Signature for the HMG-1 box DNA-binding proteins

ARTICLE *in* BIOESSAYS · AUGUST 1993

Impact Factor: 4.73 · DOI: 10.1002/bies.950150807 · Source: PubMed

2 AUTHORS:

David Landsman
National Institutes of Health
142 PUBLICATIONS **9,689** CITATIONS
SEE PROFILE

Michael Bustin
National Institutes of Health
217 PUBLICATIONS **9,418** CITATIONS
SEE PROFILE

A Signature for the HMG-1 Box DNA-Binding Proteins

David Landsman and Michael Bustin

Summary

A diverse group of DNA-binding regulatory proteins share a common structural domain which is homologous to the sequence of a highly conserved and abundant chromosomal protein, HMG-1. Proteins containing this HMG-1 box regulate various cellular functions involving DNA binding, suggesting that the target DNA sequences share a common structural element. Members of this protein family exhibit a dual DNA-binding specificity: each recognizes a unique sequence as well as a common DNA conformation. The highly conserved HMG-1/-2 proteins may modulate the binding of other HMG-1 box proteins to bent DNA. We examine the structural and functional relationships between the proteins, identify their signature† and describe common features of their target DNA elements.

Introduction

Many DNA-binding proteins can be grouped into families which share common motifs such as zinc fingers, cysteine-rich zinc cluster domains, leucine zippers, helix-turn-helix domains, helix-loop-helix domains, homeodomains, β-sheet DNA-binding motifs and cold shock domains (for recent reviews see[2,3]). Although these proteins are found predominantly in the nucleus associated with chromatin, in some cases they have also been found in the cytoplasm. Recent studies identified a new group of DNA-binding proteins which have structural elements distinctive from those mentioned above and are related to the **h**igh **m**obility **g**roup (HMG) non-histone chromosomal proteins. The HMG proteins recognize unique DNA structures and may be involved in diverse functions, including determination of nucleosome structure and stability, as well as in transcription and/or replication[4,5,6].

The HMG proteins were first characterized by Johns and Goodwin as chromatin components with a high electrophoretic mobility in polyacrylamide gels[4]. Higher eukaryotes contain three families of HMG proteins, the HMG-14/-17 family, the HMG-1/-2 family and the HMG-I/-Y family. Each of these families of proteins contain distinct sequence motifs[6] and therefore are easily distinguishable by sequence analytical methods. Although the families are different in size and DNA-binding properties, HMG proteins are similar in their physical properties. These ubiquitous and abundant proteins are extractable from chromatin in 0.35 M NaCl and are soluble in 5% perchloric or trichloroacetic acid (for reviews see refs 4, 5 and 6).

The renewed interest in one of these families, the HMG-1/-2 family, stems from the observation that a unique RNA polymerase I transcription factor, known as UBF (**u**pstream **b**inding **f**actor), contains several regions which are similar to HMG-1[7]. Subsequent studies revealed the existence of a large family of DNA-binding proteins related to the HMG-1/-2 proteins. This family is highly diverse and includes the transcription factor UBF[7], the mammalian testis-determining factor, SRY[8,9], the lymphoid enhancer binding factor, LEF-1/TCF1α[10,11,12], the mitochondrial transcription factors, ABF2 and mtTF1[13,14], a component of the V-(D)-J recombinase, T160[15], a structure-specific recognition protein that binds cisplatin-modified DNA, SSRP1[16], and several other proteins involved in mating type determinations and sexual development (see Table 1). The only apparent common structural attribute among these proteins is the HMG-1 domain.

The Scope of the Review

In this review we briefly summarize the information available on the structure and function of the canonical HMG-1 and HMG-2 chromosomal proteins. However, the emphasis of this review is to analyze the sequences of the proteins containing the HMG-1 box and to identify common motifs characteristic of this class of proteins. In addition, we summarize current hypotheses concerning their mode of DNA binding. Three-dimensional aspects of the structure of the HMG-1 box proteins are not included since results of several NMR structural determinations will soon be published. We regret that the limit on the number of references precludes us from giving appropriate credit to all the contributors in the field.

Structure and Function of HMG-1/-2 Proteins

HMG-1 and its homolog, HMG-2, are closely related proteins which show intraspecies sequence identities of approximately 80%[6]. They have a tripartite structure[17], consisting of two internal repeats of a charged domain of about 80 residues and an acidic C-terminal domain, containing a stretch of approximately 30 continuous glutamic or aspartic acid residues. Each of the positively-charged internal repeats has been called an ***HMG box***[7]. This term should be redefined as the more specific ***HMG-1 box*** to distinguish this family from the other HMG proteins which also have highly conserved DNA-binding motifs.

Secondary structure predictions indicate that each HMG-1 box domain contains at least two regions of α-helical structure preceded by a short N-terminal, β-sheet region[17,18]. However, unpublished structural data indicates that there are three α-helical regions in this domain. Each HMG-1 domain contains several internal hydrophobic residues. The C-terminal domain has a high negative charge and lacks hydrophobic

†A signature is an amino acid pattern which unambiguously identifies members of related protein families[1].

Table 1. *Properties of HMG-1 box containing proteins*

Protein	Species*	DNA conformation recognized†	Number of HMG-1 boxes	Footprint size or binding site	Function
HMG-1	Human, rat, trout, hamster, pig, calf	1×, 2×, crux	2	14±3 bp	?
HMG-2	Pig, human, chicken, mouse	1×, 2×	2	n.d.	?
UBF (also TFIS and NOR-90)	Human, *Xenopus laevis*, rat, mouse	2×	4-6	>30 bp	Transcription
HMG-B	*Tetrahymena thermophila* and *pyriformis*	n.d.	1	n.d.	Macronuclear?
HMG-C	*Tetrahymena thermophila*	n.d.	1	n.d.	Macronuclear?
delta	*Tetrahymena thermophila*	n.d.	1	n.d.	Micronuclear?
mat1-Mc	*Schizosaccharomyces pombe*	2×	1	n.d.	Mating type/meiosis
mt a-1	*Neurospora crassa*	2×	1	n.d.	Mating type
FPR1	*Podospora anserina*	n.d.	1	n.d.	Sexual cell fertilization
NHP6A/NHP6B	*Saccharomyces cerevisiae*	n.d.	1	n.d.	?
SRY	Human, mouse, +	2×, bent	1	CCATTGTTCT	Sex determination
TCF-1α	Human	2×	1	[CT]CCTTTG[AT]A	Transcription in T-cells
LEF-1	Mouse	2×, bent	1	~15 bp	Transcription in T- and pre-B-cells
mtTF1	human	2×, bent	2	~30 bp	Mitochondrial transcription
ABF2	*Saccharomyces cerevisiae*	2×	2	25-30 bp	Mitochondrial replication
Maize HMG	Maize	n.d.	1	n.d.	?
T160 V-(D)-J recombinase	Mouse	2×	1	CACAGTG..ACAAAAACC	Recombination signal sequence binding
SSRP1 (almost identical to T160)	human	2×, bent	1	cisplatin-modified DNA	Binds cisplatin-modified DNA
SB11	Soybean embryo	n.d.	1	n.d.	?
TDP-1	*Trypanosoma brucei*	n.d.	2	ATTTGCAT	?
ROX1	*Saccharomyces cerevisiae*	2×	1	YYATTGTTCTC	Transcription repression
ste11	*Schizosaccharomyces pombe*	2×, bent (?)	1	TTCTTTGTTY	Sexual development inducer
IRE-ABP	Rat	2×	1	CCTTTGA	Insulin response element binding
HMG-D	*Drosophila melanogaster*	n.d.	1	n.d.	?
SpHMG-1‡	*Strongylocentrotus purpuratus*	n.d.	2	n.d.	?

HMG-1 box containing proteins, identified as described in Fig. 2, are listed. The columns describe, from the left, the protein names, the species from which the cDNA or protein was isolated, the type of DNA to which they bind, the number of HMG-1 boxes in each protein, the size or sequence of the DNA to which each protein binds and the function of the protein, if known.

*References are not included.

†1× = single stranded, 2× = double stranded, crux = cruciform, n.d. = not determined.

‡Unpublished data from C. C. Niemeyer and C. Flytzanis.

+ = several other species.

residues, therefore it is not expected to form a compact globular domain. This acidic domain could be involved in stabilizing the protein structure by interacting with one of the positively charged HMG-1 box domains. Indeed, the structure of HMG-1/-2 proteins is stable at various ionic strengths and pH conditions[4].

In chromatin, HMG-1/-2 proteins bind to the linker DNA between nucleosomes (reviewed in refs 4 and 5). Early studies suggested that these proteins preferentially bind single-stranded DNA[4,5]. Further investigations suggested that the proteins recognize, and bind with different affinities, to a variety of non-B-DNA structures such as palindromes, B-Z DNA junctions, cruciforms, four-way junctions, and *cis*-diamminedichloroplatinum(II) (cisplatin)-modified DNA[19,20,21,22]. The DNA-binding site of HMG-1 covers about 14 base pairs or bases for double- or single-stranded DNA, respectively[23].

The cellular function of HMG-1/-2 proteins is not yet clearly understood. It has been suggested that they may be involved in transcription, replication, chromatin assembly and stabilization of chromatin structure (reviewed in refs 4, 5 and 6). These abundant proteins may bind to DNA junctions, thereby stabilizing loop structures in chromatin[18,24]. The possibility that these proteins have a role in transcription was investigated by several groups. HMG-1/-2 can stimulate the formation of transcription initiation complexes of RNA polymerase II and III[25,26,27], probably by facilitating the binding of transcription factors to template DNA. However, in a direct test, HMG-1 was not capable of activating transcription[28]. Recent studies[29] suggest that the canonical HMG-1 may modulate transcriptional processes by influencing the DNA binding of other HMG-1 box proteins (as discussed below). The role of HMG-1 in replication and chromatin assembly has not been studied in detail. The current status of these functions has been previously reviewed[4,5,6]. The data available does not lead to a clear conclusion.

Functional Diversity of HMG-1 Box Containing Proteins

HMG-1 box proteins have been implicated in a wide variety of cellular functions (see Table 1). The RNA polymerase I transcription factor, UBF, is required for efficient transcription of vertebrate ribosomal genes[7]. UBF is a component of the transcription initiation complex which includes the TATA-binding protein (TBP) and the TBP-associated factors (TAFs)[30]. UBF may be counteracting histone H1-mediated repression of RNA polymerase I transcription[31]. The first HMG-1 box and the acidic tail are required for nucleolar localization of UBF[32]. LEF-1 or TCF1 are members of a family of T-lymphocyte enhancer-binding factors which have been shown to be involved in the regulation of the T-cell receptor enhancer a [10,11,12]. Although these proteins seem to be required for full enhancer activity, their function is contingent on the presence of additional proteins[10,33]. IRE-ABP, a protein isolated from rat adipocytes, is involved in the control of tissue specificity of the insulin response by regulating the transcription of several genes including aldolase, glucokinase and amylase in lipogenic tissues[34]. ROX1, isolated from *S. cerevisiae*, is involved in the repression of several genes, including those of transcriptional activators[35]. mtTF1 stimulates transcription in human mitochondria[14]. ABF2, isolated from *S. cerevisiae* mitochondria, is the homolog of mtTF1. This protein seems to be involved in replication and higher order organization of mitochondrial DNA[36,37]. T160 is part of a protein complex which modulates the functions of the V-(D)-J recombinase in mice[15]. The function of T160 in this complex is to recognize recombination signal sequences which are proposed to form stem-loop structures[15]. SSRP1, a protein with close similarity to T160 (95.5% identity), was isolated from human B-cells by screening a cDNA expression library with cisplatin-modified DNA[16]. Interestingly, HMG-1 and HMG-2 also recognize cisplatin-modified DNA with high specificity[21,22].

Several HMG-1 box containing proteins are involved in sexual differentiation. The ste11 protein, which is induced by nitrogen starvation, regulates the expression of several mating type genes involved in sexual development in *S. pombe*[38]. The mat1-Mc protein is necessary for mating in *S. pombe* and is involved in meiosis[39]. mat a-1 protein[40] and the FPR1 gene product[41] have been implicated in mating type selection in *N. crassa* and in the related fungus, *P. anserina*, respectively. In higher eukaryotes, the SRY gene product has been shown to be the male sex-determining factor in several mammals[42]. Sox-5, an SRY-related protein isolated from mice, is expressed only in testis, and is most abundant in round spermatids which are early pre-meiotic haploid cells[43]. In view of the above findings, it is tempting to speculate that HMG-1 box proteins are involved in sex determination and mating-type selection in a wide variety of eukaryotic species. Conceivably, both developmental processes may involve similar pathways and use related protein components.

The number of HMG-1 boxes per protein varies (Table 1). Most proteins have a single HMG-1 box. The canonical HMG-1 and HMG-2 proteins, the two mitochondrial transcription factors, mtTF1 and ABF2, and the trypanosome protein TDP-1, each contain two of these domains. The number of HMG-1 boxes in UBF varies from 4 to 6 among the various species examined so far[44]. Conceivably, the variation in HMG-1 boxes arose from gene duplications of a single copy of an archetypal HMG-1 box.

Table 1 also indicates that the size of the DNA site to which the HMG-1 box binds, varies from 9 to 30 bp. This wide spread could be misleading as footprints might only be accurately assessed when all the components involved in a protein-DNA complex are included. In the case of UBF, only the first HMG-1 box is sequence-specific while the others modulate binding efficiency[30]. This protein has an unusually large binding site, suggesting that some promoter and enhancer DNA sequences are wrapped around UBF[45].

Thus, HMG-1 box containing proteins seem to be involved in functions related to transcription regulation, mating-type selection and sexual determination. The HMG-1 box is the only common feature in all these proteins. Members of this family use the HMG-1 domain to bind specifically to DNA.

HMG-1 Box Proteins Bind to Bendable DNA

The canonical HMG-1/-2 proteins preferentially bind to non-B-DNA conformations such as B-Z junctions, stem-loops, cruciforms, four-way junctions, and cisplatin-modified DNA[18,19,20,21] without any sequence specificity, suggesting that the proteins recognize specific DNA conformations. Indeed, it has been shown that HMG-1 binds to bent DNA. A 34° bend towards the major groove has been shown to be an optimal angle for HMG-1 binding[21].

Most HMG-1 box containing proteins such as LEF-1, SRY, mtTF1 and ABF2 bind, and recognize, specific sequences as well as non-B-DNA structures[36,37,42,46]. The binding of several HMG-1 box proteins to DNA has been shown to be ***sequence specific***, independent of any known requirements for a specific ***DNA conformation***[34,47]. On the other hand, T160, the murine homolog of SSRP1, binds stem-loop structures in recombination signal sequences[15]. The HMG-1 box domain proteins TCF-1, LEF-1 and SRY contact the bases in the recognition sequence on both strands in the minor groove with no major groove contacts[46,48]. The interaction of LEF-1, SRY, ABF2 and mtTF1, induces bends in the bound DNA, similar to that proposed in *E. coli* for the interaction of the integration host factor (IHF) with DNA[46], and for the wrapping of promoter and enhancer DNA around UBF (also called TFIS)[45].

The HMG-1 boxes in the HMG-1 box containing proteins are independent functional domains, are not interchangeable, and each box specifically recognizes its target[46]. Conceivably, these proteins recognize a common DNA structural motif embedded in a distinct sequence. Most probably, the binding to the structural element is not sufficiently strong to ensure a stable protein-DNA complex. Analysis of all known HMG-1 box recognition sites indicates that they contain central TG (or CA) residues (see Table 1). A study of the local helix geometry of several B-DNA crystals containing TG

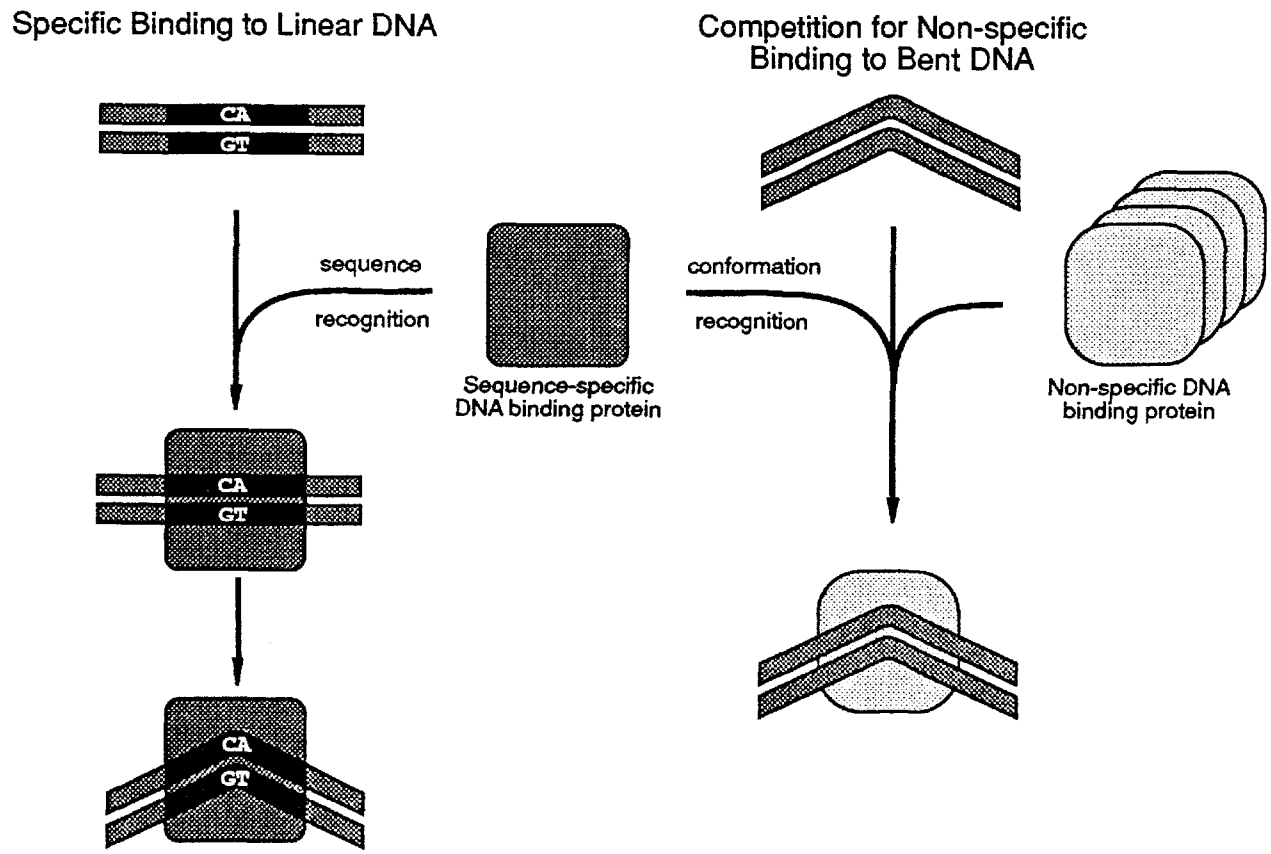

Fig. 1. Sequence and conformation specific interactions for HMG-1 box proteins binding to DNA. Most HMG-1 box proteins bind linear DNA containing a central CA or TG dinucleotide, in a sequence-specific fashion. Binding of these proteins induces a conformational change in their target DNA. The same proteins bind to bent DNA without sequence-specificity. HMG-1/-2 recognize bent DNA. The cellular concentration of the latter proteins is higher than that of the sequence-specific HMG-1 box proteins. Therefore, the canonical HMG-1/-2 proteins will inhibit the binding of HMG-1 box proteins to bent DNA. This model is based on that of Bianchi and colleagues[29].

and CA dinucleotides[49,50] indicated that these dinucleotides play a role in bending and/or deforming DNA[51,52]. Thus, the common target recognized by the HMG-1 boxes may be a DNA deformation caused by the TG or CA dinucleotides and, possibly, the bases immediately adjacent to these dinucleotides[49,50]. A relationship between binding affinity and bending angle has been observed for several HMG-1 box proteins[46].

Once bound, the sequence-specific HMG-1 box proteins induce a sharp bend in the DNA[29,46]. Bianchi and collaborators have demonstrated[29] that the SRY protein, which is a sequence-specific DNA-binding protein, binds to four-way junction DNA without any sequence specificity, in a similar fashion to the HMG-1 binding mentioned above. In fact, the binding of the sequence-specific SRY to four-way junction DNA is indistinguishable from the binding of HMG-1, which binds non-specifically to DNA. Thus, another common feature in all HMG-1 box proteins may be their ability to bind DNA bends or two separate DNA strands lying across and close to each other in space, as in the case of four-way junctions[24].

These observations may be relevant to the understanding of the cellular function of HMG-1/-2. As suggested in the model presented by Bianchi and colleagues[29], HMG-1/-2, a non-specific DNA-binding protein, binds to bent DNA, to form a stable protein-DNA complex. The sequence-specific HMG-1 box proteins can bind and induce bends in linear DNA which contains the central dinucleotides CA or TG. In addition, these HMG-1 box containing proteins bind already bent DNA without any sequence specificity[29]. Therefore, these proteins bind non-specifically to many sites in chromatin. This non-specific binding may decrease the free concentration of the HMG-1 box proteins below the threshold needed to bind their sequence-specific target. HMG-1/-2 could compete and inhibit the non-specific binding of these sequence-specific proteins and effectively, decrease the concentration of non-specific binding sites. Thus, HMG-1/-2 ensures that these sequence-specific proteins find their target. A model illustrating this concept is presented in Fig. 1.

Cellular Localization of HMG-1 Box Proteins

Chromosomal proteins HMG-1/-2 and HMG-1 box containing proteins are found in several cellular compartments. Traditionally, HMG-1/-2 are prepared from washed nuclei or chromatin[4]; however, immunofluorescence and biochemical fractionation studies have shown that HMG-1/-2 are

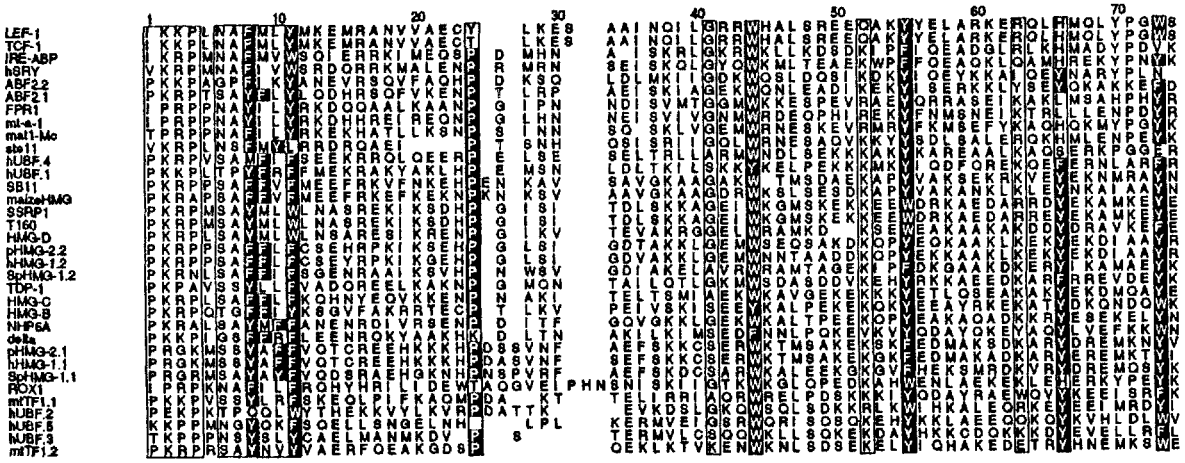

Fig. 2. Multiple alignment of HMG-1 box domains from several proteins. Sequences were extracted from the SWISS-PROT database (Version 23.0), Protein Identification Resource (PIR) (Version 34.0), and GenPept, the translated GenBank database (Version 73.1), and aligned by several methods. The search cutoffs used to identify homologs of HMG-1 are a Karlin/Altschul score for two aligned sequence segments >70 with a probability of $<10^{-3}$ using the BLAST software. The multiple alignment formatting tool, ALSCRIPT, was used to print the alignment[66]. By these criteria, all the domains presented, here and in Table I, are homologs of a single HMG-1 motif. Boxed regions show regions of higher similarity. Shaded blocks at single sites identify the regularly spaced, aromatic residues. Abbreviated protein names are cross referenced to Table 1. The protein names hSRY, hHMG-1 and hUBF are the human and pHMG-2 the porcine, homolog of these proteins. The decimal points are used to identify distinct HMG-1 box domains in the same protein.

present in both the nucleus and the cytoplasm of Chinese hamster lung, rat liver and bovine trachea cells[5]. HMG-1/-2 migrate between the nucleus and cytoplasm in a cell cycle-dependent fashion[4]. The cellular levels of these proteins vary among tissues. Their levels are increased in the nuclei and cytoplasm of lymphoid and testis tissues, whereas in hepatic and brain cells the proteins accumulate in the cytosol[53]. Several studies suggest that the cellular levels of these proteins change during cell differentiation[54]. The possible significance of HMG-1/-2 in both the nucleus and the cytoplasm, and their shuttling between these cellular compartments, has not been studied in detail. Conceivably, the cytoplasm might act as storage for these proteins before they are transferred to the nucleus for a specific and temporal function. It is also possible that HMG-1/-2 may play some role in the transport of cellular components across the nuclear membrane.

Protein p30 (amphoterin), a heparin-binding protein with adhesive and neurite outgrowth-promoting properties, is associated with membranes in the cytoplasm and filopodia of neuroblastoma cells[55]. Neurite outgrowth in neuroblastoma cells is inhibited by anti-amphoterin antibodies. Interestingly, the sequence of p30 is identical to rat HMG-1 and the two proteins cross-react immunologically[55,56]. This finding is puzzling since HMG-1/-2 have not been detected as membrane-associated proteins in other tissues.

ABF2, from *S. cerevisiae*, and mtTF1, from *H. sapiens*, are located predominantly in mitochondria, although significant fractions are also located in the nucleus[13,14]. UBF, which is involved in the transcription regulation of RNA polymerase I-transcribed genes, is located predominantly in the nucleolus[7]. The cellular localization of other HMG-1 box proteins which are involved in transcriptional regulation, has not been studied in detail. Thus, the presence of an HMG-1 box does not affect the cellular localization of these proteins.

Sequence Similarities Between HMG-1 Box Proteins

To obtain a set of HMG-1 box containing proteins, we used the BLAST series of programs[57] to search a non-redundant protein sequence database, comprising a merge of the SWISS-PROT (version 23.0), PIR (Version 34.0) and GenPept (Version 73.1) databases, with the human HMG-1 protein sequence[58]. At present (November, 1992), these databases contain over 90 related sequences which can be identified as true members of this family of proteins. About half of these proteins are homologs of HMG-1/-2 or SRY. The most distantly related family members show at least 29% identity over any stretch of 46 contiguous amino acids. A comprehensive multiple alignment of the HMG-1 box domains is shown in Fig. 2. The alignment contains only one representative of each of the known HMG-1 box motifs (e.g. HMG-1, HMG-2 and SRY in Table 1).

Regions of similarity in the alignment are highlighted in Fig. 2. The N-terminal region is characterized by two proline residues separated by two basic amino acids (positions 1-4). A hydrophobic cluster of five amino acids is located at positions 6-11. Within this cluster, positions 8 and 11 are invariably aromatic amino acids. Residue 24 is almost invariant in proline and may delineate a new region of secondary structure. The region following this proline is highly variable. At position 41, there is a glycine in at least 70% of the sequences. This amino acid, which is located at the center of the motif, could potentially be the beginning of a second, α-helical sub-domain[18]. Other highly conserved positions are

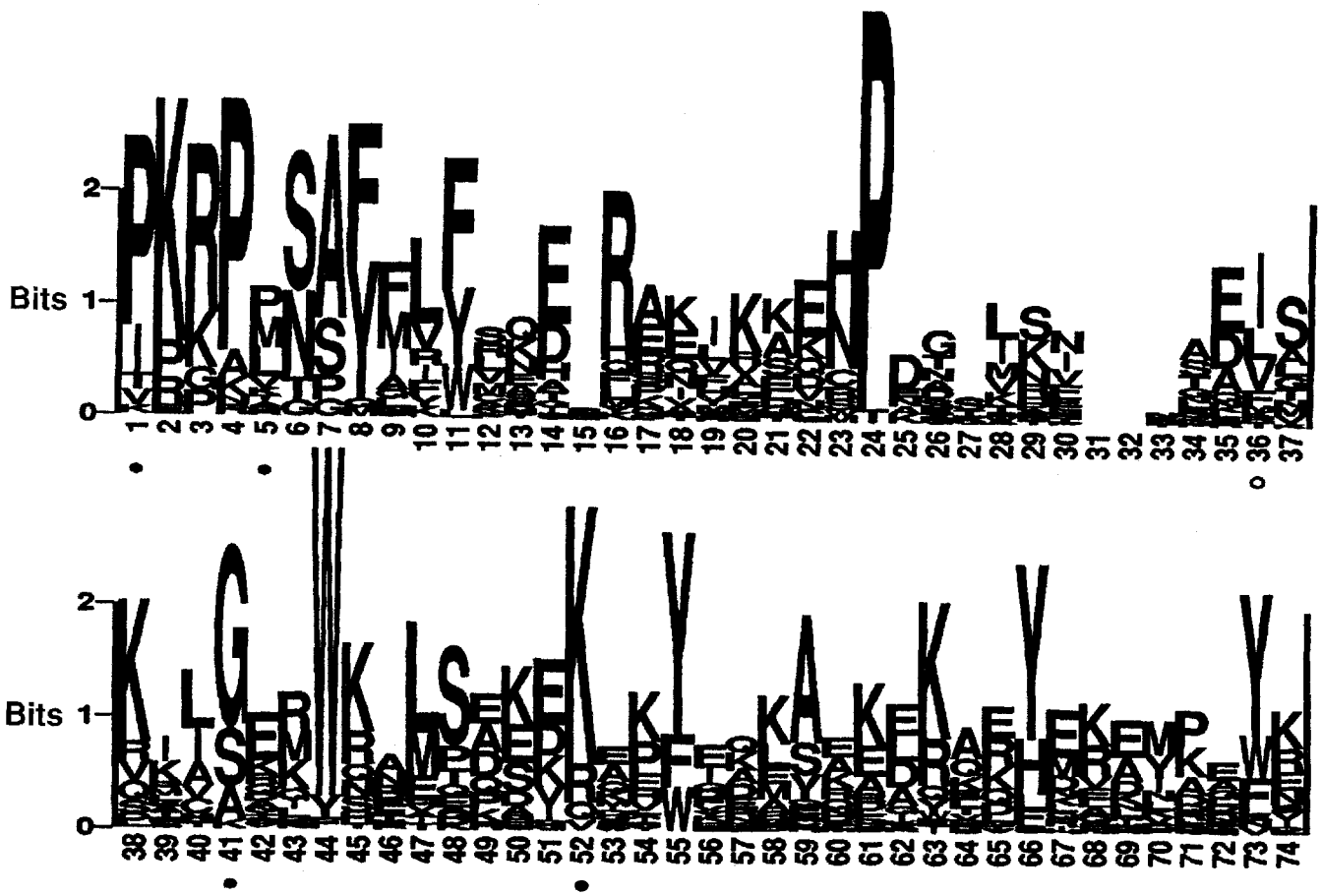

Fig. 3. Sequence logo of multiple alignment of HMG-1 box domains. The sequence logo is derived from the alignment in Fig. 2. The information content, in bits, is determined at each position[63]. The sizes of the letters are proportional to the information content (in bits) for that amino acid, which is a graphical representation of the frequency of an amino acid at a given position. Thus, taller letters represent a high information content (e.g. position 24) as do taller piles of letters (e.g. positions 1-4). The five circles are the sites at which SRY mutations have been identified in sex-reversal individuals. These mutations alter the binding of the SRY HMG-1 box domain to the DNA[42,64,65]. The numbers indicate the positions of the residues, as in Fig. 2.

44, 52, 55, 63, 66 and 73. Of these, the residues at positions 44, 55, 66 and 73 are predominantly the aromatic amino acids W, Y and F. Analysis of the alignment defines a distinct motif in HMG-1 box domains with the following signature:

[GSA][YF].^{2}.[YFW].*[GSA].^{2}.[WYF]...^{7}....[KRQ].^{2}.[YFW]...^{7}....[KRQ].^{2}.[YFH]...^{6}...[YFW]‡

A search of the databases with this signature using a pattern-matching program, identified approximately 50 HMG-1 box containing proteins with very few false positives (<4) (i.e. not known DNA-binding proteins). A more discerning signature such as:

[KR][PA].^{3}..[YF].^{2}.[YFW].....^{12}.....[P].*[GSA].^{2}.W...^{7}....[KR].^{2}.[YFW].....^{10}.....[YFH]...^{6}...[YFW]

identifies approximately 35 HMG-1 box proteins with no false positives. The most noticeable characteristic of this motif is the conservation of the position and spacing between the aromatic amino acids. Several proteins currently classified as HMG-1 box containing do not fit these criteria (e.g. CCG1[59], SIN1/SPT2[60], TFC3[61] and PF16[62]).

Conserved residues are at positions critical for a protein to maintain the structural requirements necessary for its function. From an information theoretical perspective, the positions at these conserved residues are characterized by a relatively large amount of information. The amount of information at various positions can be visualized by a sequence logo[63].

From the HMG-1 box sequence logo depicted in Fig. 3, it is apparent that positions 1-11 have a high total information content. In addition, there are specific sites (e.g. positions 1-4, 6-8, 11, 16, 24, 38, 41, 44, 52, 55, 63, 66 and 73) which also have a high information content (i.e. >2 bits). The potential importance of these high information positions for the function of the HMG-1 box proteins could be tested by site-directed mutagenesis experiments. Indeed, mutations in the human sex-determining SRY, at positions 1, 5, 36, 41 and 52 (see circles in Fig. 3), altered the binding specificity of the

‡[ ] enclose residues which could be at that position. A period could be any amino acid. The number of the periods between conserved positions is indicated. The asterisk (*) indicates that this spacing is not fixed.

protein and are associated with sex reversal phenotypes[42,64,65].

Conclusions and Future Studies

It is now established that the HMG-1 box is a distinct structural motif present in a family of DNA-binding regulatory proteins. The members of this family of proteins contain a unique signature characterized by regularly spaced, aromatic amino acids. This signature should be useful for identification of new members of this family of proteins. The emerging consensus DNA structure for HMG-1 box binding, is a local deformation of the path of the DNA helix at the site of interaction. The proteins deform the DNA but also recognize bent DNA, perhaps with little sequence specificity. The canonical HMG-1/-2 proteins recognize bent DNA and may be necessary for quenching non-specific binding of HMG-1 box proteins[29]. The structure of the HMG-1 box from several proteins is currently under investigation. These and future studies on the binding of HMG-1 domains to chromatin and DNA will clarify the structural and functional relationships between various members of the HMG-1 box family of DNA-binding proteins. In addition, the cellular functions and mode of interaction will elucidate the common features between all these HMG-1 box proteins. However, we feel that the real challenge is to determine the cellular function of the canonical HMG-1 and HMG-2 proteins. The high sequence conservation of these latter proteins suggests a stringent requirement for a defined structure that may be involved in multiple interactions. The experimental approaches which will help solve these questions include the identification of functional homologs in lower eukaryotes and altering the cellular content of the proteins in tissue culture or in more complex systems. The latter could be accomplished by either gene transfer or gene inactivation.

Acknowledgements

We would like to thank Drs A. D. Baxevanis, E. Koonin and M. S. Boguski for critically reviewing this manuscript and Dr G. D. Schuler for help with the preparation of the figures.

Note Added In Proof

Subsequent to submission of the manuscript, new information has been published which describes details of a novel three-dimensional structure for one of the rat HMG-1 box domains, as determined by 2-D ^{1}H-NMR spectroscopy of a peptide expressed in *Escherichia coli*[67]. The data shows that there are three α-helices forming an L-shaped structure with an angle of about 80° between the anti-parallel helices 1 and 2, and helix 3. There is a cluster of four aromatic residues (2 phenylalanines, 1 tryptophan and 1 tyrosine) and a single lysine residue in the bend of the L-shape. These residues correspond to high information content peaks which we have identified in Fig. 3 (i.e. positions 8, 11, 44, 52, 55). These evolutionarily conserved positions could be the minimal requirements for some of the structural determinants of other HMG-1 box domains.

References

1 Bairoch, A. (1991). PROSITE: a dictionary of sites and patterns in proteins. *Nucl. Acids Res.* **19,** 2241-5.

2 Pabo, C. O. and Sauer, R. T. (1992). Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.* **61,** 1053-95.

3 Wolffe, A. P., Tafuri, S., Ranjan, M. and Familari, M. (1992). The Y-box factors: a family of nucleic acid binding proteins conserved from *Escherichia coli* to man. *New Biol.* **4,** 290-8.

4 Johns, E. W. (1982). *The HMG Chromosomal Proteins.* Academic Press, London.

5 Einck, L. and Bustin, M. (1985). The intracellular distribution and function of the high mobility group chromosomal proteins. *Exp. Cell Res.* **156,** 295-310.

6 Bustin, M., Lehn, D. A. and Landsman, D. (1990). Structural features of the HMG chromosomal proteins and their genes. *Biochim. Biophys. Acta* **1049,** 231-43.

7 Jantzen, H. M., Admon, A., Bell, S. P. and Tjian, R. (1990). Nucleolar transcription factor hUBF contains a DNA-binding motif with homology to HMG proteins. *Nature* **344,** 830-6.

8 Sinclair, A. H., Berta, P., Palmer, M. S., Hawkins, J. R., Griffiths, B. L., Smith, M. J., Foster, J. W., Frischauf, A. M., Lovell, B. R. and Goodfellow, P. N. (1990). A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif [see comments]. *Nature* **346,** 240-4.

9 Gubbay, J., Collignon, J., Koopman, P., Capel, B., Economou, A., Münsterberg, A., Vivian, N., Goodfellow, P. and Lovell, B. R. (1990). A gene mapping to the sex-determining region of the mouse Y chromosome is a member of a novel family of embryonically expressed genes [see comments]. *Nature* **346,** 245-50.

10 Travis, A., Amsterdam, A., Belanger, C. and Grosschedl, R. (1991). LEF-1, a gene encoding a lymphoid-specific protein with an HMG domain, regulates T-cell receptor alpha enhancer function [corrected] [published erratum appears in *Genes Dev.* (1991) **5:** following 1113]. *Genes Dev.* **5,** 880-94.

11 van de Wetering, M., Oosterwegel, M., Dooijes, D. and Clevers, H. (1991). Identification and cloning of TCF-1, a T lymphocyte-specific transcription factor containing a sequence-specific HMG box. *EMBO J.* **10,** 123-32.

12 Waterman, M. L., Fischer, W. H. and Jones, K. A. (1991). A thymus-specific member of the HMG protein family regulates the human T cell receptor C alpha enhancer. *Genes Dev.* **5,** 656-69.

13 Diffley, J. F. and Stillman, B. (1991). A close relative of the nuclear, chromosomal high-mobility group protein HMG1 in yeast mitochondria. *Proc. Natl. Acad. Sci. USA* **88,** 7864-8.

14 Parisi, M. A. and Clayton, D. A. (1991). Similarity of human mitochondrial transcription factor 1 to high mobility group proteins. *Science* **252,** 965-9.

15 Shirakata, M., Hüppi, K., Usuda, S., Okazaki, K., Yoshida, K. and Sakano, H. (1991). HMG1-related DNA-binding protein isolated with V-(D)-J recombination signal probes. *Mol. Cell. Biol.* **11,** 4528-36.

16 Bruhn, S. L., Pil, P. M., Essigman, J. M., Housman, D. E. and Lippard, S. J. (1992). Isolation and characterization of human cDNA clones encoding a high mobility group box protein that recognizes structural distortions to DNA caused by binding of the anticancer agent cisplatin. *Proc. Natl. Acad. Sci. USA* **89,** 2307 2311.

17 Reeck, G. R., Isackson, P. J. and Teller, D. C. (1982). Domain structure in high molecular weight high mobility group nonhistone chromatin proteins. *Nature* **300,** 76-8.

18 Bianchi, M. E., Falciola, L., Ferrari, S. and Lilley, D. M. (1992). The DNA binding site of HMG1 protein is composed of two similar segments (HMG boxes), both of which have counterparts in other eukaryotic regulatory proteins. *EMBO J.* **11,** 1055-63.

19 Hamada, H. and Bustin, M. (1985). Hierarchy of binding sites for chromosomal proteins HMG 1 and 2 in supercoiled deoxyribonucleic acid. *Biochemistry* **24,** 1428-33.

20 Bianchi, M. E., Beltrame, M. and Paonessa, G. (1989). Specific recognition of cruciform DNA by nuclear protein HMG1. *Science* **243,** 1056-9.

21 Pil, P. M. and Lippard, S. J. (1992). Specific binding of chromosomal protein HMG1 to DNA damaged by the anticancer drug cisplatin. *Science* **256,** 234-237.

22 Hughes, E. N., Engelsberg, B. N. and Billings, P. C. (1992). Purification of nuclear proteins that bind to cisplatin-damaged DNA. Identity with high mobility group proteins 1 and 2. *J. Biol. Chem.* **267,** 13520-7.

23 Butler, A. P., Mardian, J. K. and Olins, D. E. (1985). Nonhistone chromosomal protein HMG 1 interactions with DNA. Fluorescence and thermal denaturation studies. *J. biol. Chem.* **260,** 10613-20.

24 Lilley, D. M. (1992). DNA-protein interactions. HMG has DNA wrapped up. *Nature* **357,** 282-3.

25 Tremethick, D. J. and Molloy, P. L. (1988). Effects of high mobility group proteins 1 and 2 on initiation and elongation of specific transcription by RNA polymerase II in vitro. *Nucl. Acids Res.* **16,** 11107-23.

26 Waga, S., Mizuno, S. and Yoshida, M. (1990). Chromosomal protein HMG1 removes the transcriptional block caused by the cruciform in supercoiled DNA. *J. biol. Chem.* **265,** 19424-8.

27 Singh, J. and Dixon, G. H. (1990). High mobility group proteins 1 and 2 function as general class II transcription factors. *Biochemistry* **29,** 6295-302.

28 Landsman, D. and Bustin, M. (1991). Assessment of the transcriptional activation potential of the HMG chromosomal proteins. *Mol. Cell. Biol.* **11,** 4483-9.

29 Ferrari, S., Harley, V. R., Pontiggia, A., Goodfellow, P. N., Lovell-Badge, R. and Bianchi, M. E. (1992). SRY, like HMG1, recognizes sharp angles in DNA. *EMBO J.* **11,** 4497-4506.

30 Jantzen, H.-M., Chow, A. M., King, D. S. and Tjian, R. (1992). Multiple domains of the RNA polymerase-I activator hUBF interact with the TATA-binding protein complex hSL1 to mediate transcription. *Genes Dev.* **6,** 1950-63.
31 Kuhn, A. and Grummt, I. (1992). Dual role of the nucleolar transcription factor UBF: trans-activator and antirepressor. *Proc. Natl. Acad. Sci. USA* **89,** 7340-4.
32 Maeda, Y., Hisatake, K., Kondo, T., Hanada, K.-I., Song, C.-Z., Nishimura, T. and Muramatsu, M. (1992). Mouse rRNA gene transcription factor mUBF requires both HMG-box1 and an acidic tail for nucleolar accumulation: molecular analysis of the nucleolar targeting mechanism. *EMBO J.* **11,** 3695-3705.
33 Waterman, M. L. and Jones, K. A. (1990). Purification of TCF-1 alpha, a T-cell-specific transcription factor that activates the T-cell receptor C alpha gene enhancer in a context-dependent manner. *New Biol.* **2,** 621-36.
34 Alexander-Bridges, M., Ercolani, L., Kong, X. F. and Nasrin, N. (1992). Identification of a core motif that is recognized by three members of the HMG class of transcriptional regulators: IRF-ABP, SRY, and TCF-1 alpha. *J. Cell. Biochem.* **48,** 129-35.
35 Zitomer, R. S. and Lowry, C. V. (1992). Regulation of gene expression by oxygen in *Saccharomyces cerevisiae*. *Microbiol. Rev.* **56,** 1-11.
36 Diffley, J. F. and Stillman, B. (1992). DNA binding properties of an HMG1-related protein from yeast mitochondria. *J. biol. Chem.* **267,** 3368-74.
37 Fisher, R. P., Lisowsky, T., Parisi, M. A. and Clayton, D. A. (1992). DNA wrapping and bending by a mitochondrial high mobility group-like transcriptional activator protein. *J. biol. Chem.* **267,** 3358-67.
38 Sugimoto, A., Iino, Y., Maeda, T., Watanabe, Y. and Yamamoto, M. (1991). *Schizosaccharomyces pombe* ste11+ encodes a transcription factor with an HMG motif that is a critical regulator of sexual development. *Genes Dev.* **5,** 1990-9.
39 Kelly, M., Burke, J., Smith, M., Klar, A. and Beach, D. (1988). Four mating-type genes control sexual differentiation in the fission yeast. *EMBO J.* **7,** 1537-47.
40 Staben, C. and Yanofsky, C. (1990). *Neurospora crassa* a mating-type region. *Proc. Natl. Acad. Sci. USA* **87,** 4917-21.
41 Debuchy, R. and Coppin, E. (1992). The mating types of *Podospora anserina*: functional analysis and sequence of the fertilization domains. *Mol. Gen. Genet.* **233,** 113-21.
42 Harley, V. R., Jackson, D. I., Hextall, P. J., Hawkins, J. R., Berkovitz, G. D., Sockanathan, S., Lovell, B. R. and Goodfellow, P. N. (1992). DNA binding activity of recombinant SRY from normal males and XY females. *Science* **255,** 453-6.
43 Denny, P., Swift, S., Connor, F. and Ashworth, A. (1992). An SRY-related gene expressed during spermatogenesis in the mouse encodes a sequence-specific DNA-binding protein. *EMBO J.* **11,** 3705-12.
44 Ner, S. S. (1992). HMGs everywhere. *Current Biology* **2,** 208-210.
45 Dunaway, M. (1989). A transcription factor, TFIS, interacts with both the promoter and enhancer of the *Xenopus* rRNA genes. *Genes Dev.* **3,** 1768-78.
46 Giese, K., Cox, J. and Grosschedl, R. (1992). The HMG domain of lymphoid enhancer factor 1 bends DNA and facilitates assembly of functional nucleoprotein structures. *Cell* **69,** 185-95.
47 Oosterwegel, M. A., van de Wetering, M., Holstege, F. C., Prosser, H. M., Owen, M. J. and Clevers, H. C. (1991). TCF-1, a T cell-specific transcription factor of the HMG box family, interacts with sequence motifs in the TCR beta and TCR delta enhancers. *Int. Immunol.* **3,** 1189-92.
48 van de Wetering, M. and Clevers, H. (1992). Sequence-specific interaction of the HMG box proteins TCF-1 and SRY occurs within the minor groove of a Watson-Crick double helix. *EMBO J.* **11,** 3039-44.
49 Yanagi, K., Privé, G. G. and Dickerson, R. E. (1991). Analysis of local helix geometry in three B-DNA decamers and eight dodecamers. *J. Mol. Biol.* **217,** 201-14.
50 Heinemann, U. and Hahn, M. (1992). C-C-A-G-G-C-m5C-T-G-G. Helical fine structure, hydration, and comparison with C-C-A-G-G-C-C-T-G-G. *J. biol. Chem.* **267,** 7332-41.
51 Travers, A. A. (1991). DNA bending and kinking - sequence dependence and function. *Curr. Op. Struc. Biol.* **1,** 114-22.
52 Bolshoy, A., McNamara, P., Harrington, R. E. and Trifonov, E. N. (1991). Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles. *Proc. Natl. Acad. Sci. USA* **88,** 2312-6.
53 Mosevitsky, M. I., Novitskaya, V. A., Iogannsen, M. G. and Zabezhinsky, M. A. (1989). Tissue specificity of nucleo-cytoplasmic distribution of HMG1 and HMG2 proteins and their probable functions. *Eur. J. Biochem.* **185,** 303-10.
54 Bustin, M., Crippa, M. P. and Pash, J. M. (1992). Expression of HMG chromosomal protein during the cell cycle and differentiation. *Crit. Rev. Gene Expression* **2,** 137-43.
55 Merenmies, J., Pihlaskari, R., Laitinen, J., Wartiovaara, J. and Rauvala, H. (1991). 30-kDa heparin-binding protein of brain (amphoterin) involved in neurite outgrowth. Amino acid sequence and localization in the filopodia of the advancing plasma membrane. *J. biol. Chem.* **266,** 16722-9.
56 Daston, M. M. and Ratner, N. (1991). Expression of P30, a protein with adhesive properties, in Schwann cells and neurons of the developing and regenerating peripheral nerve. *J. Cell Biol.* **112,** 1229-39.
57 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215,** 403-10.
58 Wen, L., Huang, J. K., Johnson, B. H. and Reeck, G. R. (1989). A human placental cDNA clone that encodes nonhistone chromosomal protein HMG-1. *Nucl. Acids Res.* **17,** 1197-214.
59 Sekiguchi, T., Nohiro, Y., Nakamura, Y., Hisamoto, N. and Nishimoto, T. (1991). The human CCG1 gene, essential for progression of the G1 phase, encodes a 210-kilodalton nuclear DNA-binding protein. *Mol. Cell. Biol.* **11,** 3317-25.
60 Kruger, W. and Herskowitz, I. (1991). A negative regulator of HO transcription, SIN1 (SPT2), is a nonspecific DNA-binding protein related to HMG1. *Mol. Cell. Biol.* **11,** 4135-46.
61 Lefebvre, O., Carles, C., Conesa, C., Swanson, R. N., Bouet, F., Riva, M. and Sentenac, A. (1992). *TFC3*: Gene encoding the B-block binding subunit of the yeast transcription factor IIIC. *Proc. Natl. Acad. Sci. USA* **89,** 10512-6.
62 Guntaka, R. V., Kandala, J. C. and Reddy, V. D. (1992). Cloning and characterization of a highly conserved HMG-like protein (PF16) gene from *Plasmodium falciparum*. *Biochem. Biophys. Res. Commun.* **182,** 412-9.
63 Schneider, T. D. and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucl. Acids Res.* **18,** 6097-100.
64 Nasrin, N., Buggs, C., Kong, X. F., Carnazza, J., Goebl, M. and Alexander, B. M. (1991). DNA-binding properties of the product of the testis-determining gene and a related protein. *Nature* **354,** 317-20.
65 Hawkins, J. R., Taylor, A., Berta, P., Levilliers, J., Van der Auwera, B. and Goodfellow, P. N. (1992). Mutational analysis of SRY: nonsense and missense mutations in XY sex reversal. *Hum. Genet.* **88,** 471-4.
66 Barton, G. J. (1993). ALSCRIPT: A tool to format multiple sequence alignments. *Protein Eng.* **6,** 37-40.
67 Weir, H. M., Kraulis, P. J., Hill, C. S., Raine, A. R. C., Laue, E. D. and Thomas, J. O. (1993). Structure of the HMG box motif in the B-domain of HMG1. *EMBO J.* **12,** 1311-1319.

David Landsman* is at the National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, Room 8N-807, Bethesda, MD 20894, USA. **Michael Bustin** is at the Laboratory of Molecular Carcinogenesis, National Cancer Institute, National Institutes of Health, Building 37, Room 3D12, Bethesda, MD 20892, USA.
*Corresponding author.