Nucleic Acids Research, 1993, Vol. 21, No. 10 **2493–2501**

Ancestry and diversity of the HMG box superfamily

Vincent Laudet, Dominique Stehelin and Hans Clevers[1]
CNRS URA 1160, Institut Pasteur, 1 Rue Calmette, 59019 Lille Cedex, France and [1]Department of Immunology, University Hospital Utrecht, PO Box 85500, 3508 GA, Utrecht, The Netherlands

Received December 7, 1992; Revised and Accepted April 22, 1993

ABSTRACT

The HMG box is a novel type of DNA-binding domain found in a diverse group of proteins. The HMG box superfamily comprises a.o. the High Mobility Group proteins HMG1 and HMG2, the nucleolar transcription factor UBF, the lymphoid transcription factors TCF-1 and LEF-1, the fungal mating-type genes mat-Mc and MATA1, and the mammalian sex-determining gene SRY. The superfamily dates back to at least 1,000 million years ago, as its members appear in animals, plants and yeast. Alignment of all known HMG boxes defined an unusually loose consensus sequence. We constructed phylogenetic trees connecting the members of the HMG box superfamily in order to understand their evolution. This analysis led us to distinguish two subfamilies: one comprising proteins with a single sequence-specific HMG box, the other encompassing relatively non sequence-specific DNA-binding proteins with multiple HMG boxes. By studying the extent of diversification of the superfamily, we found that the speed of evolution was very different within the various groups of HMG-box containing factors. Comparison of the evolution of the two boxes of ABF2 and of mtTF1 implied different diversification models for these two proteins. Finally, we provide a tree for the highly complex group of SRY-like ('Sox' genes), clustering at least 40 different loci that rapidly diverged in various animal lineages.

INTRODUCTION

A large proportion of the eukaryotic DNA-binding proteins cloned to date can be grouped into a small number of families, defined by the presence of conserved structural motifs such as the zinc finger (1), the basic leucine zipper (2), the homeodomain (3) and the helix-loop-helix motif (4). Tjian and co-workers recently recognized a novel type of DNA-binding domain repeated six times in the RNA polymerase I transcription factor UBF. This repeated domain is homologous to two regions in High Mobility Group 1 (HMG1) proteins, and was therefore coined the HMG box (5). One of the HMG boxes of UBF was shown to be sufficient for binding to a DNA-affinity column (5). Several HMG box containing proteins have since been identified, including the products of the fungal mating type genes Mat-Mc of *S.pombe* (6) and Mt A1 of *N.crassa* (7), the mammalian sex-determining gene SRY (8,9), the lymphoid transcription factors TCF1 and LEF1 (10–12), and the mitochondrial transcription factor mtTF1 (13). The consensus HMG box comprises approximately 80 amino acid residues; average sequence identity between individual HMG boxes is close to 25%. The HMG box is believed to interact with DNA as a monomer (14; M. Van de Wetering and H.C., unpublished).

Most HMG box proteins contain two or more HMG boxes and appear to bind DNA in a relatively sequence-aspecific manner (5, 13, 15, 16 and references therein). A curious property was described recently for HMG1 and SRY, that can interact with cruciform DNA irrespectiveLY of sequence (16, 17). A smaller number of these proteins contain a single HMG box, and bind in a highly sequence-specific manner as exemplified by various footprinting experiments. The latter group includes the yeast mating type gene products MC, MATA1 and the STE11 gene; SRY and its homologues in insects and vertebrates, and the TCF-like genes (TCF-1, -3, -4 and LEF-1). Despite the relatively low level of homology between mating type genes, SRY- and TCF-like genes, they all appear to bind to the minor groove of the A/T A/T C A A A G-motif (10, 14, 18–20).

Examples of both types of HMG box proteins have been found in yeast, plants, insects and vertebrates, implying an ancient evolutionary history for this gene family. We have collected the sequences of the HMG boxes of proteins from plants, yeast and animals and have constructed evolutionary trees for the HMG box family of DNA-binding proteins. We conclude that the HMG box superfamily appeared more than 1,000 millions years ago and since this time was organized into two subfamilies: the TCF/SOX subfamily and the UBF/HMG subfamily.

MATERIALS AND METHODS

Sequence sources and alignment

Sequences used for this study are shown in Table I and II. For each mammalian gene, the human or rodent sequences were used indifferently when available. We have checked that the introduction of various mammalian versions of these genes does not change the topology of the trees (data not shown). M1P, M2P, M4P, M5P, M6P and M8P were cloned by PCR using guessmer

* To whom correspondence should be addressed

```
                        10        20        30        40        50        60        70        80
TCF1          KPTIKK-PLNAFMLYMKEMRAKVIAECTLKE--SAAINQILGRRWHALSREEQAKYYELARKERQLHMQLYPGWSAR-DNYGKKK          81
LEF1          RPHIKK-PLNAFMLYMKEMRANVVAECTLKE--SAAINQILGRRWHALSREEQAKYYELARKERQLHMQLYPGWSAR-DNYGKKK          81
MATA1         KAKIPR-PPNAYILYRKDHHREIREQNPGLH--NNEIAVIVGNMWRDEQPHIREKYFNMSNEIKTRLLLENPDYRYN-PRRSQDI          81
MC            TERTPR-PPNAFILYRKEKHATLLKSNPSIN--NSQVSKLVGEMWRNESKEVRMRYFKMSEFYKAQHQKMYPGYKYQ-PRKNKVK          81
STE11         KSSVKR-PLNSFMLYRRDRQAEI----PTSN--HQSISRIIGQLWRNESAQVKKYYSDLSALERQKHMLENPEYKYT-PKKRSTV          77
HSRY          QDRVKR-PMNAFIVWSRDQRRKMALENPRMR--NSEISKQLGYQWKMLTEAEKWPFFQEAQKLQAMHREKYPNYKYR-PRRKAKM          81
MSRY          EGHVKR-PMNAFMVWSRGERHKLAQQNPSMQ--NTEISKQLGCRWKSLTEAEKRPFFQEAQRLKILHREKYPNYKYQ-PHRRAKV          81
MA1           QDRVKR-PMNAFMVWSRGQRRKMAQENPKMH--NSEISKRLGAEWKVMSEAEKRPFIDEAKRLRALHMKEHPDYKYR-PRRKTKT          81
MA2           PDRVKR-PMNAFMVWSRGQRRKMAQENPKMH--NSEISKRLGAEWKLLSETEKRPFIDEAKRLRALHMKEHPDYKYR-PRRKTKT          81
MA3           QDRVKR-PMNAFMVWSRGQRRKMALENPKMH--NSEISKRLGADWKLLTDAEKRPFIDEAKRLRAVHMKEYPDYKYR-PRRKTKT          81
MA4           SGHIKR-PMNAFMVWSQIERRKIMEQSPDMH--NAEISKRLGKRWKLLKDSDKIPFIQEAERLRLKHMADYPDYKYR-PRKKVKS          81
ABF2-1        KQGPKR-PTSAYFLYLQDHRSQFVKENPTLR--PAEISKIAGEKWQNLEADIKEKYISERKKLYSEYQKAKKEFDEKLPP---KK          79
ABF2-2        KLPPKK-PAGPFIKYANEVRSQVFAQHPDKS--QLDLMKIIGDKWQSLDQSIKDKYIQE-------YKKAIQEYNARYPL---N-          71
T160          PNAPKR-PMSAYMLWLNASREKIKSDHPGIS--ITDLSKKAGEIWKGMSKEKKEEWDRKAEDARREYEKAMKEYEGGRGDSSKRD          82
SSRP1         PNAPKR-PMSAYMLWLNASREKIKSDHPGIS--ITDLSKKAGEIWKGMSKEKKEEWDRKAEDARRDYEKAMKEYEGGRGESSKR-          81
DROHMGD       SDKPKR-PLSAYMLWLNSARESIKRENPGIK--VTEVAKRGGELWRAMK--DKSEWEAKAAKAKDDYDRAVKEFEAN-GGSSAAN          79
NHP6A         PNAPKR-ALSAYMFFANENRDIVRSENPDIT--FGQVGKKLGEKWKALTPEEKQPYEAKAQADKKRYESEKELYNA------TLA          76
NHP6B         PNAPKR-GLSAYMFFANENRDIVRSENPDVT--FGQVGRILGERWKALTAEEKQPYESKAQADKKRYESEKELYNA------TRA          76
ARA           PNAPKR-AMSGFMFFSQMERDNIKKEHPGIA--FGEVGKVLGDKCVKCLLMIKSHMKPRLQVDKQRYKDEISDYKN------PQP          76
HMG1-2        PNAPKR-PPSAFFLFCSEYRPKIKGEHPGLS--IGDVAKKLGEMWNNTAADDKQPYEKKAAKLKEKYEKDIAAYRA-KGKPDAAK          81
HMG2-2        PNAPKR-PPSAFFLFCSEHRPKIKSEHPGLS--IGDTAKKLGEMWSEQSAKDKQPYEQKAAKLKEKYEKDIAAYRA-KGKSEAGK          81
HMGT-2        PNAPKR-PSSAFFIFCADFRPQVKGETPGLS--IGDVAKKLGEKWNNLTAEDKVPYEKKASKLKEKYEKDITAYRN-KGKVPVSM          81
HMGMAIZE      PNKPKR-APSAFFVFMEEFRKEFKEKNPKNK-SVAAVGKAAGDRWKSLSESDKAPYVAKANKLKLEYNKAIAAYN--KGESTAAK          81
HMGSOY        PNKPKR-PPSAFFVFMEEFRKVFNKEHPENK-AVSAVGKAAGAKWKTMSDAEKAPYVAKSEKRKVEYEKNMRAYN--KKQAEGPT          81
TETHMGB       PSKPKR-PQTGFFIYKSEVFAKRRTECPTLK--VPEIVSKISEEYKALPEKEKQKYEEAYRKEKATYDKQNDQWKEKYGDIEKS-          81
TETHMGC       PAPPKR-PLSAFFLFKQHNYEQVKKENPNAK--ITELTSMIAEKWKAVGEKEKKKYETLQSEAKAKYEKALQAYEKKYGKPEKQ-          81
LG1           PAPPKR-PLSAFFLFKQHNYEQVKKENPNAK--ITELTSMIAEKWKAVGEKEKKKYETLQSEAKAKYEKDMQAYEKKYGKPEKQ-          81
HMG1-1        PKKPRG-KMSSYAFFVQTCREEHKKKHPDASVNFSEFSKKCSERWKTMSAKEKGKFEDMAKADKARYEREMKTYIPPKGET-KKK          83
HMG2-1        PNKPRG-KMSSYAFFVQTCREEHKKKHPDSSVNFAEFSKKCSERWKTMSAKEKSKFEDMAKSDKARYDREMKNYVPPKGDK-KGK          83
HMGT-1        PRKPRG-KMSSYAYFVQTRREEHKKKHPEASVNFSEFSKKCSERWKTMSAKEKGKFEDLAKLDKVRYEREMRSYIPPKGEK-KKR          83
MTTF1-2       LGKPKR-PRSAYNVYVAERFQEAKGDSPQEK--L----KTVKENWKNLSDSEKELYIQHAKEDETRYHNEMKSWEEQMIEVGR-K          77
CCG1          TTKDEE-FIRKFALFDEQHREEMRKER-----------RRIQEQLRRLKRNQEKEKLKGPPEKKPKKMKERPDLKLKCGACGAIG          73
XUBF-1        PEFPKK-PLTPYFRFFMEKRAKYAKLHPEMS--NLDLTKILSKKYKELPEKKKMKYIQDFQREKLEFERNLARFREEHPD---LM          79
HUBF-1        PDFPKK-PLTPYFRFFMEKRAKYAKLHPEMS--NLDLTKILSKKYKELPEKKKMKYIQDFQREKQEFERNLARFREDHPD---LI          79
HUBF-3        SEKPKR-PVSAMFIFSEEKRRQLQEERPELS--ESELTRLLARMWNDLSEKKKAKYKAREAALKAQSERKPGGEREERGK---LP          79
XUBF-3        DGRPTKPPPNSYSMYCAELMAN----MKDVP--STERMVLCSQRWKLLSQKEKDAYNKKCEQRKKDYEVELMRFLESLPEEEQQR          79
HUBF-2A       DGRPTKPPPNSYSLYCAELMAN----MKDVP--STERMVLCSQQWKLLSQKEKDAYHKKCDQKKKDYEVELLRFLESLPEEEQQR          79
HUBF-5        QGEPKKPPMNGYQKFSQELLSN--GELNHLP--LKERMVEIGSRWQRISQSQKEHYKKLAEEQQKQYKVHLDLWVKSLSPQDRAA          81
XUBF-5        LGEPKKAPMNGYQKFSQELLSN--GELNHLP--LKERMVEIGSRWHRISPSQKDYYKKLAEDQQRVYRTQFDTWMKGLSSQDRAA          81
MTTF1-1       ASCPKK-PVSSYLRFSKEQLPIFKAQNPDAK--TTELIRRIAQRWRELPDSKKKIYQDAYRAEWQVYKEEISRFKEQLTPSQIMS          82
HUBF-2        SDIPEK-PKTPQQLWYTHEKKVYLKVRPDAT--TKEVKDSLGKQWSQLSDKKRLKWIHKALEQRKEYEEIMRDYIQKHPELNI-S          81
XUBF-2        SDVPEK-PKTPQQLWYNHERKVYLKLHADAS--TKDVKDALGKQWSQLTDKKRLKWIHKALEQRKQYEGIMREYMQKHPELNI-A          81
HUBF-4        GKLPES-PKRAEEIWQQSVIGDYLARFKNDR--VKALK-AMEMTWNNMEKKEKLMWIKKAAEDQKRYERELSEMRAPPAATNS-S          80
XUBF-4        AKLPET-PKTAEEIWQQSVIGDYLARFKNDR--AKALK-SMEGTWLNMEKKEKIMWIKKAAEDQKRYERELSDMRATPTPTTA-G          81
CONS          ...PK..P..A........R.......P.............G..W.......K..Y...A......Y......Y..........
HELIX         .........**********.....**********..**********.....**********.....**********.....**********.....
```

Figure 1. Alignment of the HMG boxes. Each box (even boxes from the same protein) was treated as a single taxonomic unit. The bottom 'CONS' line indicates some amino acid positions less divergent than the rest of the box. No real consensus can be derived from this alignment. The residus implicated in alpha helix on the NMR structure recently described (65) are noted by stars on the 'HELIX' line at the bottom of the figure.

primers (M. Van de Wetering and H.C., unpublished). Numerous SRY-related genes were found in the GenBank data base (see ref 64).

The HMG sequences were used to conduct a computer alignment procedure using the CLUSTAL package available on the CITI-2/Bisance network (21, 22). Because of the low sequence similarities we have chosen to confer a substantial penalty for the insertion of gaps. The alignment of the SRY family genes is available from V.L. upon request.

Construction of phylogenetic trees

Our method was similar to the one used by Laudet et al., (1992) (23). The percent divergence values for all pairwise comparisons of the aligned sequences were calculated by dividing the number of different residues by the total number of compared residues. Gaps were treated as mismatches. Before tree construction, all values were transformed into distances (d) with Poisson correction $d = \ln(1-S)$ where S is the proportion of sites that differ (24). These values were then used to construct phylogenetic trees by the Fitch least Squares method (25) and by the NeigHbor-Joining (NJ) method described by Saitou and Nei (1987). We have preferred NJ to UPGMA (26) since the UPGMA method implies an equal rate of change along all sequences (27), an assumption which might not be true for genes encoding HMG boxes. The distance matrix as well as the trees not shown in the paper are available from V.L. upon request.

RESULTS

Alignment of HMG box sequences

Our first aim was to produce a list of all HMG-box sequences through a data base search (Genbank, EMBL data library and NBRF) with the Fasta procedure. We have used different members of the family as well as short peptides signatures (data not shown) to find all possible sequences. The result of this exhaustive search is given in Tables 1 and 2. Two sequences that have been described as HMG boxes were excluded from our sequence list since they do not appear to belong to the superfamily. The ACP2 protein (28) was described as a *Saccharomyces cerevisiae* HMG1-like protein. Indeed, this sequence shows 19% amino acid identity with Calf thymus HMG1 but numerous amino acids well conserved within the HMG superfamily are not present in ACP2, suggesting that the observed homology only reflects a sequence convergence. Since its first description, it has been demonstrated that ACP2 actually encodes a subunit of the RNA polymerase III from *Saccharomyces* and that the described homology corresponds to a non significant convergence of sequences rich in acidic amino acids (29). Based on similar arguments, we do not consider the histone H1 from *Tetrahymena thermophila* as an HMG box protein (30, 31).

The sequences listed in Table 1 were aligned using the procedure described in ref 23 and the Clustal V program (21). In the cases where a protein contains more than one HMG box, we have treated each box separately. This allowed us to compare the evolution of the various boxes present in proteins such as UBF or 'classical' HMGs (i.e. HMG1 and HMG2 proteins). The Clustal V alignment was slightly modified in the central part of the HMG domain where some variations appear in the length of the sequences. In this region, the gaps were introduced exactly at the same place in the various sequences assuming only one insertion event common to all HMG1 group members. The final alignment of all the 44 HMG boxes used in this study is shown in Figure 1. The length of the HMG domain varies from 71 (for

Table 1. Sequences used in this study.

Factor	Box number	Abbreviation	Species	References
TCF1	1	TCF1	Human	10
LEF1	1	LEF1	Human	11, 12
MATA1	1	MATA1	Neurospora	7
MC	1	MC	Schizosaccharomyces	6
STE11	1	STE11	Schizosaccharomyces	19
SRY	1	HSRY	Human	9
		MSRY	Mouse	8
MA1	1	MA1	Mouse	8
MA2	1	MA2	Mouse	8
MA3	1	MA3	Mouse	8
MA4	1	MA4	Mouse	8
CCG1/P250	1	CCG1	Human	61, 62
ABF2	2	ABF2-1	Saccharomyces	15
		ABF2-2		
T160	1	T160	Mouse	49
SSRP1	1	SSRP1	Human	50
DROHMGD	1	DROHMGD	Drosophila	51
NHP6A	1	NHP6A	Saccharomyces	52
NHP6B	1	NHP6B	Saccharomyces	52
HMGARA	1	ARA	Arabidopsis	63
HMG1	2	HMG1-1	Human	53, 54
		HMG1-2		
HMG2	2	HMG2-1	Human	53
		HMG2-2		
HMGT	2	HMGT-1	Trout	55
		HMGT-2		
HMGMAIZE	1	HMGMAIZE	Maize	56
HMGSOY	1	HMGSOY	Soybean	57
TETHMGB	1	TETHMGB	Tetrahymena	31
TETHMGC	1	TETHMGC	Tetrahymena	31
LG1	1	LG1	Tetrahymena	35
MTTF1	2	MTTF1-1	Human	13
		MTTF1-2		
HUBF	6	HUBF-1	Human	5
		HUBF-2		
		HUBF-2A		
		HUBF-3		
		HUBF-4		
		HUBF-5		
XUBF	5	XUBF-1	Xenopus	58
		XUBF-2		
		XUBF-3		
		XUBF-4		
		XUBF-5		

Abbreviations used in the text are indicated as well as the number of HMG boxes contained in each factor. Sequences are given in the order of the Fitch tree of Fig. 3.

box 2 of ABF2) to 83 amino acids (for box 1 of the 'classical' HMGs).

From the alignment presented in Fig. 1, it appears that the sequence of the HMG box is highly variable. This may explain why it is sometimes difficult to determine whether a given protein belongs to the HMG box superfamily. There are no strictly conserved amino acids between all the superfamily members: it is thus impossible to propose an unequivocal signature of the superfamily. Furthermore only three amino acids are conserved in more than 80% of the sequences: a P at position 8 of our alignment, a W at position 45 and a K at position 53. Several other positions are also conserved throughout the family, but to a lesser extent (see Fig.1).

The extreme structural and functional diversity of the HMG box superfamily may explain why several authors have suggested a relationship between HMG members and unrelated factors (see 32 for a review) *e.g.* it has been proposed that the HMG box of LEF-1 shares homology with the conserved DNA binding domain of the ETS family of transcription factors (12). Careful examination of sequence alignments between various HMG boxes and ETS domains (33, 34) leads us to conclude that there is no significant relationship between these two superfamilies. In the phylogenetic trees relating these sequences, the connection between the two families was highly variable and very sensitive to subtle variations on the alignment or on the tree reconstruction procedure used (data not shown). This indicates an absence of real homology between these sequences. Based on the same argument a relation between the eukaryotic hsp70 gene, the *E.coli* dnaK gene and the central portion of the HMG box (15, 32) was ruled out.

Generation of phylogenetic trees

Using the alignment shown in Fig. 1, we have constructed phylogenetic trees relating the 44 HMG-boxes using two different programs (Neighbor-Joining and Fitch Least Square analyses) both based on distance matrix calculation (Fig. 2, 3 and 4). The

Table 2. SOX-related sequences used in this study

Abbreviation	Species	Reference	Size	GenBank	Putative locus
CKCH1	Chicken	64	54	CHKCH1DNA	AMA3 ?
CKCH7	Chicken	64	54	CHKCH7DNA	AMA3 ?
AMA3	Alligator	64	54	ALLAMA3DNA	AMA3 ?
CKCH31	Chicken	64	54	CHKCH31DNA	AMA2
AMA2	Alligator	64	54	ALLAMA2DNA	AMA2
CKCH4	Chicken	64	54	CHKCH4DNA	AMA1
AMA1	Alligator	64	54	ALLAMA1DNA	AMA1
CKCH2	Chicken	64	54	CHKCH2DNA	CH2 ?
CKCH60	Chicken	64	54	CHKCH60	CH60 ?
CKCH32	Chicken	64	54	CHKCH32DNA	CH32
DM17	Drosophila	64	54	DRODM17DNA	CH32
CKCH3	Chicken	64	54	CHKCH3DNA	CH3 ?
DM64	Drosophila	64	54	DRODM64GEN	DM Cluster
DM23	Drosophila	64	54	DRODM23DNA	DM Cluster
DM33	Drosophila	64	54	DRODM33DNA	DM Cluster
DM36	Drosophila	64	54	DRODM36DNA	DM Cluster
DM10	Drosophila	64	54	DRODM10DNA	DM Cluster
DM63	Drosophila	64	54	DRODM64DNA	DM Cluster
M6P	Mouse	H.C., unpublished	52	–	Mouse SOX1 Cluster
M4P	Mouse	H.C., unpublished	52	–	Mouse SOX1 Cluster
MA1	Mouse	8	81	–	Mouse SOX1 Cluster
MA2	Mouse	8	81	–	Mouse SOX1 Cluster
MMSOX14	Mouse	67	56	MMSOX14	Mouse SOX1 Cluster
M5P	Mouse	H.C., unpublished	52	–	Mouse SOX1 Cluster
HUMSOX10	Human	20	54	HUMSOX10	SOX10
HUMSOX9	Human	20	54	HUMSOX9	SOX8P
M8P	Mouse	H.C., unpublished	52	–	SOX8P
MA3	Mouse	8	81	–	SOX3
XELSOX11	Xenopus	20	54	XELXSOX11	??
MMSOX15	Mouse	68	54	–	SOX15
HSRY	Human	9	81	HUMSRY	SRY
RSRY	Rabbit	36	79	–	SRY
MSRY	Mouse	8	81	MUSSRYLOC	SRY
SMSRY	Sminthopsis macroura	36	78	–	SRY
MESRY	Macropus eugenii	36	79	–	SRY
HUMSOX5	Human	20	54	HUMSOX5	SOX5
MUSSOX5	Mouse	20, 37	54	MUSSOX5P	SOX5
XELSOX5	Xenopus	20	54	XELXSOX5	SOX5
HUMSOX6	Human	20	54	HUMSOX6	SOX6
MUSSOX6	Mouse	20	54	MUSSOX6	SOX6
MMSOX13	Mouse	67	56	MMSOX13	??
XELSOX12	Xenopus	20	54	XELXSOX12	??
MUSSOX7	Mouse	20	54	MUSSOX7	SOX7
HUMSOX8	Human	20	54	HUMSOX8	SOX8
DROSOX15	Drosophila	20	54	DRODSOX15	SOX8
MMSOX10	Mouse	67	56	MMSOX10	SOX10
MMSOX8	Mouse	67	56	MMSOX8	??
MMSOX9	Mouse	67	56	MMSOX9	SOX9
MA4	Mouse	8	81	–	SOX4
HUMSOX4	Human	20	54	HUMSOX4	SOX4
MMSOX11	Mouse	67	56	MMSOX11	SOX11
XELSOX13	Xenopus	20	54	XELSOX13	SOX13
M1P	Mouse	H.C., unpublished	52	–	SOX13
AES4	Alligator	64	72	ALLAES4	AES Cluster ?
AES1	Alligator	64	72	ALLAES1DNA	AES Cluster ?
AES2	Alligator	64	72	ALLAES2DNA	AES Cluster ?
AES6	Alligator	64	72	ALLAES6DNA	AES Cluster ?
TMG44	Gecko	64	72	TELMG44	Reptilia SOX12 Cluster
TMG42	Gecko	64	72	TELMG42DNA	Reptilia SOX12 Cluster
TMG43	Gecko	64	72	TELMG43DNA	Reptilia SOX12 Cluster
LG28	Eublepharis macularis	64	71	EULLG27DNA	Reptilia SOX12 Cluster
LG27	Eublepharis macularis	64	71	EULLG28DNA	Reptilia SOX12 Cluster
ADW4	Alligator	64	72	ALLADW4DNA	Reptilia SOX12 Cluster
ADW5	Alligator	64	71	ALLADW5DNA	Reptilia SOX12 Cluster
ADW2	Alligator	64	72	ALLADW2DNA	Reptilia SOX12 Cluster
MMSOX12	Mouse	67	56	MMSOX12	SOX12
DROSOX14	Drosophila	20	54	DRODSOX14	SOX14

The size in amino acid is indicated for each sequence as well as the GenBank code and/or the reference when available. The species are also indicated. *Eublepharis macularis* is a Reptilia, *Sminthopsis macroura* and *Macropus eugenii* are two Marsupial species. The 'Putative locus' column indicates homologous genes. HUMSOX5, MUSSOX5 and XELSOX5 are obvious homologues of the Sox5 locus in human, mouse and Xenopus respectively. But, from the evolutionary tree it appears that other genes may also be considered as homologues (CKCH31 and AMA2 or XELSOX13 and M1P, for example). Species-specific gene clusters as defined in Fig.6 are also indicated in this column. Sequences are given in the order of the Fitch tree of Fig.6

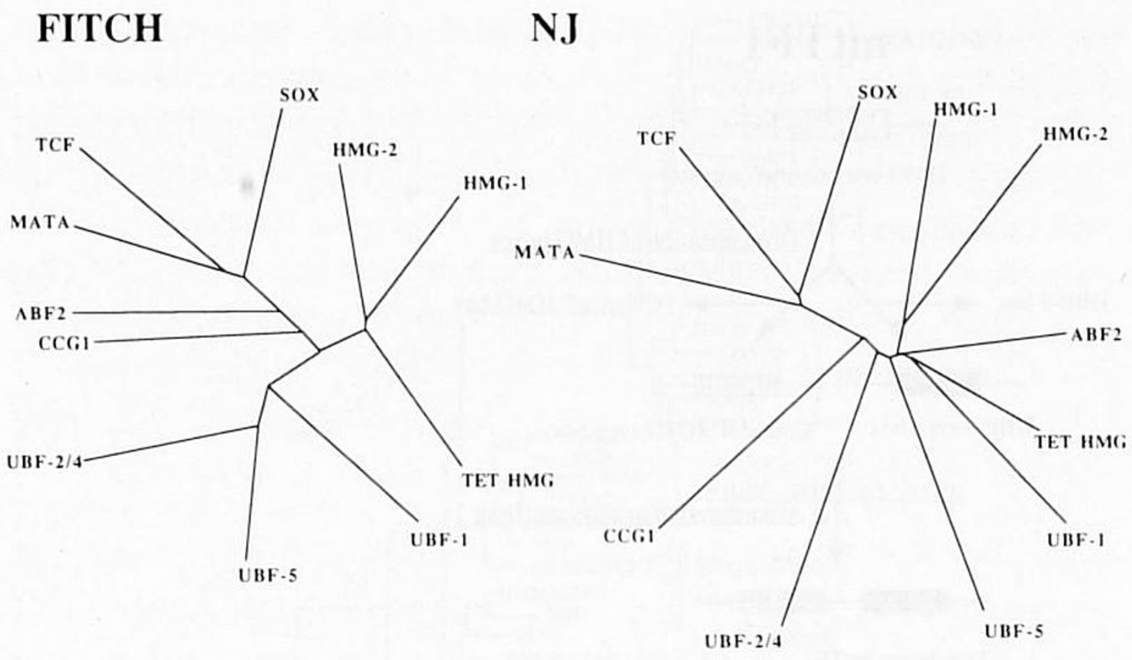

Figure 2. Comparison of the unrooted phylogenetic Fitch (left) and Neighbour-Joining (right) trees. For clarity, only the eleven groups of genes have been indicated. The position of the TETHMG and ABF2 groups are the most variable.

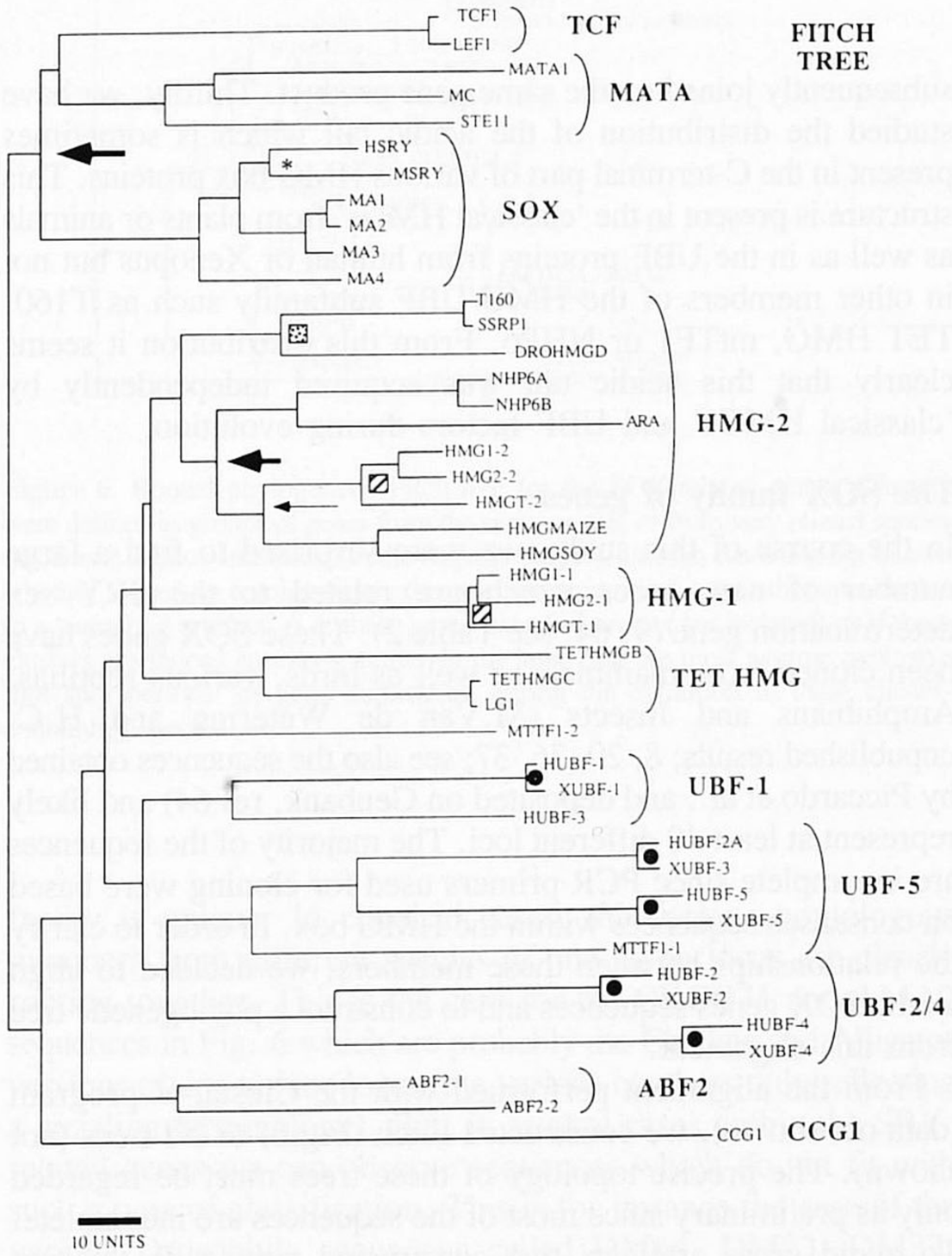

Figure 3. Unrooted phylogenetic Fitch tree for the HMG box superfamily. The bar represents a branch length of 10 units which reflects relative sequence divergence. Groups are indicated by brackets on the right. Large arrows indicate the dichotomy between animal and fungal genes; small arrows between animal and plant genes (all approx 1,000 million years ago). Stitched squares show dichotomy between Arthropods and Vertebrates genes; Hatched squares show the dichotomy between Fish and Mammalian genes (all ca. 400 million years ago). Circles indicate dichotomy between Human and Xenopus lineages (ca. 400 million years ago) and the star between HSRY and MSRY indicates the Primates/Rodents dichotomy (100 million years ago).

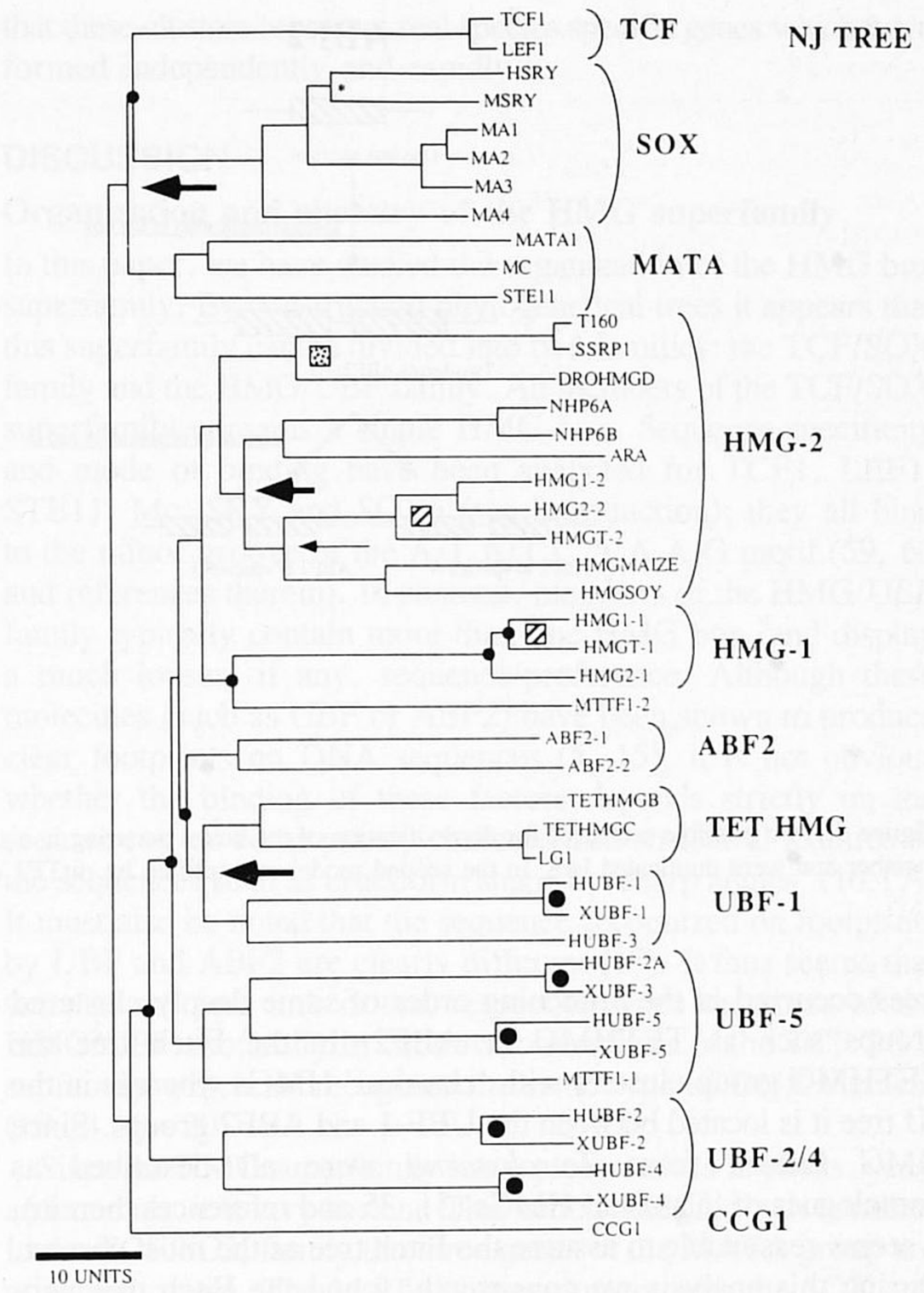

Figure 4. Unrooted phylogenetic Neighbour-Joining tree for the HMG box superfamily. The bar represents a branch length of 10 units which reflects sequence divergence. Groups are indicated by brackets on the right. The symbols are the same as in the Fig. 3 except for the small black circles which point out the important differences between Fitch and NJ trees.

two programs allow a division of the HMG superfamily into two subfamilies and a subdivision into 11 groups of genes (see Fig. 2). The two families are *(i)* the TCF, MATA, SOX family which will be referred to as the TCF/SOX family and *(ii)* a family which comprises all other HMG boxes from 'classical' HMG to UBF factors which will be referred to as the HMG/UBF subfamily. The 11 groups are the following: TCF, MATA, and SOX in the TCF/SOX family and CCG1, ABF2, HMG-2 (our terminology for box 2 of 'classical' HMG proteins), TET-HMG ('classical' HMG from Tetrahymena), HMG-1 (box 1 of 'classical' HMG) and finally UBF-1, UBF-5 and UBF-2/4 which correspond to the various HMG boxes from the Xenopus and Human UBF (see Figure 2, 3 and 4). The groups were invariable between the trees constructed with the NJ or the Fitch program, with the exception of MTTF1−2 which belongs to the HMG-1 group for NJ tree but not for the Fitch tree. This strongly argues for the validity of the trees.

The trees obtained by the Fitch (Fig.3), or the NJ (Fig.4) methods were compared and the result of this comparison is presented in Fig.2. As the composition of the groups of genes is identical with both methods only group names were indicated in the Figure 2. The most important differences between the two

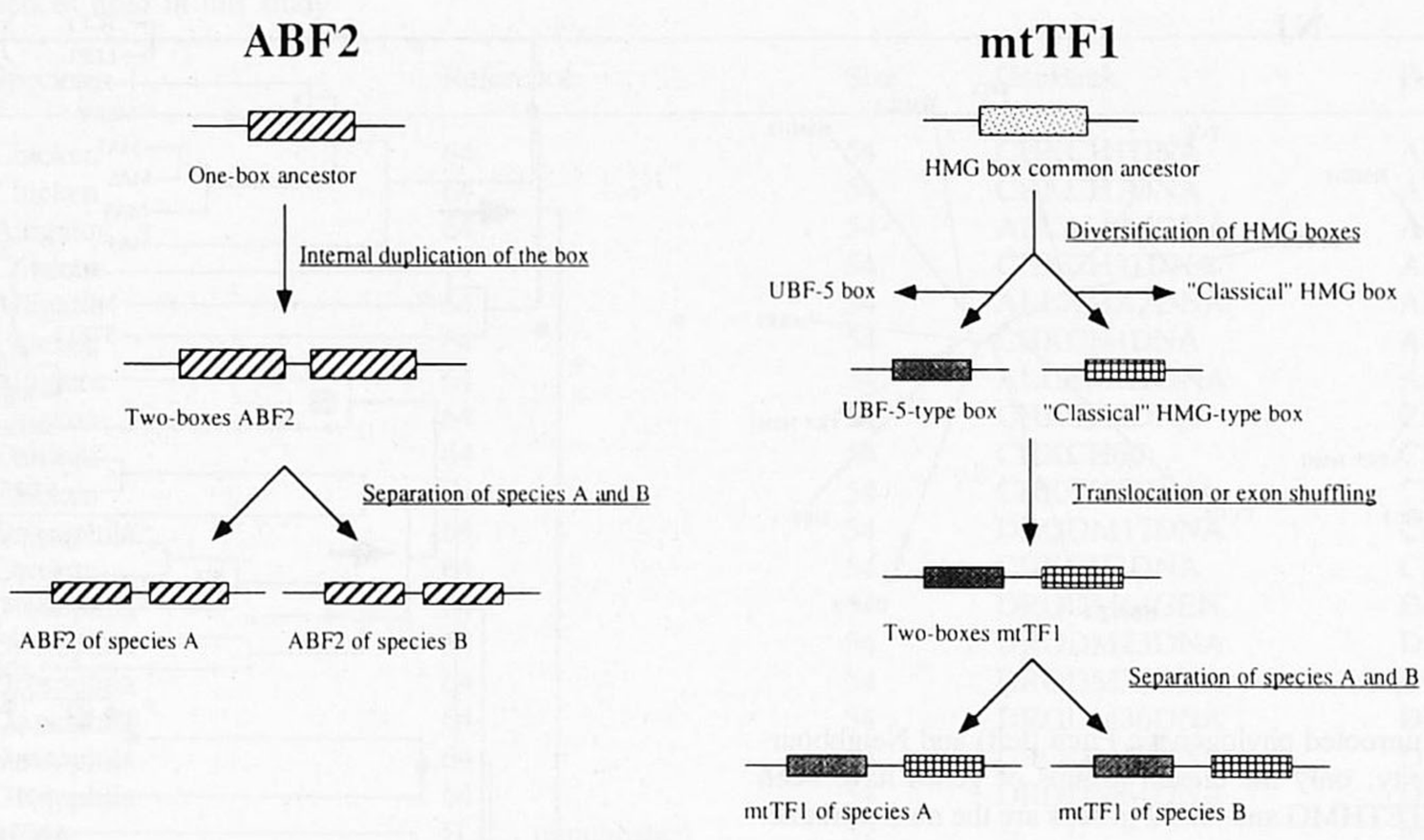

Figure 5. Two possible scenarios for diversification of the boxes occuring in a single factor. In the case exemplified by ABF2 (left), the two boxes are highly related together and were duplicated late. In the second model exemplified by mtTF1 (right) the story is much more complex and the two boxes have different ancestors.

trees occurred in the branching order of some deeply clustered groups such as TETHMG or ABF2. In the Fitch tree the TETHMG group clusters with 'classical' HMGs whereas in the NJ tree it is located between the UBF-1 and ABF2 groups. Since HMG factors from *Tetrahymena* were all described as homologues of 'classical' HMGs (31, 35 and references therein), it seems reasonable to assume the Fitch tree as the most correct. During this analysis we consistently found the Fitch tree very less sensitive to the addition of new members or to slight variations in the alignment than the NJ tree (not shown). Furthermore, in the analysis of nuclear receptors or ETS family members we also consistently found the Fitch trees to be more valuable and solid than NJ trees (23, 34).

It is important to mention that the trees presented in Fig.2 are unrooted *i.e.* that the position of the putative common ancestor to all the HMG boxes is not known. As we have no correct outgroup (i.e. genes distantly related to HMG box) we cannot determine the position of this ancestor. Although it is tempting to place the root between the two subfamilies (TCF/SOX and HMG/UBF), this remains a pure speculation and for that reason the trees presented in Fig.3 and 4 are also unrooted. For these reasons it is hard to know whether CCG1 belongs to the HMG/UBF or to the TCF/SOX subfamilies.

Examination of the trees (Fig. 3 and 4) leads to other interesting observations. Firstly, using the fact that HMG box sequences have been described in a broad spectrum of species (from vertebrates to insects, yeast and even plants), it is possible to estimate the age and the speed of diversification of the superfamily. The dichotomies between sequences originating from different organisms are indicated in the Fig. 3 and 4. This clearly illustrates the age of the HMG box superfamily. Secondly, comparison of the evolution of different boxes occurring in a single factor allows us to propose different diversification models for the appearance of these boxes in the final product. Two extreme cases of intra gene diversification are given in Fig.5: in the case of ABF2 (15) the two boxes are duplicated from a single box ancestor to form the final product; in the case of mtTF1 (13) the two boxes were duplicated very early on and subsequently joined on the same gene product. Thirdly, we have studied the distribution of the acidic tail which is sometimes present in the C-terminal part of various HMG box proteins. This structure is present in the 'classical HMGs' from plants or animals as well as in the UBF proteins from human or Xenopus but not in other members of the HMG/UBF subfamily such as T160, TET HMG, mtTF1 or NHP6. From this distribution it seems clearly that this acidic tail was acquired independently by 'classical HMGs' and UBF factors during evolution.

The SOX family of genes

In the course of this study, we were surprised to find a large number of new genes which are related to the SRY sex determination gene (9, 64; see Table 2). These SOX genes have been cloned from mammals as well as birds, various reptilias, Amphibians and Insects (M.Van de Wetering and H.C. unpublished results; 8, 20, 36, 37; see also the sequences obtained by Piccardo et al., and deposited on Genbank, ref 64) and likely represent at least 40 different loci. The majority of the sequences are incomplete since PCR primers used for cloning were based on consensus sequences within the HMG box. In order to clarify the relationships between these members, we decided to align all the SOX genes sequences and to construct a phylogenetic tree from this alignment.

From the alignment performed with the Clustal V program (data not shown), we constructed Fitch (Fig.6) or NJ trees (not shown). The precise topology of these trees must be regarded only as preliminary since most of the sequences are incomplete. To avoid gross artefacts, we constructed trees with the two programs (Fitch and NJ) and with two different alignments: one containing the entire HMG box with gaps at the beginning and at the end of the sequences when information was not available and the other where only the smallest common part to all SOX sequences was used. The four different trees constructed allowed us to divide the SOX family into various groups.

From trees such as the Fitch tree constructed from the sequences of all the SOX genes presented (Fig.6) it is interesting to note that the evolutionary pattern of some members of this

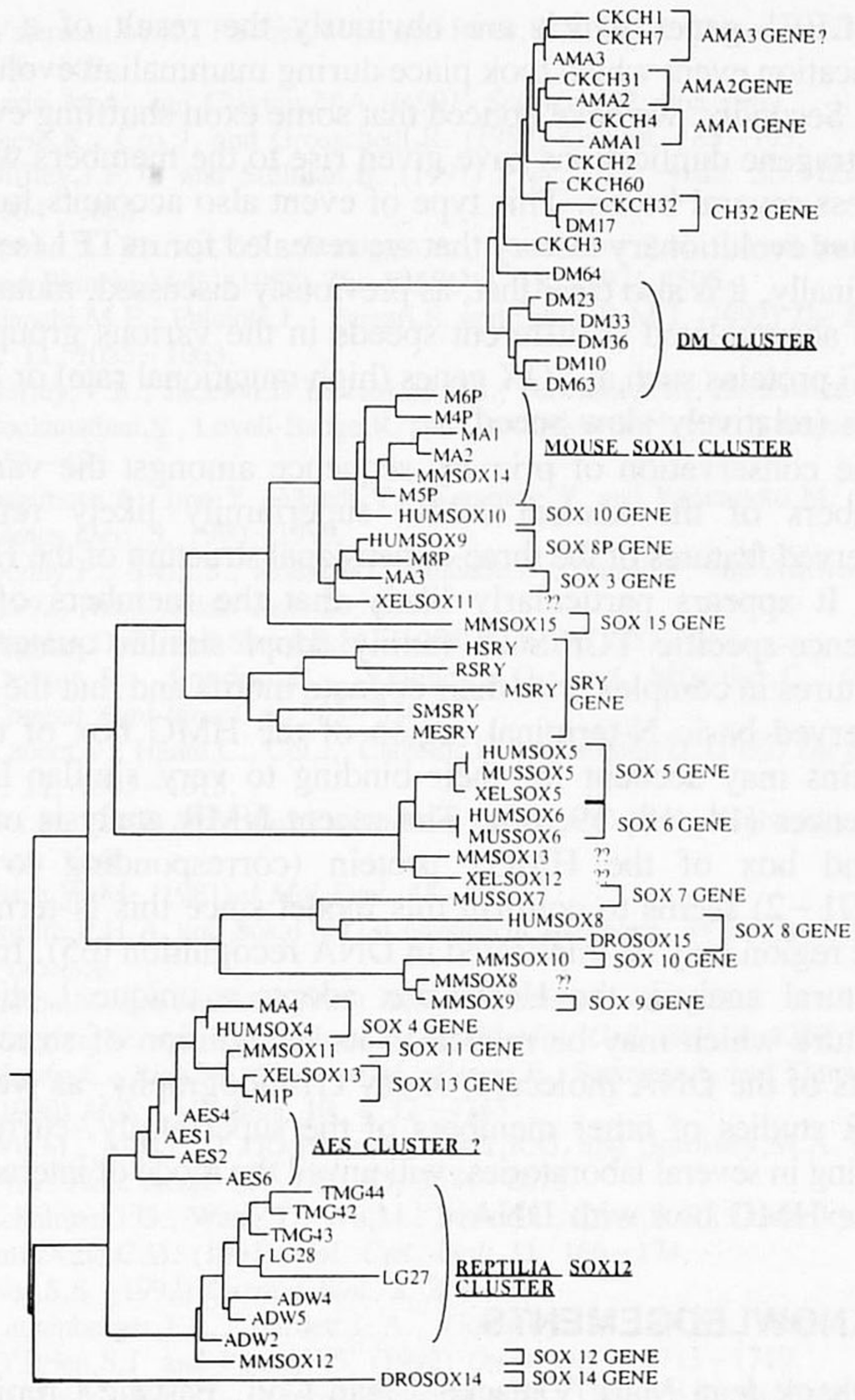

Figure 6. Rooted phylogenetic Fitch tree for the SOX-related genes. Clusters were defined as groups of genes from the same species or from very related species (such as Alligator and Gecko in the Reptilia cluster). In such cases it is not known whether sequences coming from other species were missing for other species due to a 'sampling artefact' (i.e. these genes exist but are not yet isolated) or if these clusters are indeed species-specific. In the later case we must assume explosive 'hot spot periods' of gene duplication during the evolution of these cluster-containing species.

family is unique. In classical cases, the various homologous members from different species of one given locus are clearly rooting together. This is the case for the CKCH31 and AMA2 sequences in Fig. 6 which are probably the Chicken and Alligator versions of one unique locus (the various numbers only reflecting a puzzling terminology). But, in several cases within the SRY-related genes we can observe sequences which do not fit with such a type of classification. This is for instance the case of the various Drosophila sequences called DM64, DM23, DM33, DM36, DM10 or DM63 (see the 'DM cluster' in Fig. 6; other clusters of that type are also shown for mouse or Reptilians sequences in that Fig. 6). Such a cluster of genes may have two different origins: First, it may represent a sampling artefact. This interpretation assumes that we presently lack the sequences for the vertebrates homologues of these Drosophila genes or that (as noted by the authors) some of the very close homologues inside these clusters represent sequence artefacts caused by the PCR amplification process (64). An alternate interpretation may be that these clusters represent real species specific genes which were formed independently and rapidly.

DISCUSSION

Organization and ancestry of the HMG superfamily

In this paper, we have studied the organization of the HMG box superfamily. By constructing phylogenetical trees it appears that this superfamily can be divided into two families: the TCF/SOX family and the HMG/UBF family. All members of the TCF/SOX superfamily contains a single HMG box. Sequence-specificity and mode of binding have been analyzed for TCF1, LEF1, STE11, Mc, SRY and SOX5 (see Introduction); they all bind to the minor groove of the A/T A/T C A A A G motif (59, 60 and references therein). In contrast, members of the HMG/UBF family typically contain more than one HMG box, and display a much looser, if any, sequence-preference. Although these molecules (such as UBF or ABF2) have been shown to produce clear footprints on DNA sequences (5, 15), it is not obvious whether the binding of these factors depends strictly on the sequence or on some poorly characterized structural features of the sequences such as cruciform shape or 'sharp angles' (16,17). It must also be noted that the sequence recognized on footprints by UBF and ABF2 are clearly different (15). It thus seems that there is no clear link in that respect between members of the HMG/UBF subfamily in contrast to members of the TCF/SOX subfamily which all recognize the same type of DNA element (10,14,18–20).

Since the HMG superfamily has representatives in plants, yeast and animals, it is very ancient. This notion holds for both families. In the TCF/SOX family the three genes of the MATA group are of fungal origin. The MATA group clusters with the SOX and TCF groups indicating that the separation between the fungal genes and SOX and TCF genes predates the divergence of fungal and animal lineages. Since this dichotomy occured approximately 1,000 million years ago (38, 39), it implies that the three groups of the TCF/SOX family existed before this date. The same conclusion can be drawn for the 'classical' HMG and UBF with regard to the dichotomy between yeast NHP6 genes and the second box of 'classical' HMG. The putative HMG box ancestor should predate the appearance of the TCF/SOX and the UBF/HMG families and is thus obligatory older than 1,000 million years.

Few transcription factor families have been studied from an evolutionary point of view. Two of these families, the nuclear receptors (23, 40), and the ETS oncogene family (33, 34) are known only in animals. This makes it difficult to know if these families are older than 500–600 million years. In contrast, homeobox genes have been cloned from plants and animals suggesting that this family, like the HMG box superfamily, is much older (41–43). The same type of conclusion can be drawn for the jun family since the GCN4 factor is a yeast homologue of the jun oncogenes (44).

Speed of diversification of HMG superfamily members

It is interesting to note that the speed of evolution (i.e. of accumulation of mutations in newly duplicated genes) is very different from one gene group to another. In trees such as those presented in Fig. 3 and 4, the length of the horizontal branches is not proportional to time, but to the divergence between sequences. This means that a long branch will be obtained in a short time period if the two sequences have diverged very

rapidly. Alternatively, when two sequences accumulate mutations very slowly, the short branches connecting them may correspond to long periods of times. Thus, it is difficult to assess the speed of evolution without knowing the precise dates corresponding to the different dichotomies. As the HMG superfamily contains homologues belonging to different species, we can use the dichotomies between these homologues to date some parts of the trees. For example, the dichotomies between human and Xenopus UBF box sequences correspond to the dichotomy between mammals and amphibians, *i.e.* approx 400 million years (45). This 400 million year periods corresponds to 6% divergence between human and Xenopus UBF box 1 so to 1.5% divergence per 100 million years (1.5%/100MY). The same calculation resulted in 21% divergence in 400 million years for human and Xenopus box 5 thus 5%/100MY. This means that box 5 evolved approximately 3−4 fold more rapidly than box 1. The same calculation can be made for other parts of the trees. It gives a rough estimate of sequence divergence for each gene group when it is possible to date at least one dichotomy in the group. For example, we can observe by comparing the speed of evolution between HMG1-2 or HMG2-2 to either HMGT-2 (HMG box from trout, time of divergence with mammals of approx 500 million years) or maize or soybean 'classical' HMG (from plants, time of divergence with animals of approx 1,000 million years) that the speed of evolution in this group was fairly constant and may be estimated to 6%/100MY i.e. twice that for the HMG box 1 (3%/100MY). By comparison, the speed of evolution between mouse and human SRY is 30%/100MY since the rodent and primates were separated 100 million years ago (46). It means that the evolution of SRY was 5 to 10 fold more rapid than the speed of evolution of 'classical' HMGs.

The fact that two boxes contained within the same protein may diverge at different speeds highlights the fact that these boxes evolved independently. From our tree it is clear that the 5 boxes shared by UBF factors from human and Xenopus or the two boxes from 'classical' HMGs originated very early in evolution. It is probable that these boxes have always been joined since all UBF or 'classical' HMGs boxes share a common and specific ancestor. Nevertheless, this rule (illustrated in Fig. 5A) has at least one exception: the two boxes of the mitochondrial mtTF1 protein. The first box of this protein belongs to the UBF-5 group while the second box belongs to the divergent HMG-1 group. This situation is reminiscent of the two conserved domains of some nuclear receptors such as the vitamin D or the ecdysone receptors (23). In these cases, the two domains belong to different subfamilies of nuclear receptors suggesting that a kind of recombination or exon shuffling event (47) has joined them to form a new gene which is an 'evolutive chimaera'. It seems probable that one such event created the mtTF1 gene by juxtaposing a UBF-5-like box and an HMG-1-like box. The resulting chimaera was then retained by natural selection during the evolution of the organisms. This type of shuffling mechanism may represent a simple way to increase gene diversity during evolution and is reminiscent of the recently proposed 'overprinting' gene diversification model (48).

Mechanisms of evolution of the HMG box superfamily

Our analysis suggests that the presently observed diversity of the HMG box superfamily was acquired by at least three types of mechanisms acting cooperatively. The first type of mechanism is gene duplication which took place very often during the evolution of the family. This is for example the case for the TCF1 and LEF1 genes which are obviously the result of a gene duplication event which took place during mammalian evolution (66). Secondly, we have noticed that some exon shuffling events or intragene duplications have given rise to the members which possess several boxes. This type of event also accounts for the curious evolutionary history that we revealed for mtTF1 (see fig 5). Finally, it is also clear that, as previously discussed, mutations were accumulated at different speeds in the various groups of HMG proteins such as SOX genes (high mutational rate) or UBF genes (relatively slow speed).

The conservation of primary sequence amongst the various members of the ancient HMG superfamily likely reflects conserved features of the three-dimensional structure of the HMG box. It appears particularly likely that the members of the sequence-specific TCF/SOX family adopt similar quaternary structures in complex with their cognate motifs and that the well conserved basic N-terminal region of the HMG box of these proteins may account for their binding to very similar DNA sequences (14, 18, 59, 60). The recent NMR analysis of the second box of the HMG1 protein (corresponding to our HMG1−2) seems to confirm this model since this N-terminal basic region may be implicated in DNA recognition (65). In this structural analysis the HMG box adopts a unique L-shaped structure which may be related to its recognition of structural motifs of the DNA molecule. X-ray cristallography, as well as NMR studies of other members of the superfamily, currently ongoing in several laboratories, will unveil the mode of interaction of the HMG box with DNA.

ACKNOWLEDGEMENTS

We thank Jean Marc Vanacker, Jean Coll, Pascale Crepieux, Catherine Hänni and Dominique Leprince for critical reading of the manuscript. We are also grateful to Nicole Devassine and Marie Christine Bouchez for patient typing. We thank Claire Valencien from the CITI-2/Bisance network for computer help. V.L. holds a fellowship from the French Ministère de la Recherche et de la Technologie. H.C. holds a Pionier-grant from the Dutch scientific organization 'NWO'. We thank the Association pour la Recherche contre le Cancer, the Centre National de la Recherche Scientifique, and the Institut Pasteur de Lille for financial support.

REFERENCES

1. Evans,R.M. and Hollenberg,S.M.(1988), *Cell*, **52**, 1−3.
2. Landschulz,W.H., Johnson,P.F. and McKnight,S.L. (1988) *Science* **240**, 1759−1764.
3. Levine, M. and Hoey, T. (1988) *Cell*, **55**, 537−540.
4. Murre, C., McCaw,P.S., and Baltimore,D. (1989) *Cell*, **56**, 777−783.
5. Jantzen,H-M., Admon,A., Bell,S.P. and Tjian,R. (1990) *Nature* **344**, 830−836.
6. Kelly,M., Burke,J., Smith,M., Klar,A. and Beach,D. (1988) *EMBO J.* **7**, 1537−1547.
7. Staben,C. and Yanofsky,C. (1990) *Proc.Natl.Acad.Sci USA* **87**, 4917−4921.
8. Gubbay,J., Collignon,J., Koopman,P., Capel,B., Economou,A., Münsterberg,A., Vivian,N., Goodfellow,P. and Lovell-Badge,R. (1990) *Nature* **346**, 245−250.
9. 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.
10. Van de Wetering,M., Oosterwegel,M., Dooijes,D. and Clevers,H. (1991) *EMBO J.* **10**, 123−132.
11. Travis,A., Amsterdam,A., Belanger,C. and Grosschedl,R. (1991) *Genes Dev.* **5**, 880−894.

12. Waterman,M.K., Fischer,W.H. and Jones,K.A. (1991) *Genes and Dev.* **5**, 656–669.
13. Parisi,M.A. and Clayton,D.A. (1991) *Science* **252**, 965–969.
14. Giese,K., Cox,J. and Grosschedl,R. (1982) *Cell* **69**, 185–195.
15. Diffley,J.F.X. and Stillman,B. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 7864–7868.
16. Ferrari,S., Harley,V.R., Pontiggia,A., Goodfellow,P.N., Lovell-Badge,R. and Bianchi,M.E. (1992) *The EMBO J.* **11**, 4497–4506.
17. Bianchi,M.E., Falciola,L., Ferrari,S. and Lilley,D.M.J. (1992) *The EMBO J.* **11**, 1055–1063.
18. Harley,V.R., Jackson,D.I., Hextall,P.J., Hawkins,J.R., Berkowitz,G.D., Sockanathan,S., Lovell-Badge,R. and Goodfellow,P.N. (1992) *Science* **255**, 453–456.
19. Sugimoto,A., Iino,Y., Maeda,T., Watanabe,Y. and Yamamoto,M. (1991) *Genes Dev.* **5**, 1990–1999.
20. Denny,P., Swift,S., Brand,N., Dabhade,N., Barton,P. and Ashworth,A. (1992) *Nucl. Acids. Res.* **20**, 2887.
21. Higgins,D.G. and Sharp,P.M. (1988) *Gene* **73**, 237–244.
22. Dessen,P., Fondrat,C., Valencien,C. and Mugnier,C. (1990) *Comput.Appl.Biosci.* **6**, 355–356.
23. Laudet,V., Hänni,C., Col,J., Catzeflis,F. and Stéhelin D. (1992) *The EMBO J.* **11**, 1003–1013.
24. Nei,M. (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York.
25. Fitch,W.M. (1981) *J.Mol.Evol.* **18**, 30–37.
26. Sneath,P.H.A. and Sokai (1973) *Numerical Taxonomy*. W.H.Freeman, San Francisco.
27. Saitou,N. and Nei,M. (1987). *Mol.Biol.Evol.* **4**, 406–425.
28. Haggren,W. and Kolodrubetz,D. (1988) *Mol. Cell. Biol.* **8**, 1282–1289.
29. Mosrin,C., Riva,M., Beltrame,M., Cassar,E., Sentenac,A. and Thuriaux,P. (1990) *Mol. Cell. Biol.* **10**, 4737–4743.
30. Wu,M., Allis,C.D., Richman,R., Cook,R.G. and Gorovsky,M.A. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 8674–8678.
31. Schulman,I.G., Wang,T., Wu,M., Bowen,J., Cook,R.G., Gorovsky,M.A. and Allis,C.D. (1991) *Mol. Cell. Biol.* **11**, 166–174.
32. Ner,S.S. (1992) *Current Biol.* **2**, 208–210.
33. Lautenberger,J.A., Burdett,L.A., Gunnell,M.A., Qi,S., Watson,D.K., O'Brien,S.J. and Papas,T.S. (1992) *Oncogene* **7**, 1713–1719.
34. Laudet,V., Niel,C., Duterque-Coquillaud,M., Leprince,D. and Stéhelin,D. (1993) *Biochem. and Biophys.Res.Com.* **190**, 8–14.
35. Hayashi,T., Hayashi,H. and Iwai,K. (1989) *J.Biochem.* **105**, 577–581.
36. Foster,J.W., Brennan,F.E., Hampikian,G.K., Goodfellow,P.N., Sinclair,A.H., Lovell-Badge,R., Selwood,L., Renfree,M.B., Cooper,D.W. and Graves,J.A.M. (1992) *Nature* **359**, 531–533.
37. Denny,P., Swift,S., Connor,F. and Ashworth,A. (1992) *The EMBO J.* **11**, 3705–3712.
38. Knoll,A.H. (1992) *Science* **256**, 622–627.
39. Erwin,D.H. (1991) *TREE* 6, 131–134.
40. Amero,S.A., Kretsinger,R.H., Moncrief,N.D., Yamamoto,K.R. and Pearson,W.R. (1992) *Mol.Endo* **6**, 3–7.
41. Kappen,C., Schughart,K. and Ruddle,F.H. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 5459–5463.
42. Schughart,K., Kappen,C. and Ruddle,F.H. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 7067–7071.
43. Murtha,M.T., Leckman,J.F. and Ruddle,F.H. (1991) *Proc.Natl.Acad.Sci USA* **88**, 10711–10715.
44. Vogt,P.K. and Tjian,R. (1988) *Oncogene* **3**, 3–7.
45. Kobel,H.R. and Du Pasquier,L. (1986) *TIG* **2**, 310–315.
46. Novacek,M.J. (1992) *Nature* **356**, 121–125.
47. Patthy,L. (1991) *Cur. Opinion Genet. Dev.* **1**, 351–361.
48. Keese,P.K. and Gibbs,A. (1992) *Proc.Natl.Acad.Sci.USA* **89**, 9489–9493.
49. Shirakata,M., Hüppi,K., Usuda,S., Okazaki,K., Yoshida,K. and Sakano,H. (1991) *Mol. Cell. Biol.* **11**, 4528–4536.
50. Bruhn,S.L., Pil,P.M., Essigmann,J.M., Housman,D.E. and Lippard,S.J. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 2307–2311.
51. Wagner,C.R., Hamana,K. and Elgin,S.C.R. (1992) *Mol. Cell. Biol.* **12**, 1915–1923.
52. Kolodrubetz,D. and Burgum,A. *J.Biol.Chem.* (1990) **265**, 3234–3239.
53. Walker,J.M. (1982) In E.W.Johns (ed.) *Primary structures*. p 69–88. Academic Press, London.
54. Tsuda,K.I., Kiruchi,M., Mori,K., Waga,S. and Yoshida,M. (1988) *Biochemistry* **27**, 1197–1214.
55. Pentecost,B.T., Wright,J.M. and Dixon,G.H. (1985) *Nucleic Acids Res.* **13**, 4871–4888.
56. Grasser,K.D. and Feix,G. (1991) *Nucl. Acids. Res.* **19**, 2573–2577.
57. Laux,T. and Goldberg,R.B. (1992) *Nucl. Acids Res.* **19**, 4769.
58. Bachvarov,D. and Moss,T. (1991) *Nucl. Acids. Res.* **19**, 2331–2335.
59. Van de Wetering,M. and Clevers,C. (1992) *The EMBO J.* **11**, 3039–3044.
60. Haqq,C.M., King,C.Y., Donahae,P.K. and M.A. Weiss. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 1097–1101.
61. Hisatake,K., Hasegawa,S., Takada,R., Nakatani,Y., Horikoshi,M., and Roeder,R.G. (1993) *Nature,* **362**, 179–181.
62. Sekiguchi,T., Nohiro,Y., Nakamura,Y., Hisamoto,N. and Nishimoto,T. (1991) *Mol. Cel. Biol.*, **11**, 3317–3325.
63. Yamaguchi-Shinozaki,K. and Shinozaki, K. (1992) *Nucleic Acids Res.*, **20**, 6737.
64. Coriat, A.M., Müller,U., Harry,J.L., Uwanogho,D. and Sharpe, P.T. (1993) *PCR Meth and Applic.*, **2**, 218–222.
65. Weir, H.M., Kraulis,P.J., Hill,C.S., Raine,A.R.C., Laue,E.D. and Thomas,J.O. (1993) *The EMBO J.* **12**, 1311–1319.
66. Gastrop,J., Hoevenagel,R., Young,J.R. and Clevers,H.C. (1992) *Eur. J. Immunol.* **22**, 1327–1330.
67. Wright,E.W., Snopek,B. and Koopman,P. (1993) *Nucleic Acids Res.*, **21**, 744
68. Van de Wetering,M. and Clevers,H.C. (1993) in press.