The HMG-1 box protein family: classification and functional relationships

Andreas D. Baxevanis and David Landsman*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, Room 8N-805, Bethesda, MD 20894, USA

Received November 30, 1994; Revised and Accepted February 27, 1995

ABSTRACT

The abundant and highly-conserved nucleoproteins comprising the high mobility group-1/-2 (HMG-1/-2) family contains two homologous basic domains of about 75 amino acids. These basic domains, termed HMG-1 boxes, are highly structured and facilitate HMG-DNA interactions. Many proteins that regulate various cellular functions involving DNA binding and whose target DNA sequences share common structural characteristics have been identified as having an HMG-1 box; these proteins include the RNA polymerase I transcription factor UBF, the mammalian testis-determining factor SRY and the mitochondrial transcription factors ABF2 and mtTF1, among others. The sequences of 121 HMG-1 boxes have been compiled and aligned in accordance with thermodynamic results from homology model building (threading) experiments, basing the alignment on structure rather than by using traditional sequence homology methods. The classification of a representative subset of these proteins was then determined using standard least-squares distance methods. The proteins segregate into two groups, the first consisting of HMG-1/-2 proteins and the second consisting of proteins containing the HMG-1 box but which are not canonical HMG proteins. The proteins in the second group further segregate based on their function, their ability to bind specific sequences of DNA, or their ability to recognize discrete non-B-DNA structures. The HMG-1 box provides an excellent example of how a specific protein motif, with slight alteration, can be used to recognize DNA in a variety of functional contexts.

INTRODUCTION

Various metabolic processes in which DNA is involved are dependent upon many different types of non-histone chromosomal proteins. The most prevalent of these non-histone nucleoproteins in eukaryotic cells are the high mobility group, or HMG proteins. These proteins were originally operationally defined by their physical properties, such as their behavior on SDS-polyacrylamide gels (1), their ability to be extracted from chromatin

in 0.35 M NaCl and their solubility in 5% perchloric or trichloroacetic acid (2,3). While all HMG proteins share these characteristics, they can be divided into three groups based on their size, sequence similarities and DNA binding properties: HMG-1/-2, HMG-14/-17 and HMG-I/Y (3). Of these groups, interest has focused lately on the highly-conserved HMG-1/-2 proteins and their homologs.

HMG-1 and HMG-2 have a molecular weight of ~25 000 and have a tripartite structure (4-6). The N-terminal A-domain (residues 1-79) and central B-domain (residues 90-163) are both highly basic and substantially structured. Circular dichroism studies indicate that domains A and B contain 30 and 50% β -helix, respectively (4,5). In contrast, the C-domain, beginning at position 164, contains a very high proportion of acidic amino acids, including a run of 30 aspartic and glutamic acid residues that ends at the C-terminus. Domains A and B are homologous to one another, each possessing a conserved block of 75 residues containing a signature of aromatic and basic amino acids termed the HMG-1 box (7). Recent NMR spectroscopy studies on the structure of the HMG-1 box in the B-domain of HMG-1 (8,9), in the *Drosophila* HMG-D protein (10) and in the SRY-related protein SOX-5 (11) show that the box consists of three α -helices, accounting for 75% of the total residues in the box. Furthermore, the HMG-1 box is L-shaped, the angle of ~80° between the two arms being stabilized by the conserved aromatic residues (8,10). It is possible that this structure facilitates the binding of HMG-1 to DNA.

HMG-1/-2 can bind to both single- and double-stranded DNA, with a preference for the former (2,12). Interestingly, these proteins can also bind to a variety of non-B-DNA structures, such as B-Z DNA junctions (13) and platinated DNA (14,15). It has been suggested that these proteins may play a role in the stabilization or sealing of loop structures in chromatin through their binding to these non-B-DNA structures (16). While the overall structure of HMG-1 is similar to that of many transcriptional activators, it has been shown that HMG-1 is incapable of activating transcription in transfected yeast cells (17), although it can promote the formation of transcription initiation complexes of RNA polymerases II and III (18,19).

Increasing numbers of proteins that regulate various cellular functions involving DNA binding and whose target DNA sequences share common structural characteristics are being identified as containing the HMG-1 box. These proteins include

* To whom correspondence should be addressed

the RNA polymerase I transcription factor UBF (7,20–24), the mammalian testis-determining factor SRY (25,26), and the mitochondrial transcription factors ABF2 and mtTF1 (27,28), among others. The number of HMG-1 boxes found in each protein varies; most proteins have a single HMG-1 box, but proteins such as the UBFs can contain as many as six. Based on this diverse group of proteins, a distinct common signature was deduced for the HMG-1 box (29). The diversity in the cellular roles of these proteins strongly suggests that this common signature (and, therefore, this common structural domain) can be used to accomplish a wide variety of productive protein–DNA interactions.

In this study, the sequences of 121 full-length HMG-1 boxes have been compiled and aligned in accordance with thermodynamic results from homology model building (threading) experiments (30), basing the alignment on structure rather than by using traditional sequence homology methods. This new structural alignment was then used to determine the clustering relationships between the HMG-1 box proteins. The proteins cluster into groups consisting primarily of either HMG-1/-2 proteins or non-HMG proteins containing the HMG-1 box. The proteins in the second group further segregate based on their function, their ability to bind specific sequences of DNA, and their ability to recognize discrete non-B-DNA structures. The large number of proteins identified as containing the HMG-1 box and the range of cellular functions in which they are involved points to a critical role for this conserved DNA-binding region.

METHODS

Database searches and sequence alignment

The Swiss-Prot version 29.0 (31), PIR version 41.0 (32), EMBL release 39.0 (33) and GenPept version 83.0 (34) databases were searched using the BLAST (35) algorithm, with the human HMG-1 sequence used as the basis for comparison (36). BLAST search cutoffs used to identify the homologs of human HMG-1 were a Karlin–Altschul score for two aligned sequences ≥ 70 with a probability $\leq 10^{-3}$. These database searches identified 121 proteins containing the target HMG-1 box motif. Twenty-five of these proteins contain more than one HMG-1 box (Table 1), bringing the total number of HMG-1 boxes to 164. Sequence fragments were eliminated from the data set, leaving 121 full domain-length sequences for analysis. The energetic information from homology model building (threading) experiments was used to perform an alignment of all 121 of the full-length HMG-1 boxes (30). Final positioning of residues within core segment regions represent the alignments producing optimal threading energies (37). The alignment minimizes the total number of gaps present and is consistent with the previously-reported signature for the HMG-1 box (29). ALSCRIPT version 1.4.4 (38) was used to format the final alignment (Fig. 1). The sequences used in this study can be accessed through the World Wide Web (<http://www.ncbi.nlm.nih.gov/Baxevani/HMG>) and are available for anonymous ftp at the National Center for Biotechnology Information ([ncbi.nlm.nih.gov, directory /pub/baxevanis/HMG](http://ncbi.nlm.nih.gov/directory/pub/baxevanis/HMG)).

Classification analysis

Tree diagrams were constructed using algorithms contained within the PHYLIP Phylogeny Inference Package, version 3.5c (39). Specifically, 57 of the 164 HMG-1 boxes found in the database

searches described above were selected to form a representative subset of the entire class. PROTDIST was used on the 57 boxes in this subset to calculate a distance matrix according to the Dayhoff PAM probability model (40). The distances computed represent the expected fraction of amino acid substitutions between each pair of sequences. The distance matrix was then used to estimate phylogenies using the Fitch–Margoliash least-squares distance method (41). Under this method, the sum of the branch lengths between any two species is expected to equal the distances between species found in the calculated matrix. Also, no evolutionary clock is implied using this method; assumption of an evolutionary clock implies that all of the sequences from all of these different proteins have changed at the same rate, an assumption which *a priori* cannot be supported. The results from the FITCH runs were selected over those obtained using the protein parsimony method PROTPARS and the neighbor-joining method NEIGHBOR (data not shown) due to the rigor of the FITCH method: the parsimony method does not support global rearrangements (below), providing a less exhaustive search for the best tree as does FITCH. In the neighbor-joining method, global rearrangement is again not supported and, more importantly, simulation studies indicate that NEIGHBOR does not provide as accurate an estimate of the phylogeny as is obtained using FITCH (39).

All FITCH runs were performed with global rearrangement and multiple jumbles (reordering of the data set 25 times) to evaluate the effect of different input orders on the derived trees and to assure that none of the subtrees has become caught in a region of the tree representing a statistical local minimum. Each data set was examined five times in this fashion, producing trees with identical sum-of-squares and average percent standard deviation statistics. CONSENSE was used to compute the consensus tree by the majority-rule method. The final unrooted tree diagram was generated using TREETOOL, a free-standing editor and tree formatter (42).

RESULTS AND DISCUSSION

Database searches and sequence alignment

In order to identify proteins belonging to the HMG-1 superfamily, the GenPept, EMBL, PIR and Swiss-Prot peptide sequence databases were searched using the BLAST (35) program to identify sequences significantly similar that of human HMG-1 (36). In this way, 121 proteins were identified as containing HMG-1 box sequences (Table 1). The proteins are from a variety of sources, ranging from simple organisms such as the protozoan *Tetrahymena* and yeast to plants and animals. More interestingly, the proteins are involved in a variety of cellular functions involving DNA recognition. Upon further inspection, 25 of the 123 proteins were found to contain more than one HMG-1 box, bringing the total number of HMG-1 boxes to 164. Of these 164, only 121 of the sequences were of full-domain length, and only these 121 sequences were used in the analysis that follows. Each HMG-1 box was treated as a separate unit, allowing them to be analyzed independently of other HMG-1 boxes present within the same protein.

Multiple alignment of the HMG-1 boxes was performed based on the results of threading experiments that used the structure of the HMG-1 box in the B-domain of HMG-1 as a model for protein structure prediction (30). The method of alignment used here is significantly different than those traditionally used in that the sequences are aligned based on their predicted structure rather

Figure 1. Multiple sequence alignment of HMG-1 box domains from members of the HMG-1 superfamily. The abbreviation for each protein corresponds to those listed in Table 1. Decimals appended to the protein name denote multiple HMG-1 boxes within the same protein, the numbers being assigned in the same order as the HMG-1 boxes appear within the protein (N- to C-terminal). The numbering scheme at the top of the figure refers to amino acid positions within the HMG-1 box. For each subgroup, the most prevalent residue at each position is shown in inverse type. The positions of the three α -helices found in the NMR study (8) are shown at the bottom of the figure. ALSCRIPT version 1.4.4 (38) was used to format the alignment.

than through the use of sequence similarity scores or other established procedures [cf.(43) for a review of various structure-based protein modeling methods]. This method produces an alignment with few gaps which also remains compatible with the HMG-1 box signature described above (29). The final alignment of the HMG-1 boxes is presented in Figure 1, with the positions

of the three α -helices determined in the ^1H -NMR study (8) and the positions of the core segments defined in the threading experiments shown at the bottom of the figure. (Most of the sequences from the SOX family of proteins were omitted due to the overwhelming similarity of most of the SOX-related proteins, particularly those from *Drosophila* and alligator.)

Table 1. The HMG-1 superfamily

Abbreviation	Protein/locus	Species/organ	No. of boxes	Accession no.	References
ABF2	ARS-binding factor 2	<i>Saccharomyces cerevisiae</i>	1	sp Q02486	(27)
ADW2	ADW2	Alligator	1	gb M86310	(66)
ADW4	ADW4	Alligator	1	gb M86311	(66)
ADW5	ADW5	Alligator	1	gb M86312	(66)
AES1	AES1	Alligator	1	gb M86313	(66)
AES2	AES2	Alligator	1	gb M86314	(66)
AES4	AES4	Alligator	1	gb M86315	(66)
AES6	AES6	Alligator	1	gb M86316	(66)
AMA2	AMA2	Alligator	1	gb M86318	(66)
AMA3	AMA3	Alligator	1	gb M86319	(66)
Amph	Amphoterin	Rat	2	sp P27428	(67)
atHMG	HMG	<i>Arabidopsis</i>	1	pir S35511	(68)
bbHMG1rel	HMG-containing protein 1	<i>Babesia bovis</i>	1	pir JQ1490	(69)
bHMG1	HMG-1	Bovine	2	sp P10103	(70)
btHMG1	HMG-1	Bovine testis	1	gb M26110	(71)
cCIIDBP	CIIDBP	Chicken	1	gb L08815	(72)
cfXSX5	XSX-5	<i>Xenopus</i>	1	emb X65653	(25)
cfXSX11	XSX-11	<i>Xenopus</i>	1	emb X65654	(25)
cfXSX12	XSX-12	<i>Xenopus</i>	1	emb X65655	(25)
cfXSX13	XSX-13	<i>Xenopus</i>	1	emb X65656	(25)
CH1	CH1	Chicken	1	gb M86320	(66)
CH3	CH3	Chicken	1	gb M86322	(66)
CH4	CH4	Chicken	1	gb M86325	(66)
CH31	CH31	Chicken	1	gb M86323	(66)
CH60	CH60	Chicken	1	gb M86326	(66)
Chilo	HMG homolog	Chilo iridescent virus	1	gb L22300	(73)
cHMG1	HMG-1	Chicken	2	sp P36194	(74)
cHMG2	HMG-2	Chicken	2	sp P26584	(75)
cHMG2a	HMG-2a	Chicken	2	pir S22359	(76)
cTCF	T cell-specific transcription factor	Chicken	1	—	(77)
ctHMG1a	HMG-1a	<i>Chironomus tentans</i>	1	gb M93253	(78)
ctHMG1b	HMG-1b	<i>Chironomus tentans</i>	1	pir A43436	(78)
delta	delta	<i>Tetrahymena thermophila</i>	1	gb M87306	(64)
DM10	DM10	<i>Drosophila</i>	1	gb M86328	(66)
DM17	DM17	<i>Drosophila</i>	1	gb M86329	(66)
DM23	DM23	<i>Drosophila</i>	1	gb M86330	(66)
DM33	DM33	<i>Drosophila</i>	1	gb M86331	(66)
DM36	DM36	<i>Drosophila</i>	1	gb M86332	(66)
DM63	DM63	<i>Drosophila</i>	1	gb M86333	(66)
DM64	DM64	<i>Drosophila</i>	1	gb M86334	(66)
dmDSOX14	DSOX-14	<i>Drosophila</i>	1	emb X65667	(25)
dmDSOX15	DSOX-15	<i>Drosophila</i>	1	emb X65668	(25)
DSP1	DNA-binding protein DSP1	<i>Drosophila</i>	2	gb U13881	(79)
DssRP	Single-strand recognition protein	<i>Drosophila</i>	1	sp Q05344	(47)
FPR1	FPR1	<i>Podospira anserina</i>	1	sp P35693	(80)
haHMG1	HMG-1	Chinese hamster	2	sp P07156	(81)
hHMG1	HMG-1	Human	2	sp P09429	(36)
hHMG1ret	HMG-1retropseudogene	Human	1	gb L08048	(82)
hHMG2	HMG-2	Human	2	sp P26583	(83)
hHMG2B	HMG-2B	Human	1	emb Z17240	(84)
HMGB	HMG-B	<i>Tetrahymena thermophila</i>	1	pir B39668	(85)
HMGC	HMG-C	<i>Tetrahymena thermophila</i>	1	pir A39668	(85)
HMGD	dHMG-D	<i>Drosophila</i>	1	pir A44382	(86)
HMGZ	dHMG-Z	<i>Drosophila</i>	1	pir S32725	(87)
hSOX4	SOX-4	Human	1	emb X65661	(25)
hSOX5	SOX-5	Human	1	emb X65662	(25)
hSOX6	SOX-6	Human	1	emb X65663	(25)
hSOX8	SOX-8	Human	1	emb X65664	(25)
hSOX9	SOX-9	Human	1	emb X65665	(25)
hSOX10	SOX-10	Human	1	emb X65666	(25)

Table 1. continued

hSOX11	SOX-11	Human	1	sp P35716	(88)
hSOX12	SOX-12	Human	1	sp P35717	(88)
hSRY	Testis-determining factor	Human	1	sp Q05066	(89)
hSRYmut	Testis-determining factor peptide mutant	Human	1	gb S53156	(26)
hUBF1	Upstream binding factor 1	Human	6	sp P17480	(7)
IRE-ABP	Insulin-response element binding protein	Rat	1	pir S19597	(52)
Ixr1	Ixr1	<i>Saccharomyces cerevisiae</i>	2	sp P33417	(90)
LEF-1	Lymphoid enhancer binding factor 1	Mouse	1	sp P27782	(91)
LG-1		<i>Tetrahymena thermophila</i>	1	sp P11873	
LG27	LG27	<i>Eublepharis macularis</i>	1	gb M86335	(66)
LG28	LG28	<i>Eublepharis macularis</i>	1	gb M86336	(66)
marsTDF	Testis-determining factor	<i>Sminthopsis macroura</i>	1	sp P36395	(92)
mat1-Mc	Mating type protein mat1-Mc	<i>Schizosaccharomyces pombe</i>	1	sp P10840	(93)
Mat1-2	Mating type protein Mat1-2	<i>Cochliobolus heterostrophus</i>	1	pir S34810	(94)
MG42	MG42	<i>Tarentola mauritanica</i>	1	gb M86337	(66)
MG43	MG43	<i>Tarentola mauritanica</i>	1	gb M86338	(66)
mHMG1	HMG-1	Mouse	2	sp P27109	(95)
mHMG2	HMG-2	Mouse	2	sp P30681	(96)
mHP1	Hypothetical protein	Mouse	1	pir S10947	(97)
mHP2	Hypothetical protein	Mouse	1	pir S10950	(97)
mSDP1	Sex-determining protein	Mouse	1	pir S10949	(97)
mSDP2	Sex-determining protein	Mouse	1	pir S10938	(97)
mSDP3	Sex-determining protein	Mouse	1	pir S10948	(97)
mSOX1	SOX-1	Mouse	1	—	(25)
mSOX2	SOX-2	Mouse	1	—	(25)
mSOX3	SOX-3	Mouse	1	—	(25)
mSOX4	SOX-4	Mouse	1	pir S22938	(25)
mSOX5	SOX-5	Mouse	1	emb X65658	(25)
mSOX6	SOX-6	Mouse	1	emb X65659	(25)
mSOX7	SOX-7	Mouse	1	emb X65660	(25)
mSRY	Testis-determining factor	Mouse	1	emb X55491	(97)
mt-a-1	Mating type protein mt-a-1	<i>Neurospora crassa</i>	1	gb M54787	(98)
mtsHMG	Testis-specific HMG	Mouse testis	2	gb L07107	(45)
mTCF1	T cell-specific factor 1	Mouse	1	pir JH0402	(99)
mtTF1	Transcription factor 1	Human mitochondria	1	sp Q00059	(28)
mUBF1	Upstream binding factor 1	Mouse	5	sp P25976	(20)
mzHMG1rel	HMG1-like protein	Maize	1	gp X66077	(100)
NHP6A	Nonhistone protein 6A	<i>Saccharomyces cerevisiae</i>	1	sp P11632	(101)
NHP6B	Nonhistone protein 6B	<i>Saccharomyces cerevisiae</i>	1	sp P11633	(101)
NOR90	Autoantigen NOR-90	Human	4	pir S18193	(102)
pHMG1	HMG-1	Porcine	2	sp P12682	(103)
pHMG2	HMG-2	Porcine	2	sp P17741	(104)
rabSRY	Testis-determining factor	Rabbit	1	—	(25)
rCIIDBP		Rat	1	pir S35637	(72)
rHMG1	HMG-1	Rat	2	sp P07155	(8,105)
ROX1	Heme-dependent repression factor	<i>Saccharomyces cerevisiae</i>	1	sp P25042	
rUBF1	Upstream binding factor-1	Rat	5	sp P25977	(21)
rUBF2	Upstream binding factor-2	Rat	4	sp P25978	(21)
SB11	SB11	Soybean	1	sp P26585	(106)
SpHMG1	HMG1	<i>Strongylocentrotus purpuratus</i>	1	gb L06453	
SSRP1	Single-strand recognition protein 1	Human	1	gb M86737	(46)
ste11	Sexual development regulator ste11+	<i>Schizosaccharomyces pombe</i>	1	sp P36631	(107)
T160	Recombination factor T160	Mouse	1	pir A41265	(48)
tHMG1	HMG-T	Trout	2	sp P07746	(108)
tpHMG	HMG	<i>Tetrahymena pyriformis</i>	1	pir JU0038	(109)
Tryp	HMG-1-like protein	<i>Trypanosoma brucei</i>	2	sp P26586	(110)
vfHMG	HMG	<i>Vicia faba</i>	1	pir S39556	(111)
whHMG	HMG	Wheat	1	pir S18991	
xUBF1	Upstream binding factor 1	<i>Xenopus</i>	5	sp P25979	(24)
xUBF2	Upstream binding factor 2	<i>Xenopus</i>	4	sp P25980	(22,23)

Figure 2. Clustering relationships between the members of the HMG-1 box family. Unrooted FITCH trees were generated using the PHYLIP Phylogeny Inference Package, version 3.5 (39) and TREE TOOL, a free-standing editor and tree formatter (42). Major subgroups (colored regions) and abbreviations correspond to those in the alignment presented in Figure 1.

It is immediately apparent that, while there may not be absolute identity between all of the sequences in question, discrete regions of high similarity are present. At the N-terminus, a short basic region precedes the beginning of the first helix, with most of the proteins having prolines at positions 1 and 4. This is followed by a group of hydrophobic residues at positions 6–11, a large proportion of which are phenylalanine, tyrosine or tryptophan; helix 1 begins at position 7, within this hydrophobic tract. With very few exceptions, there is a proline at position 24, effectively disrupting the first helix. A number of aromatic residues are highly conserved in the C-terminal region of the HMG-1 box, such as the tryptophan at position 41, a phenylalanine, tryptophan or tyrosine at positions 52 and 70; and a phenylalanine, tryptophan or histidine at position 63. The high conservation of these residues has major implications in the stability of the tertiary structure of the HMG-1 box: the NMR solution structure shows that, for rat HMG1.2 (rHMG1.2 in Fig. 1), ⁸Phe, ¹¹Phe, ⁴¹Trp, ⁴⁹Lys and ⁵²Tyr maintain the angle between the two arms of the HMG-1 box, ¹Pro and ⁴Pro at the N-terminus interacts with ⁶³Tyr

at the C-terminus, and ²⁴Pro helps to maintain the non-helical conformation between helix 1 and helix 2 (8). Threading experiments identifying the critical contacts within the HMG-1 box also confirm the involvement of most of these aromatic residues in maintaining the three-dimensional structure of the domain (30).

Classification of HMG-1 box sequences

The relationships between the proteins belonging to the HMG-1 superfamily were determined using a representative subset of 57 of the 121 aligned proteins shown in Figure 1. The non-linear least squares algorithm FITCH (41) from the PHYLIP Phylogeny Inference Package (39) were used to perform the analysis, using a Dayhoff PAM distance matrix (40) as the basis for the calculations. The resulting unrooted tree is presented in Figure 2.

The first feature regarding the distribution of the component HMG-1 boxes over the tree is that, with minor exception, the HMG-1 and HMG-2 proteins cluster away from the non-HMG

proteins containing the HMG-1 box. The HMG-1/2 proteins themselves form two separate subgroups, corresponding to HMG-1 box 1 (HMG1.1) and HMG-1 box 2 (HMG1.2). In addition to all of the HMG1.1 boxes, the HMG1.1 subgroup contains the second box from the human mitochondrial transcription factor mtTF1, boxes 2 and 5 from the human upstream binding factor UBF (hUBF1.2/1.5), and the HMG boxes from both *Tetrahymena thermophila* and *Tetrahymena pyriformis* (tpHMG and HMG-C, respectively). The HMG1.2 subgroup contains all of the HMG1.2 boxes, as well as the fourth box from UBF (hUBF1.4) and the boxes from various plant sources (maize, wheat and soybean). The plant HMG-1 boxes occupy an independent branch within the HMG1.2 subgroup, an observation that may be important from an evolutionary standpoint. The common functional feature of the proteins in the HMG1.1 and HMG1.2 subgroups is their ability to bind to non-B-DNA structures, as previously noted.

The UBF subgroup is so named since it contains three of the six HMG-1 boxes found in the UBF protein (boxes 1, 3 and 6). Also included in the UBF subgroup is the first box from mtTF1, the first box from the mouse testis-specific HMG (mtsHMG), and two boxes from *T.thermophila* (HMG-B and delta). Of these, HMG-B is found in the macronucleus of *T.thermophila*, while delta is to be localized to the micronucleus (29). As with the HMG1.1 and HMG1.2 subgroups, the members of this subgroup can recognize non-B-DNA structures, with the majority of proteins in the subgroup being involved in transcription. mtTF1 has been shown to bind in a sequence-specific fashion to a conserved segment in the D-loop region of human mitochondrial DNA (44), a sequence very similar to that bound by mtsHMG (45) (Table 2).

The SSRP subgroup has amongst its members the HMG-1 boxes from both the human and *Drosophila* single-stranded recognition proteins, the V(D)J recombinase T160 and DBP from both rat and chicken. On a separate branch within the subgroup are the boxes of the *Chironomus* and *Drosophila* HMG proteins. A common functional feature in this subgroup is the recognition of discrete non-B-DNA structures; in the case of both SSRP and DssRP, there is an elevated affinity for platinated DNA (46,47). It is not unusual for several HMG proteins to be found within such a subgroup, as the HMG proteins in general do recognize platinated DNA (14,15). It is likely that the SSRP proteins do indeed bind to specific DNA sequences, as T160 has been shown to bind specifically to the sequence CACAGTC (Table 2; 48).

The members of the SRY subgroup are unlike those of the subgroups previously discussed in that they are able to bind to DNA in a sequence-specific fashion, rather than in a conformation-specific fashion (Table 2). Included in this subgroup is the transcription repressor ROX1, the mating-type selection proteins ste11, mat1-Mc and FPR1; transcription proteins LEF-1, cTCF and mTCF1; the insulin response element binding protein IRE-ABP, the testis-determining factor marsTDF, as well as the sex-determining protein SRY for which the subgroup is named. Footprinting studies on TCF1 and SRY indicate that the recognition of their specific DNA targets occurs primarily through nucleotide contacts in the minor groove (49). A recent study confirmed the sequence-specific nature of binding through the use of chimeric constructs containing portions of HMG-1 boxes from both TCF1 and HMG-1 (50); this study proposes a model for binding where N-terminal residues and part of helix 1 of the HMG-1 box makes contact with the minor groove on the outside of a bent DNA duplex.

Table 2. Specific binding sequences recognized by HMG-1 box proteins

Protein	Binding site ^a	Reference
SRY:		
IRE-ABP	TTCAAAG	(52)
LEF-1	TTCAAAGG	(112)
mat1-Mc	AACAAAG	(113)
mt-a-1	CAAAG	(114)
ROX1	RRRTAACAAAGAG	(115)
SOX-5	AACAAT	(11)
SRY	WWCAAAG	(51,52)
ste11	AACAAAGAA	(107)
TCF-1	WWCAAAG	(99,116)
SSRP:		
T160, heptamer site	CACAGTG	(48)
T160, nonamer site	ACAAAACC	(48)
UBF1-3-6:		
mtTF1	TTTTGACAT	(44)
mtsHMG	AGGTTTTTTACAT	(45)

^aThe IUPAC degenerate nucleotide base codes are: R (G or A); W (A or T); Y (C or T).

Each of the known recognition sites for proteins in the SRY subgroup have a central CA/TG dinucleotide pair (Table 2). It is also interesting to note that the binding sites containing this CA/TG pair are flanked by similar downstream sequences. In contrast, T160 (from the SSRP subgroup) exhibits sequence-specific binding to two recombination signal sequences having a central CA/TG pair but whose binding site and flanking regions are otherwise different from those seen for the SRY subgroup (Table 2). The binding sites for mtTF1 and mtsHMG (from the UBF1-3-6) also contain the central CA/TG pair but are otherwise different from those seen in the other subgroups. While the published SRY binding sites contain this central CA dinucleotide pair (51,52), it is important to note that the most detailed study on the interaction of SRY with its binding site shows that ⁹Ile interacts with an AA pair rather than with the CA pair in the binding site (53). This interaction involves a partial sidechain intercalation into the minor groove of the DNA. Further studies indicate that mutation of the AA pair has a greater effect on SRY binding than does mutation of the CA pair (54).

The proteins in the SRY subgroup are capable of bending free DNA; for example, gel electrophoretic experiments show that a 130° bend is induced by LEF-1 (55). Studies of the local geometry of B-DNA crystals containing either the CA/TG pair indicate that this DNA is susceptible to bending or deformation (56–59), and may thereby facilitate the binding of the sequence-specific HMG-1 box proteins. The interplay between sequence-specific binding to linear DNA and non-specific binding to bent DNA by HMG-1 box proteins has previously been addressed (13,29). It is possible that the alignment of HMG-1 proteins on the basis of structure allows for the identification of similar target DNA sequences after the classification of the protein sequences.

Despite containing most of the elements of the overall HMG-1 box signature (29), several proteins containing HMG-1 boxes did not fall into any of the clustering groups discussed above. These include the HMG protein from *Arabidopsis*, the HMG homolog from Chilo iridescent virus, and a number of yeast proteins: Ixr1, ARS-binding factor 2 (ABF2) and the non-histone proteins 6A

Figure 3. Signatures for each of the major subgroups in the HMG-1 superfamily. Signatures are based on the multiple sequence alignment in Figure 1 and indicate positions where very high similarity is observed. The signatures unambiguously identify each of the members found within each subgroup. For each signature, the most prevalent residue is shown in the first line. Residues shown in italic represent positions where gaps may be tolerated in the alignment. Positions corresponding to those identified in the global HMG-1 box signature (29) are shown in inverse type.

and 6B. At first, it appears unusual that the *Arabidopsis* HMG-1 box falls away from the other plant HMG-1 boxes located within the HMG1.2 group. However, inspection of the sequences indicates that the *Arabidopsis* HMG-1 box is only 30% identical to that from soybean. In contrast, the soybean HMG-1 box is 63% identical to that from wheat and 65% identical to that from maize; the maize and wheat HMG-1 boxes themselves are 90% identical.

The use of a structural alignment has produced a tree significantly different from those previously proposed for the HMG-1 superfamily (60,61). These studies relied upon a combination of both automated and manual methods to derive the alignment subsequently used in the clustering analysis. As such, gaps have been introduced into regions that are defined as α -helical in the solved NMR structure (8–10), gaps which may not necessarily be consistent with the maintenance of required contacts within the HMG-1 box. The structure-based threading alignment also remains consistent with the overall HMG-1 box signature (29), a feature which is not consistently seen in either of the alignments previously put forth. The differences in the alignment may explain the dissimilarity seen in the clustering of the six UBF boxes [only three are presented by Griess *et al.* (61)]. Similarities are seen in the clustering of the HMG-1 boxes 1 and 2 from different organisms, a result that would be expected regardless of alignment technique due to the high degree of sequence identity of these proteins. This study demonstrates how the structure-based alignment, which inherently contains more information than its sequence-based counterparts, produces a significantly different picture of the overall organization of the HMG-1 box family.

Sequence signatures for the HMG-1 box subgroups

Despite the identification of a common signature for the HMG-1 family (29), members of this family greatly differ in their degree of sequence similarity to one another. Based on the structural alignment of the HMG-1 box proteins, individual amino acid signatures were deduced for each of the major subgroups in the HMG-1 superfamily (Fig. 3). These amino acid signatures unambiguously identify each of the members found within each subgroup.

The individual signatures do retain the major features of the overall HMG-1 box signature (29). Hydrophobic residues predominate at positions 6–11, phenylalanine residues being particularly prevalent. The signatures indicate two almost invari-

ant positions: a proline at position 24 and a tryptophan at position 41. The highly conserved aromatic residues at the C-terminus are also observed in the individual signatures: phenylalanine, tryptophan or tyrosine residues are seen at positions 52 and 70, while phenylalanine, tryptophan or histidine residues are seen at position 63.

While these significant similarities exist across all five of the HMG-1 subgroup signatures, wide variation is observed in the nature of the individual signatures. The highly-conserved positions discussed above show differing degrees of degeneracy from subgroup to subgroup; for example, position 41 can be a tryptophan, phenylalanine or tyrosine in the UBF1-3-6 subgroup, but must be a tryptophan in the other four subgroups. The number of positions needed to unambiguously define each subgroup is very different: the signature for the SSRP subgroup is defined by residues at 62 positions, while that for the HMG1.1 subgroup is defined by residues at only 14 positions. The distribution of signature positions also varies substantially; most of the information contained in the signature for the HMG1.1 subgroup is at positions towards the C-terminus, while the signature positions in the HMG1.2 and SSRP subgroups are fairly evenly distributed across the entire length of the HMG-1 box. The differences between each of the subgroups are biologically important, as small variations in the otherwise conserved three-dimensional motif are ultimately critical to the maintenance of the separate cellular role of each individual HMG-1 box protein.

Evolutionary considerations

The threading experiments performed on the HMG-1 family using the NMR structure of the second HMG-1 box from rat HMG-1 (8) indicate that there are no rigid sequence requirements for the formation of the HMG-1 box motif (30), an observation that also becomes obvious upon examining the individual HMG-1 box subgroup signatures (Fig. 3). These observations further reinforce the concept that structure is conserved to a greater extent than sequence (62,63). The use of a structural alignment rather than one generated through traditional homology methods (60,61,64) will therefore provide more reliable information due to the evolutionary pressure to maintain the basic three-dimensional structure of the HMG-1 box.

The radial tree presented in Figure 2 is necessarily unrooted, since neither the identity or the relative position of the ancestral HMG-1 box is known. However, some speculations can be made

regarding the origin of the HMG-1 box. We believe that the second HMG-1 box (HMG1.2) is the most likely candidate for being the ancestral HMG-1 box. First, as mentioned above, plant species contain only one HMG-1 box, and these boxes (with the exception of the one from *Arabidopsis*) occupy a discrete branch within the HMG1.2 subgroup. A second consideration is the relative position of the HMG1.2 subgroup within the tree. As this subgroup occupies the most central position of the tree, it is the most closely-related in terms of sequence to the other subgroups. (It is important to make the distinction that the distances in the tree are not representative of time but rather strictly representative of the differences between sequences.) However, since the *Arabidopsis* HMG-1 box did not cluster with the HMG-1.2 group due to its lower extent of sequence similarity with other plant HMG-1 boxes, and since no yeast sequences are found within the HMG1.2 group, it is possible that there are alternative explanations. If the HMG-1 box 2 is indeed the ancestral HMG-1 box, it is then possible that the fourth UBF box (UBF1.4) is the HMG-1 box from which the remaining UBF boxes arose.

For the proteins that possess multiple HMG-1 boxes, there are two possible scenarios which could have led to the duplications (cf. 60,61). The first case is represented by a protein such as ABF2, whose two HMG-1 boxes are located on adjacent branches. Here, it is most likely that ABF2 arose from an ancestor with a single HMG-1 box gene and that the single gene was internally duplicated; the individual HMG-1 boxes are joined on the final protein. In the second case, represented by proteins such as mtTF1 and Ixr1, the individual HMG-1 boxes are located in entirely different parts of the tree. Here, a single, common HMG-1 box ancestor gave rise to the different HMG-1 box classes. In turn, the boxes from these two different classes subsequently joined through either a translocation or exon shuffling event, giving rise to the final protein product with multiple HMG-1 boxes. This second case is similar to that previously reported for selected nuclear receptors (65). Both of these scenarios allow for each of the HMG-1 boxes involved to evolve independently of one another.

Future perspectives

We have shown how homology model building can be applied to the problem of multiple sequence alignment in clustering studies, thereby emphasizing the structure of a conserved motif over its sequence alone. By emphasizing structural constraints, a more convincing and realistic portrait of the relationship between individual proteins or motifs can be deduced. The protocol is simple in principle and has great potential for its application in other cases where three-dimensional structural information is available. In the case of the high mobility group proteins and the HMG-1 box family in general, studies such as this will hopefully aid in the determination of the precise sites of interaction between HMG-1 box proteins and DNA, as well as in the characterization of the forces behind important processes such as DNA recognition and DNA deformation within this diverse class of DNA-binding proteins.

ACKNOWLEDGEMENTS

We wish to thank Drs Mark Boguski and Eugene Koonin for their critical review of this manuscript.

REFERENCES

- Goodwin, G.H., Johns, E.W. and Walker, J.M. (1977) *The Organization and Expression of the Eukaryotic Genome*. Academic Press, London.
- Johns, E.W. (1982) *The HMG Chromosomal Proteins*. Academic, New York.
- Bustin, M., Lehn, D.A. and Landsman, D. (1990) *Biochim Biophys Acta*, **1049**, 231–243.
- Reeck, G.R., Isackson, P.J. and Teller, D.C. (1982) *Nature*, **300**, 76–78.
- Cary, P.D., Turner, C.H., Mayes, E. and Crane-Robinson, C. (1983) *Eur. J. Biochem.*, **131**, 367–374.
- Cary, P.D., Turner, C.H., Leung, I., Mayes, E. and Crane-Robinson, C. (1984) *Eur. J. Biochem.*, **143**, 323–330.
- Jantzen, H.M., Admon, A., Bell, S.P. and Tjian, R. (1990) *Nature*, **344**, 830–836.
- Weir, H.M., Kraulis, P.J., Hill, C.S., Raine, A.R.C., Laue, E.D. and Thomas, J.O. (1993) *EMBO J.*, **12**, 1311–1319.
- Read, C.M., Cary, P.D., Crane-Robinson, C., Driscoll, P.C. and Norman, D.G. (1993) *Nucleic Acids Res.*, **21**, 3427–3436.
- Jones, D.N.M., Searles, M.A., Shaw, G.L., Churchill, M.E.A., Ner, S.S., Keeler, J., Travers, A.A. and Neuhaus, D. (1994) *Structure*, **2**, 609–627.
- Connor, F., Cary, P.D., Read, C.M., Preston, N.S., Driscoll, P.C., Denny, P., Crane-Robinson, C. and Ashworth, A. (1994) *Nucleic Acids Res.*, **22**, 3339–3346.
- Einck, L. and Bustin, M. (1985) *Exp. Cell. Res.*, **156**, 295–310.
- Bianchi, M.E., Beltrame, M. and Paonessa, G. (1989) *Science*, **243**, 1056–1059.
- Pil, P.M. and Lippard, S.J. (1992) *Science*, **256**, 234–237.
- Hughes, E.N., Engelsberg, B.N. and Billings, P.C. (1992) *J. Biol. Chem.*, **267**, 13520–13527.
- Lilley, D.M. (1992) *Nature*, **357**, 282–283.
- Landsman, D. and Bustin, M. (1991) *Mol. Cell. Biol.*, **11**, 4483–4489.
- Tremethick, D.J. and Molloy, P.L. (1988) *Nucleic Acids Res.*, **16**, 11107–11123.
- Waga, S., Mizuno, S. and Yoshida, M. (1990) *J. Biol. Chem.*, **265**, 19424–19428.
- Hisatake, K., Nishimura, T., Maeda, Y., Hanada, K.I., Song, C.Z. and Muramatsu, M. (1991) *Nucleic Acids Res.*, **19**, 4631–4637.
- O'Mahony, D.J. and Rothblum, L.I. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 3180–3184.
- Bachvarov, D., Normandeau, M. and Moss, T. (1991) *FEBS Lett.*, **288**, 55–59.
- McStay, B., Hu, C.H., Pikkard, C.S. and Reeder, R.H. (1991) *EMBO J.*, **10**, 2297–2303.
- Bachvarov, D. and Moss, T. (1991) *Nucleic Acids Res.*, **19**, 2331–2335.
- Denny, P., Swift, S., Brand, N., Dabhade, N., Barton, P. and Ashworth, A. (1992) *Nucleic Acids Res.*, **20**, 2887.
- Jäger, R.J., Anvret, M., Hall, K. and Scherer, G. (1990) *Nature*, **348**, 452–454.
- Diffley, J.F.X. and Stillman, B. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 7864–7868.
- Parisi, M.A. and Clayton, D.A. (1991) *Science*, **252**, 965–969.
- Landsman, D. and Bustin, M. (1993) *BioEssays*, **15**, 539–546.
- Baxevanis, A.D., Bryant, S.H. and Landsman, D. (1995) *Nucleic Acids Res.*, **23**, 1019–1029.
- Bairoch, A. and Boeckmann, B. (1993) *Nucleic Acids Res.*, **21**, 3093–3096.
- Barker, W.C., George, G.D., Mewes, H.-W., Pfeiffer, F. and Tsugita, A. (1993) *Nucleic Acids Res.*, **21**, 3089–3092.
- Rice, C.M., Fuchs, R., Higgins, D.G., Stoehr, P.J. and Cameron, G.N. (1993) *Nucleic Acids Res.*, **21**, 2967–2971.
- Benson, D., Lipman, D.J. and Ostell, J. (1993) *Nucleic Acids Res.*, **21**, 2963–2965.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
- Wen, L., Huang, J.K., Johnson, B.H. and Reeck, G.R. (1989) *Nucleic Acids Res.*, **17**, 1197–1214.
- Bryant, S.H. and Lawrence, C.E. (1993) *Proteins*, **16**, 92–113.
- Barton, G.J. (1993) *Protein Engng*, **6**, 37–40.
- Felsenstein, J. (1993) PHYLIP Phylogeny Inference Package 3.5. Department of Genetics, The University of Washington.
- Dayhoff, M.O. (1978) *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington.
- Fitch, W.M. and Margoliash, E. (1967) *Science*, **155**, 279–284.

- 42 Maciukenas, M. (1991) Treetool 1.0. Ribosomal RNA Database Project, The University of Illinois.
- 43 Johnson, M.S., Srinivasan, N., Sowdhamini, R. and Blundell, T.L. (1994) *Crit. Rev. Biochem. Mol. Biol.*, **29**, 1–68.
- 44 Fisher, R.P., Topper, J.N. and Clayton, D.A. (1987) *Cell*, **50**, 247–258.
- 45 Boissonneault, G. and Lau, Y.F.C. (1993) *Mol. Cell. Biol.*, **13**, 4323–4330.
- 46 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.
- 47 Hsu, T., King, D.L., LaBonne, C. and Kafatos, F.C. (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 6488–6492.
- 48 Shirakata, M., Hüppi, K., Usuda, S., Okazaki, K., Yoshida, K. and Sakano, H. (1991) *Mol. Cell. Biol.*, **11**, 4528–4536.
- 49 van de Wetering, M. and Clevers, H. (1992) *EMBO J.*, **11**, 3039–3044.
- 50 Read, C.M., Cary, P.D., Preston, N.S., Lnenicek-Allen, M. and Crane-Robinson, C. (1994) *EMBO J.*, **13**, 5639–5646.
- 51 Harley, V.R., Jackson, D.I., Hextall, P.J., Hawkins, J.R., Berkovitz, G.D., Sockanathan, S., Lovell-Badge, R. and Goodfellow, P.N. (1991) *Science*, **255**, 453–456.
- 52 Nasrin, N., Buggs, C., Kong, X.F., Carnazza, J., Goebel, M. and Alexander-Bridges, M. (1991) *Nature*, **354**, 317–320.
- 53 King, C.-Y. and Weiss, M.A. (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 11990–11994.
- 54 Haqq, C.M., King, C.Y., Ukiyama, E., Falsafi, S., Haqq, T.N., Donahoe, P.K. and Weiss, M.A. (1991) *Science*, **266**, 1494–1500.
- 55 Giese, K. and Grosschedl, R. (1993) *EMBO J.*, **12**, 4667–4676.
- 56 Bolshoy, A., McNamara, P., Harrington, R.E. and Trifonov, E.N. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 2312–2316.
- 57 Travers, A.A. (1991) *Curr. Opin. Struct. Biol.*, **1**, 114–122.
- 58 Yanagi, K., Privé, G.G. and Dickerson, R.E. (1991) *J. Mol. Biol.*, **217**, 201–214.
- 59 Heinemann, U. and Hahn, M. (1992) *J. Biol. Chem.*, **267**, 7332–7341.
- 60 Laudet, V., Stehelin, D. and Clevers, H. (1993) *Nucleic Acids Res.*, **21**, 2493–2501.
- 61 Griess, E.A., Rensing, S.A., Grasser, K.D., Maier, U.-G. and Feix, G. (1993) *J. Mol. Evol.*, **37**, 204–210.
- 62 Chothia, C. and Lesk, A.M. (1986) *EMBO J.*, **5**, 823–826.
- 63 Chothia, C. (1992) *Nature*, **357**, 543–544.
- 64 Wu, M., Allis, C.D., Sweet, M.T., Cook, R.G., Thatcher, T.H. and Gorovsky, M. (1994) *Mol. Cell. Biol.*, **14**, 10–20.
- 65 Laudet, V., Hänni, C., Col, J., Catzeffis, F. and Stéhelin, D. (1992) *EMBO J.*, **11**, 1003–1013.
- 66 Picardo, A.M., Mueller, U., Harry, J.L., Uwanogho, D. and Sharpe, P.T. (1992) *PCR Methods Appl.*, **2**, 218–222.
- 67 Merenmies, J., Pihlaskari, R., Laitinen, J., Wartiovaara, J. and Rauvala, H. (1991) *J. Biol. Chem.*, **266**, 16722–16729.
- 68 Yamaguchi-Shinozaki, K. and Shinozaki, K. (1992) *Nucleic Acids Res.*, **20**, 6737.
- 69 Dalrymple, B.P. and Peters, J.M. (1992) *Biochem. Biophys. Res. Commun.*, **184**, 31–35.
- 70 Kaplan, D.J. and Duncan, C.H. (1988) *Nucleic Acids Res.*, **16**, 10375.
- 71 Pentecost, B. and Dixon, G.H. (1984) *Biosci. Rep.*, **4**, 49–57.
- 72 Wang, L., Precht, P., Balakir, R. and Horton, W.E. (1993) *Nucleic Acids Res.*, **21**, 1493.
- 73 Schnitzler, P., Hug, M., Handermann, M., Janssen, W., Koonin, E.V., Delius, H. and Darai, G. (1993) *Nucleic Acids Res.*, **22**, 158–166.
- 74 Funahashi, J., Sekido, R., Murai, K., Kamachi, Y. and Kondoh, H. (1993) *Development*, **119**, 433–446.
- 75 Davis, D.L. and Burch, J.B.E. (1992) *Gene*, **113**, 251–256.
- 76 Ota, T., Endo, Y., Ito, M., Miyamoto, K.I., Sasakawa, T., Suzuki, I. and Natori, Y. (1992) *Biochim. Biophys. Acta*, **1130**, 224–226.
- 77 Gastrop, J., Hoevenagel, R., Young, J.R. and Clevers, H.C. (1992) *Eur. J. Immunol.*, **22**, 1327–1330.
- 78 Wisniewski, J.R. and Schulze, E. (1992) *J. Biol. Chem.*, **267**, 17170–17177.
- 79 Lehming, N., Thanos, D., Brickman, J.M., Ma, J., Maniatis, T. and Ptashne, M. (1994) *Nature*, **371**, 175–179.
- 80 Debuchy, R. and Coppin, E. (1992) *Mol. Gen. Genet.*, **233**, 113–121.
- 81 Lee, K.-L.D., Pentecost, B.T., D'Anna, J.A., Tobey, R.A., Gurley, L.R. and Dixon, G.H. (1987) *Nucleic Acids Res.*, **15**, 5051–5068.
- 82 Stros, M. and Dixon, G.H. (1993) *Biochim. Biophys. Acta*, **1172**, 231–235.
- 83 Majumdar, A., Brown, D., Kerby, S., Rudzinski, I., Polte, T., Randawa, Z. and Seidman, M.M. (1991) *Nucleic Acids Res.*, **19**, 6643.
- 84 Alexandre, S., Li, W.W. and Lee, A.S. (1992) *Nucleic Acids Res.*, **20**, 6413.
- 85 Schulman, I.G., Wang, T., Wu, M., Bowen, J.K., Cook, R.G., Gorovsky, M.A. and Allis, C.D. (1991) *Mol. Cell. Biol.*, **11**, 166–174.
- 86 Wagner, C.R., Hamana, K. and Elgin, S.C. (1992) *Mol. Cell. Biol.*, **12**, 1915–1923.
- 87 Ner, S.S., Churchill, M.E.A., Searles, M.A. and Travers, A.A. (1993) *Nucleic Acids Res.*, **21**, 4369–4371.
- 88 Gozé, C., Poulat, F. and Berta, P. (1993) *Nucleic Acids Res.*, **21**, 2943.
- 89 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.
- 90 Brown, S.J., Kellett, P.J. and Lippard, S.J. (1993) *Science*, **261**, 603–605.
- 91 Travis, A., Amsterdam, A., Belanger, C. and Grosschedl, R. (1991) *Genes Dev.*, **5**, 880–894.
- 92 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. (1992) *Nature*, **359**, 531–533.
- 93 Kelly, M., Burke, J., Smith, M., Klar, A. and Beach, D. (1988) *EMBO J.*, **7**, 1537–1547.
- 94 Turgeon, B.G., Bohlmann, H., Ciuffetti, L.M., Christiansen, S.K., Yang, G., Schaefer, W. and Yoder, O.C. (1993) *Mol. Gen. Genet.*, **238**, 270–284.
- 95 Pauken, C.M., Nagle, D.L., Bucan, M. and Lo, C.W. (1994) *Mammalian Genome*, **5**, 91–99.
- 96 Stolzenburg, F., Dinkl, E. and Grummt, F. (1992) *Nucleic Acids Res.*, **20**, 4927.
- 97 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.
- 98 Staben, C. and Yanofsky, C. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 4917–4921.
- 99 Oosterwegel, M., van de Wetering, M., Dooijes, D., Klomp, L., Winoto, A., Georgopoulos, K., Meijlink, F. and Clevers, H. (1991) *J. Exp. Med.*, **173**, 1133–1142.
- 100 Grasser, K.D. and Feix, G. (1991) *Nucleic Acids Res.*, **19**, 2573–2577.
- 101 Kolodrubetz, D. and Burgum, A. (1990) *J. Biol. Chem.*, **265**, 3234–3239.
- 102 Chan, E.K.L., Imai, H., Hamel, J.C. and Tan, E.M. (1991) *J. Exp. Med.*, **174**, 1239–1244.
- 103 Tsuda, K.-I., Kikuchi, M., Mori, K., Waga, S. and Yoshida, M. (1988) *Biochemistry*, **27**, 6159–6163.
- 104 Shirakawa, H., Tsuda, K.-I. and Yoshida, M. (1990) *Biochemistry*, **29**, 4419–4423.
- 105 Paonessa, G., Frank, R. and Cortese, R. (1987) *Nucleic Acids Res.*, **15**, 9077.
- 106 Laux, T. and Goldberg, R.B. (1991) *Nucleic Acids Res.*, **19**, 4769.
- 107 Sugimoto, A., Iino, Y., Maeda, T., Watanabe, Y. and Yamamoto, M. (1991) *Genes Dev.*, **5**, 1990–1999.
- 108 Pentecost, B.T., Wright, J.M. and Dixon, G.H. (1985) *Nucleic Acids Res.*, **13**, 4871–4888.
- 109 Hayashi, T., Hayashi, H. and Iwai, K. (1989) *J. Biochem.*, **105**, 577–581.
- 110 Erond, N.E. and Donelson, J.E. (1992) *Mol. Biochem. Parasitol.*, **51**, 111–118.
- 111 Grasser, K.D., Wohlfarth, T., Baumlein, H. and Feix, G. (1993) *Plant Mol. Biol.*, **23**, 619–625.
- 112 Giese, K., Amsterdam, A. and Grosschedl, R. (1991) *Genes Dev.*, **5**, 2567–2578.
- 113 Dooijes, D., van de Wetering, M., Knippels, L. and Clevers, H. (1993) *J. Biol. Chem.*, **268**, 24813–24817.
- 114 Philley, M.L. and Staben, C. (1994) *Genetics*, **137**, 715–722.
- 115 Lowry, C.V., Cerdan, M.E. and Zitomer, R.S. (1990) *Mol. Cell. Biol.*, **10**, 5921–5926.
- 116 Oosterwegel, M., van de Wetering, M., Holstege, F., Prosser, H.M., Owen, M.J. and Clevers, H. (1991) *Int. Immunol.*, **3**, 1189–1192.