we have determined that *Sox2-Sox14* pair maps together on chromosome three in chimpanzee, chromosome two in maccaca and chromosome one in cow (table 3; figure 3,a). By the same approach we have shown that *Sox1* and *Sox21* colocalize on chromosome 13 in chimpanzee, 17 in maccaca, 22 in dog, 12 in cow and 7 in opossum (table 3; figure 3,a). While *Sox1-Sox21* map together on chromosome 22, no linkage was found for *Sox2-Sox14* pair in dog (table 3; figure 3,b). On the other hand, no linkage of group B *Sox* genes is found in the mouse genome (http://www.informatics.jax.org/) (table 3; figure 3,c). The mapping data presented in this report suggest additional chromosome rearrangement that separated *Sox2* and *Sox14* genes on different chromosomes in dog and horse, while more complex rearrangement resulted in complete separation of all *Sox*B group genes in mouse and rat. These results also indicate that *Sox1, Sox2, Sox14* and *Sox21* genes are split into different chromosomes in mouse, after the lineage of mouse diverged from that of human.

In order to determine whether group B *Sox* genes are encompassed within syntenic regions in mammals, we have determined the chromosomal positions of additional 16 genes, that are selected based on their locations in proximity to group B *Sox* genes in human genome (table 4). The analysis of syntenic regions of the chromosomal locations that harbour the *Sox* genes have demonstrated that the genes which flank the *Sox14* orthologues and other group B *Sox* genes are also conserved in their positions in some mammalian species, except those flanking *Sox2* gene in rat and mouse (table 4). All together, the mapping data presented in this report provide the further evidence supporting the hypothesis that group B *Sox* genes have arisen by duplication.

Although location of *Sox1* gene is yet to be determined in rat and horse, based on conserved position of the neighbouring genes, we have assumed that *Sox1* genes in those two species are also located on chromosomes 16 and 17, respectively (tables 3 and 4; figure 3). It is interesting to point out that our mapping data revealed that three *Sox* B genes are localized at chromosome seven in opossum, two of them being activators (*sox1* and, *sox2*) and one being a repressor (*sox21*) (table 2; figure 3,d). It is possible that, after being organized as activator/repressor pairs, subsequent unequal crossing-over translocated *Sox2* gene at the same chromosome where *Sox1* and *Sox21* had been localized. Finally, if we assume that additional rearrangements have occurred in dog, mouse, rat and horse, the organization found in amniotes are in agreement with proposed model of group B *Sox* genes evolution (Kirby *et al*. 2002). The model proposed by Okuda *et al*. (2006) suggested that chromosomal organization of group B *Sox* genes in fish species differs from that of the other vertebrates, and implies that more complex mechanisms including duplications associated with gene losses and multiple rearrangements have operated during the evolution of fish. Therefore, it would be of great interest to

Table 3. Chromosomal positions of group B *Sox* genes[1].

Species	*Sox2/Sox14*		*Sox1/Sox21*	
Homo sapiens	3	3	13	13
Pan troglodytes	3*	3*	13*	13*
Macaca mulatta	2*	2*	17*	17*
Bos taurus	1*	1*	12*	12*
Gallus gallus	9*	9	1	1
Canis familiaris	34*	23*	22*	22*
Mus musculus	3*	9	8	14
Monodelphis domesticus	7*	4*	7*	7*
Danio rerio	22*	6*	1/9#	6*
Equus cabalus	19*	16*	NA	17*
Rattus norvegicus	2*	8*	NA	15*
Ornithorhynchus anatinus	1*	1	20	NM

[1] Chromosomal positions of group B *Sox* genes in different vertebrates are given according to literature data, NCBI Map Viewer and Ensemble databases. * Genes whose locations are first determined and shown in this report. # zebrafish Sox1b and Sox1a localized on chromosome one and nine, respectively (Okuda *et al*. 2006). NA, sequences are not available; NM, sequences are available but not mapped.

Table 4. Chromosomal positions of group B *Sox* genes and their neighbouring genes in different mammals.

Gene	*Homo*	*Pan*	*Macaca*	*Bos*	*Canis*	*Equus*	*Mus*	*Rattus*
SOX14	3q23	3	2	1	23	16	9	8
RAB6B	3q22.1	3	2	1	23	16	9	8
CLDN18	3q22.3	3	2	NA	23	16	9	NA
CEP70	3q22-23	3	2	NM	23	16	9	8
RASA2	3q22-23	3	2	1	23	16	9	8
TFDP2	3q23	3	2	1	23	16	9	8
SOX2	3q26-27	3	2	1	34	19	3	2
TNK2	3q29	3	2	NA	33	19	16	11
BCL6	3q27	3	2	1	34	19	16	X
B3GALNT1	3q25	3	NA	1	NA	19	3	2
SOX1	13q34	13	17	12	22	NA	8	NA
COL4A1	13q34	13	17	12	22	17	8	16
ATP11A	13q34	13	17	NM	NA	17	8	16
F7	13q34	13	17	12	22	17	8	16
LAMP1	13q34	13	17	12	22	17	8	16
SOX21	13q31-32	13	17	12	22	17	14	15
TGDS	13q32.1	13	17	12	22	17	14	NA
GPR180	13q32.1	13	17	12	NA	17	14	15
DZIP1	13q32.1	13	17	12	22	17	14	15
OXGR1	13q32.1	13	17	NA	22	17	14	15

Chromosomal positions are given according NCBI Map Viewer; NA, sequences are not available; NM, sequences are available but not mapped.

determine the chromosomal locations of group B *Sox* genes in fish species in order to get better insight into evolution of these genes in the lineage of fish.

In summary, we have identified *Sox14* gene of goat, cow and rat, and performed comparative analyses of SOX14 orthologues in vertebrates. We have found conserved motifs specific for SOX14 protein only. Also, we have determined chromosomal locations of numerous predicted group B *Sox* genes using currently available genome database. Comparative mapping of group B *Sox* genes in various vertebrate genomes presented in this study would contribute to refining our understanding of the evolution of *Sox* B genes. Continuing to decipher the structure and organization of *Sox* genes will increase our understanding the functions and evolution of these developmentally important genes.

Acknowledgements

This work has been supported by Ministry of Science, Republic of Serbia (Grant No. 143028).

References

Arsic N., Rajic T., Stanojcic S., Goodfellow P. N. and Stevanovic M. 1998 Characterisation and mapping of the human *SOX14* gene. *Cytogenet. Cell Genet*. **83**, 139–146.

Bernardi G. 2000 Isochores and the evolutionary genomics of vertebrates. *Gene* **241**, 3–17.

Bernardi G., Olofsson B., Filipski J., Zerial M., Salinas J., Cuny G. *et al*. 1985 The mosaic genome of warm-blooded vertebrates. *Science* **228**, 953–958.

Bowles J., Schepers G. and Koopman P. 2000 Phylogeny of the SOX family of developmental transcription factors based on sequence and structural indicators. *Dev. Biol*. **227**, 239–255.

Clay O., Caccio S., Zoubak S., Mouchiroud D. and Bernardi G. 1996 Human coding and noncoding DNA: compositional correlations. *Mol. Phylogenet. Evol*. **5**, 2–12.

Cohen N., Dagan T., Stone L. and Graur D. 2005 GC composition of the human genome: in search of isochores. *Mol. Biol. Evol*. **22**, 1260–1272.

Collignon J., Sockanathan S., Hacker A., Cohen-Tannoudji M., Norris D., Rastan S. *et al*. 1996 A comparison of the properties of *Sox-3* with *Sry* and two related genes, *Sox-1* and *Sox-2*. *Development* **122**, 509–520.

Corrochano L. M. and Ruiz-Albert J. 2004 Nucleotide composition in protein-coding and non-coding DNA in the zygomycete *Phycomyces blakesleeanus. Mycol. Res*. **108**, 858–863.

Del Sal G., Manfioletti G. and Schneider C. 1989 The CTAB-DNA precipitation method: a common mini-scale preparation of template DNA from phagemids, phages or plasmids suitable for sequencing. *Biotechniques* **7**, 514–520.

Ganesh S. and Raman R. 1997 *CvSox-4*, the lizard homologue of the human *SOX4* gene, shows remarkable conservation among the amniotes. *Gene* **196**, 287–290.

Hargrave M., James K., Nield K., Toomes C., Georgas K., Sullivan T. *et al*. 2000a Fine mapping of the neurally expressed gene *SOX14* to human 3q23, relative to three congenital diseases. *Hum. Genet*. **106**, 432–439.

Hargrave M., Karunaratne A., Cox L., Wood S., Koopman P. and Yamada T. 2000b The HMG box transcription factor gene *Sox14* marks a novel subset of ventral interneurons and is regulated by sonic hedgehog. *Dev. Biol*. **219**, 142–153.

Herrmann B., Bucan M., Mains P. E., Frischauf A. M., Silver L. M. and Lehrach H. 1986 Genetic analysis of the proximal portion of

the mouse t complex: evidence for a second inversion within t haplotypes. *Cell* **44**, 469–476.

Kamachi Y., Uchikawa M., Collignon J., Lovell-Badge R. and Kondoh H. 1998 Involvement of *Sox1, 2* and *3* in the early and subsequent molecular events of lens induction. *Development* **125**, 2521–2532.

Katoh Y. and Katoh M. 2005 Comparative genomics on *SOX2* orthologs. *Oncol. Rep*. **14**, 797–800.

Kirby P. J., Waters P. D., Delbridge M., Svartman M., Stewart A. N., Nagai K. *et al*. 2002 Cloning and mapping of platypus *SOX2* and *SOX14*: insights into *SOX* group B evolution. *Cytogenet. Genome Res*. **98**, 96–100.

Kishi M., Mizuseki K., Sasai N., Yamazaki H., Shiota K., Nakanishi S. *et al*. 2000 Requirement of Sox2-mediated signaling for differentiation of early *Xenopus* neuroectoderm. *Development* **127**, 791–800.

Koopman P., Schepers G., Brenner S., Venkatesh B. 2004 Origin and diversity of the SOX transcription factor gene family: genome-wide analysis in *Fugu rubripes. Gene* **328**, 177–186.

Kovacevic-Grujicic N., Mojsin M., Djurovic J., Petrovic I. and Stevanovic M. 2008 Comparison of promoter regions of *SOX3, SOX14* and *SOX18* orthologs in mammals. *DNA Seq*. **19**, 185–194.

Kudla G., Lipinski L., Caffin F., Helwak A. and Zylicz M. 2006 High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol*. **4**, e180.

Kuroiwa A., Uchikawa M., Kamachi Y., Kondoh H., Nishida-Umehara C., Masabanda J. *et al*. 2002 Chromosome assignment of eight *SOX* family genes in chicken. *Cytogenet. Genome Res*. **98**, 189–193.

Lander E. S., Linton L. M., Birren B., Nusbaum C., Zody M. C., Baldwin J. *et al*. 2001 Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

Malas S., Duthie S., Deloukas P. and Episkopou V. 1999 The isolation and high- resolution chromosomal mapping of human *SOX14* and *SOX21*; two members of the *SOX* gene family related to *SOX1, SOX2*, and *SOX3*. *Mamm. Genome* **10**, 934–937.

McKimmie C., Woerfel G. and Russell S. 2005 Conserved genomic organisation of Group B *Sox* genes in insects. *BMC Genet*. **6**, 26.

Okuda Y., Yoda H., Uchikawa M., Furutani-Seiki M., Takeda H., Kondoh H. *et al*. 2006 Comparative genomic and expression analysis of group B1 *sox* genes in zebrafish indicates their diversification during vertebrate evolution. *Dev. Dyn*. **235**, 811–825.

Pevny L. H. and Lovell-Badge R. 1997 *Sox* genes find their feet. *Curr. Opin. Genet. Dev*. **7**, 338–344.

Stevanovic M., Zuffardi O., Collignon J., Lovell-Badge R. and Goodfellow P. 1994 The cDNA sequence and chromosomal location of the human *SOX2* gene. *Mamm. Genome* **5**, 640–642.

Uchikawa M., Kamachi Y. and Kondoh H. 1999 Two distinct subgroups of Group B *Sox* genes for transcriptional activators and repressors: their expression during embryonic organogenesis of the chicken. *Mech. Dev*. **84**, 103–120.

Wegner M. 1999 From head to toes: the multiple facets of Sox proteins. *Nucleic Acids Res*. **27**, 1409–1420.

Received 1 July 2008; in revised form 12 August 2008; accepted 9 September 2008
Published on the Web: 5 March 2009