Two groups of HMG box proteins are distinguished. Proteins in the first group contain multiple HMG boxes, are non-sequence-specific, and recognize structural features as found in cruciform DNA and cross-over DNA. The abundant chromosomal protein HMG-1 belongs to this subgroup. Proteins in the second group carry a single HMG box with affinity for the minor groove of the heptamer motif AACAAG or variations thereof. A solution structure for the non-sequence-specific C-terminal HMG box of HMG-1 has recently been proposed. Now, we report the solution structure of the sequence-specific HMG box of the SRY-related protein Sox-4. NMR analysis demonstrated the presence of three α-helices (Val10-Gly22, Gln22-Leu41 and Phe50-Tyr65) connected by loop regions (Ser23-Ala49 and Leu42-Pro49). Helices I and II are positioned in an antiparallel mode and form one arm of the HMG box. Helix III is less rigid, makes an average angle of about 90° with helices I and II, and constitutes the other arm of the molecule. As in HMG1B, the overall structure of the Sox-4 HMG box is L-shaped and is maintained by a cluster of conserved, mainly aromatic residues.

The cloning of the RNA polymerase I transcription factor UBF1 (1) has originally led to the recognition of a novel type of DNA-binding domain, the so-called HMG box. The HMG box was named after its homology with high mobility group (HMG)-1 proteins and is defined by a loose consensus sequence of about 80 amino acids (2). At this moment, more than 60 proteins with one or more HMG boxes have been reported. An evolutionary study of the HMG box family indicated that two major subfamilies can be discriminated (3). One of these subfamilies contains proteins with a single HMG box, which binds with high sequence specificity to variants of the DNA sequence (A/T)(A/T)CAAAG. Members of this subfamily include products of the mammalian sex-determinator Sry and related Sox genes (Sry HMG box-containing genes) (4, 5), the Schizosaccharomyces pombe transcription factor Ste11+ (6), the lymphoid factors TCF-1 (7, 8) and LEF-1 (9, 10), and the products of several fungal type genes such as Mat-Mc of S. pombe (11) and Mt a1 of Neurospora crassa (12).

DNA binding occurs in the minor groove, as was shown for TCF-1, LEF-1, Mat-Mc, SRY, and Sox-4 by methylation-and diethylpyrocarbonate carboxylation interference footprinting and T(CT)A/I nucleotide substitutions (13–16) and is accompanied by the induction of a strong bend in the DNA helix (14, 16–18). A bend-swap experiment demonstrated that LEF-1 and its specific DNA-binding motif can functionally replace bending induced by the integration host factor at the attP locus in phage λ integrase reaction (16).

The other subfamily includes proteins with multiple HMG boxes and with a rather nonspecific affinity for DNA, such as the HMG-1 and -2 proteins (19), UBF (1) and mtTF1 (20). Characteristic of these HMG boxes is their affinity for the cis-platinated-GG-adduct in DNA (21, 22) and cruciform DNA (23, 24), independent of sequence determinants. This suggested that the non-sequence-specific HMG boxes recognize DNA structure instead of DNA sequence (25).

Circular dichroism measurements and secondary structure prediction methods indicated a high α-helical content for sequence-specific HMG domains (17). This is consistent with NMR studies on the tertiary structure of the second HMG box of HMG1 (26, 27) and HMG-D (28). The 60-amino acid core of these non-sequence-specific HMG boxes consists of three α-helices, which form an unusual L-shaped molecule. The angle between the two arms is 70–80° and is defined by a cluster of conserved, aromatic residues (26–28). Based on an identical secondary structure observed for the HMG box of Sox-5, a similar L-shaped structure has been suggested for this sequence-specific HMG domain (29).

Hydroporphic interactions of the HMG box of SRY with DNA by partial isoleucine side chain intercalation predicts the positioning of an α-helix into a widened minor groove and might account for sequence specificity and DNA bending (30). Using the solution structure of rat HMG1B (26) a model for the SRY-DNA complex was proposed (31).

Since a detailed structural sequence for a sequence-specific HMG box has not yet been determined, we have pursued the elucidation of the NMR solution structure of the HMG box domain of the lymphocyte transcriptional activator Sox-4. This HMG box shows high sequence-specific binding toward the AACAAAG DNA-binding motif with a Kd of 10−11 M (15). The biological significance of the Sox-4 gene has recently been underscored in a gene disruption experiment. Mice carrying two null alleles of Sox-4 fail to develop functional valves in the heart and have a severe block in early lymphoid development.2 The NMR data indicate that the secondary structure of Sox-4 HMG box is...
closely related to that of Sox-5 (29) and that the overall fold compares well with that of HMGB1 (26, 27) and HMG-D (28).

MATERIALS AND METHODS

Plasmid Construction—The Sox-4 HMG box was cloned by PCR from pCAGGS DNA using the primers 5′-ATACATATGGCTAAGCCAGG-CCACGGGCCAC-3′ and 5′-CCCGGATCTCCAGGCTTCCGCG-3′ and inserted between the NdeI and BamHI sites of pET-3c (32, 33). The identity of the subcloned HMG-box fragment was confirmed by DNA sequencing. The resulting plasmid was transformed into Escherichia coli strain BL21 (DE3).

Production and Purification of the Sox-4 HMG Box Peptide—The production and purification of the Sox-4 HMG box peptide was basically done as described for the HMG box of TCF-1 (17). For the production of unlabeled HMG box peptide the transformed cells were grown at 37°C in LB, while for the production of uniformly 15N-labeled protein the cells were grown in minimal medium containing 15NH4Cl. Both media contained 100 mg/ml ampicillin. In the midlog phase, the cells were induced with 0.3 mM isopropyl-1-thio-

Expression, Purification, and Characterization of the HMG Box of Sox-4

The HMG box of murine Sox-4 (amino acids 59–135) (15) was produced in a T7-based expression system (32, 33). For this purpose a pET-3/Sox-4 HMG box plasmid was constructed. The identity of the inserted Sox-4 HMG box fragment was con-
firmed by DNA sequencing. This recombinant plasmid was transformed into E. coli BL21(DE3), where the Sox-4 HMG box was overexpressed after isopropyl-1-thio-β-D-galactopyranoside induction (Fig. 1A). The overexpressed HMG box peptide was purified to homogeneity in a single-step cation exchange chromatographic run. A typical elution profile is presented in Fig. 1B. The procedure yielded 1–2 mg of Sox-4 HMG box protein/liter of bacterial culture with a purity greater than 95% as judged from a silver-stained SDS-polyacrylamide gel. The identity of the first 10 amino acids of the isolated peptide was confirmed by protein sequencing. The DNA binding activity of the protein was established in a gel retardation assay (Fig. 1C).

Circular Dichroism

Fig. 2 shows the CD spectrum of the HMG box peptide of Sox-4. Deconvolution of the spectrum predicted a secondary structure with 54% α-helix, 11% β-sheet, and 35% random coil. A similar high α-helical content was observed for the HMG boxes of TCF-1 (17), HMG1B (26, 27), HMG-D (28), and Sox-5 (29).

NMR Measurements

Assignment—Unlabeled as well as uniformly 15N-labeled Sox-4 HMG box samples were used for NMR spectroscopy. The NMR data were collected at pH 6.5 and at temperatures of 293 and 298 K. Conditions with a pH lower than 6.5 resulted in precipitation of the protein, while at temperatures above 298 K the protein starts to unfold. The predominantly α-helical nature of the protein results in a limited chemical shift dispersion. This causes a severe overlap in the amide and fingerprint region of the NOESY and TOCSY spectra and makes it difficult to assign these spectra completely. However, the 15N signals of the various residues are well separated in the two-dimensional 15N-1H HSQC experiment (Fig. 3). Therefore, we collected three-dimensional 15N-1H NOESY-HSQC and three-dimensional 15N-1H TOCSY-HSQC data at 293 and 298 K. The majority of sequential assignments of the amino acid side chain was found by comparison of the amide region of three-dimensional 15N-1H NOESY-HSQC and three-dimensional 15N-1H TOCSY-HSQC spectra. In some cases additional information of NOESY, TOCSY, and/or two-dimensional 15N-1H HSQC spectra was helpful. First, stretches of spin systems with sequential 15N-1H and Cα-15N NOE contacts were identified in the three-dimensional 15N-1H NOESY-HSQC spectrum. Assignment of these spin systems to specific residues was done by comparison of the amino acid side-chain resonances in the three-dimensional 15N-1H NOESY-HSQC, three-dimensional 15N-1H TOCSY-HSQC, NOESY, and TOCSY spectra. Two-dimensional spectra were especially helpful for the assignment of the side chains of the aromatic residues. Using this strategy we were able to assign more than 80% of the backbone resonances of the HMG box. The 15N and 1H chemical shift data are deposited at the BioMagResBank (University of Wisconsin, Madison) (Table I). Although we collected data sets at two temperatures (293 and 298 K), some residues at the N and C termini could not be identified, due to flexibility and/or overlap.
Also, the assignments of Asn\textsuperscript{28}, Ala\textsuperscript{29}, Lys\textsuperscript{47}, Ile\textsuperscript{48}, and Pro\textsuperscript{49}, which are located in the two loop regions, could not be established.

Secondary Structure—In total we found 426 identifiable interresidue NOE cross-peaks in the NOESY (100- and 150-ms mixing times) and the three-dimensional \textsuperscript{15}N-\textsuperscript{1}H NOESY-HSQC (150-ms mixing time) spectra: 256 sequential (i to (i + 1)), 106 medium range (i to \(\geq(i + 2)\) and \(\leq(i + 4)\)), and 56 long range (i to \(\geq(i + 5)\)). A similar low number of interresidue NOE contacts was found for the non-sequence-specific HMG-D box (28), while for the HMG1B box about twice as many interresidual NOEs were observed (26, 27).

The observation of stretches of strong \(d_{\text{HN}}(i, i + 1)\) and weak \(d_{\text{HN}}(i, i + 3)\) connectivities in combination with \(d_{\text{AN}}(i, i + 1)\), \(d_{\text{AN}}(i, i + 4)\), and \(\alpha_{\text{HN}}(i, i + 3)\) contacts (48) in the three-dimensional \textsuperscript{15}N-\textsuperscript{1}H NOESY-HSQC and NOESY spectra provided evidence for the existence of three \(\alpha\)-helical regions in the Sox-4 HMG box. The \(\alpha\)-helices are formed by residues Val\textsuperscript{10}, Glu\textsuperscript{30}, Leu\textsuperscript{41} and Phe\textsuperscript{50}, Tyr\textsuperscript{65} (Fig. 5). Based on these NMR data an \(\alpha\)-helical content of 53% was calculated for the Sox-4 HMG box. This is consistent with the analysis of the CD spectrum of the Sox-4 HMG box (Fig. 2), which revealed an \(\alpha\)-helical content of 54% (see "Circular Dichroism").

Fast and intermediate exchanging NH protons with water were identified from the difference of a NH sensitivity-enhanced \textsuperscript{15}N-HSQC experiment with and without presaturation. Fast exchanging NH protons are mainly found outside the helical regions with exception of helix III, which also contains a number of fast and intermediate exchanging NH protons (Fig. 5). A more or less similar distribution of mobile backbone NH protons was observed in a heteronuclear NOE experiment (Figs. 5 and 6A). These observations are in accordance with a less rigid and more exposed character of helix III. The most instable region of helix III is Glu\textsuperscript{55}-Arg\textsuperscript{56}-Leu\textsuperscript{57}-Arg\textsuperscript{58}-Leu\textsuperscript{59} as

![Fig. 3. \textsuperscript{15}N-\textsuperscript{1}H HSQC spectrum of the Sox-4 HMG box peptide.](http://www.jbc.org/)

![Fig. 4. Sequential \textsuperscript{15}NH-\textsuperscript{15}NH and C\textsubscript{\text{H}}-\textsuperscript{15}NH NOE contacts in helix I (Val\textsuperscript{10}-Gly\textsuperscript{22}) of the Sox-4 HMG box peptide.](http://www.jbc.org/)
| Residue | $^1$H | NH | C$^\alpha$H | C$^\beta$H | Others |
|---------|-------|----|-------------|-------------|---------|
| Ile$^4$ | 123.83 | 8.00 | 3.95 | 1.63 | C$^\alpha$H 1.24; C$^\beta$H 0.93; C$^\gamma$H 0.63 |
| Lys$^2$ | 127.21 | 8.33 | 4.16 | 1.58 | C$^\gamma$H 1.28 |
| Arg$^3$ | 117.49 | 6.73 | 3.95 | 1.64 | C$^\gamma$H 1.36,1.46; C$^\beta$H 2.98 |
| Pro$^6$ | 120.59 | 7.94 | 3.95 | 1.67 | C$^\gamma$H 1.24,1.41; C$^\beta$H 0.83; C$^\gamma$H 0.65 |
| Met$^9$ | 120.53 | 7.93 | 3.94 | 1.69 | C$^\gamma$H 1.24,1.41 |
| Ser$^{12}$ | 114.96 | 8.43 | 4.54 | 2.54 | C$^\gamma$H 1.58,3.72 |
| Glu$^{15}$ | 121.77 | 7.76 | 3.50 | 1.92 | C$^\gamma$H 3.7 |
| Gly$^{36}$ | 112.56 | 8.12 | 4.02 | 3.86 | C$^\gamma$H 1.37,1.57 |
| Lys$^{33}$ | 123.09 | 7.64 | 4.10 | 1.94 | C$^\gamma$H 1.37; C$^\beta$H 1.55 |
| Arg$^{34}$ | 120.97 | 8.21 | 4.07 | 1.90 | C$^\gamma$H 1.37; C$^\beta$H 1.57 |
| Leu$^{35}$ | 120.01 | 8.82 | 4.25 | 1.85 | C$^\gamma$H 1.37; C$^\beta$H 1.57 |
| Met$^{20}$ | 121.77 | 7.94 | 3.95 | 1.67 | C$^\gamma$H 1.37; C$^\beta$H 1.57 |
| Lys$^{11}$ | 121.17 | 8.03 | 3.55 | 2.37 | C$^\gamma$H 1.37; C$^\beta$H 1.57 |
| Trp$^{11}$ | 121.18 | 8.34 | 3.51 | 2.76 | C$^\gamma$H 1.37; C$^\beta$H 1.57 |
| Ser$^{12}$ | 114.86 | 8.77 | 3.45 | 3.03 | C$^\gamma$H 2.37 |
| Gln$^{13}$ | 120.88 | 7.32 | 3.77 | 1.77,2.02 | C$^\gamma$H 2.37 |
| Ile$^{14}$ | 119.41 | 6.73 | 3.54 | 1.76 | C$^\gamma$H 1.64; C$^\beta$H 0.98; C$^\gamma$H 0.76 |
| Gln$^{52}$ | 119.86 | 7.39 | 4.13 | 1.90,2.07 | C$^\gamma$H 1.64; C$^\beta$H 0.98; C$^\gamma$H 0.76 |
| Leu$^{54}$ | 121.77 | 8.00 | 3.94 | 2.04 | C$^\gamma$H 1.64; C$^\beta$H 0.98; C$^\gamma$H 0.76 |
| Phe$^{50}$ | 118.24 | 8.12 | 3.90 | 2.46,2.63 | C$^\gamma$H 1.64; C$^\beta$H 0.98; C$^\gamma$H 0.76 |
| Met$^{9}$ | 122.07 | 8.43 | 4.25 | 1.89,2.10 | C$^\gamma$H 1.64; C$^\beta$H 0.98; C$^\gamma$H 0.76 |
| Lys$^{47}$ | 123.09 | 7.64 | 4.10 | 1.94 | C$^\gamma$H 1.64; C$^\beta$H 0.98; C$^\gamma$H 0.76 |
| Ala$^{29}$ | 114.96 | 8.43 | 4.54 | 2.54 | C$^\gamma$H 1.64; C$^\beta$H 0.98; C$^\gamma$H 0.76 |
| Gly$^{36}$ | 112.56 | 8.12 | 4.02 | 3.86 | C$^\gamma$H 1.64; C$^\beta$H 0.98; C$^\gamma$H 0.76 |
| Gly$^{36}$ | 112.56 | 8.12 | 4.02 | 3.86 | C$^\gamma$H 1.64; C$^\beta$H 0.98; C$^\gamma$H 0.76 |
| Lys$^{33}$ | 123.09 | 7.64 | 4.10 | 1.94 | C$^\gamma$H 1.64; C$^\beta$H 0.98; C$^\gamma$H 0.76 |
| Arg$^{34}$ | 120.97 | 8.21 | 4.07 | 1.90 | C$^\gamma$H 1.64; C$^\beta$H 0.98; C$^\gamma$H 0.76 |
| Leu$^{35}$ | 120.01 | 8.82 | 4.25 | 1.85 | C$^\gamma$H 1.64; C$^\beta$H 0.98; C$^\gamma$H 0.76 |
| Glu$^{15}$ | 121.77 | 8.00 | 3.94 | 2.04 | C$^\gamma$H 1.64; C$^\beta$H 0.98; C$^\gamma$H 0.76 |
| Gly$^{36}$ | 112.56 | 8.12 | 4.02 | 3.86 | C$^\gamma$H 1.64; C$^\beta$H 0.98; C$^\gamma$H 0.76 |
| Arg$^{38}$ | 118.53 | 8.42 | 4.03 | 2.25,2.11 | C$^\gamma$H 1.64; C$^\beta$H 0.98; C$^\gamma$H 0.76 |
| Trp$^{39}$ | 121.32 | 8.51 | 3.57 | 3.03,3.23 | C$^\gamma$H 1.64; C$^\beta$H 0.98; C$^\gamma$H 0.76 |
| Lys$^{50}$ | 116.33 | 7.20 | 3.78 | 1.84 | C$^\gamma$H 1.41; C$^\beta$H 2.95 |
| Leu$^{41}$ | 117.06 | 7.12 | 4.07 | 1.68 | C$^\gamma$H 1.54; C$^\beta$H 0.77 |
| Leu$^{42}$ | 122.79 | 6.86 | 4.03 | 1.02 | C$^\gamma$H 0.45 |
| Lys$^{43}$ | 124.28 | 8.82 | 4.03 | 1.63 | C$^\gamma$H 1.46 |
| Met$^{57}$ | 120.70 | 8.43 | 3.94 | 1.82,2.13 | C$^\gamma$H 1.36; C$^\beta$H 0.81 |
| Arg$^{48}$ | 123.09 | 8.25 | 4.28 | 1.94 | C$^\gamma$H 3.07 |
| Leu$^{49}$ | 119.85 | 8.05 | 4.02 | 1.76 | C$^\gamma$H 1.42; C$^\beta$H 0.80 |
| Lys$^{50}$ | 120.89 | 7.86 | 4.03 | 1.80 | C$^\gamma$H 1.36; C$^\beta$H 1.62 |
is indicated by a patch of fast exchanging and mobile NH backbone protons. However, helix III is not flexible as indicated by the $T_1^r$ relaxation times (Fig. 6B). The different time scale of NHexchange, $1H-15N$NOE, and NH $T_1^r$ relaxation explains the seemingly contradictory results. Possible salt bridges between Arg17 and Glu21 in helix I, between Arg34 and Glu30 in helix II, and between Lys60 and Asp64 in helix III might contribute to helix stabilization (49). Loop regions are located between Ser23 and Ala29 and between Leu42 and Pro49 (Fig. 5).

The N-terminal residues Asn6–Met9 as well as the C-terminal amino acids between Pro66 and Pro72 have an extended conformation, as was indicated by the observation of strong sequential $d_{1N}$ and weak $d_{1H}$ contacts and the absence of most medium range NOE contacts (48). Turns involving 4 residues are characterized by a strong $d_{1N}(3,4)$ connectivity together with a $d_{1H}(2,4)$ contact (48, 50). Such a pattern was found in the sequence Leu42-Lys43-Asp44-Ser45. A very strong $d_{1N}$ contact was observed between Asp44 and Ser45. In addition, a $d_{1N}(i,i + 2)$ contact with medium intensity was detected between Lys33 and Ser45. Type I and type II turns can be distinguished from each other by the intensity of the $d_{1N}(2,3)$ (strong in type I, absent in type II) and $d_{1N}(2,3)$ (weak in type I, strong in type II) connectivities (48). Since Lys43 and Asp44 show a weak $d_{1N}(2,3)$ contact and a $d_{1N}(2,3)$ cross-peak with medium intensity, we were unable to classify this turn. However, in our refined structure this sequence has a type I turn conformation (see later). In the sequence Ser23-Pro24-Asp25-Met26, just after helix I, we found a very strong $d_{1N}$ contact between Asp25 and Met26 and a weak $d_{1N}(1,4)$ connectivity between Ser23 and Met26, suggesting the presence of a turn. Unfortunately, we were unable to identify a $d_{1H}(2,4)$ contact in this sequence, since the $C_{\alpha}H$ proton of Pro24 was not assigned. It is noted that this sequence has a type I turn structure in the final model (see later).

The unassigned residues Arg3–Asn6 and Arg73–Lys77 at the N and C termini are most probably flexible and unstructured.

### Table I—continued

| Residue | $^{15}N$ | NH | $^{1}H$ | $^{13}C$ | Others |
|---------|---------|----|------|-------|-------|
| His61   | 119.27  | 8.07| 4.38 | 3.11,3.24| 4H 6.93;2H 7.92 |
| Met62   | 117.80  | 7.90| 4.18 | 2.06 | C'H 2.55,2.61 |
| Ala63   | 122.06  | 7.69| 4.05 | 1.32 | |
| Asp64   | 118.53  | 7.94| 4.28 | 2.33,2.10| |
| Tyr65   | 119.85  | 7.86| 4.27 | 2.58,2.83| 2/6H 6.68;3/5H 6.68 |
| Pro66   | 119.25  | 8.35| 4.51 | 2.58,2.83| 2/6H 1.81,1.85;C'H3.43 |
| Tyr68   | 121.47  | 7.96| 4.34 | 2.99 | 2/6H 7.02;3/5H 6.72 |
| Lys69   | 123.39  | 7.88| 4.07 | 1.46 | C'H 1.06 |
| Tyr70   | 121.32  | 7.91| 4.38 | 2.88,2.94| 2/6H 7.03;3/5H 6.75 |
| Arg71   | 125.54  | 7.89| 4.42 | 1.50 | C'H3.46 |

Fig. 5. Sequential and medium range NOE contacts, NH proton exchange, backbone mobility, and $\alpha$-helical regions in the Sox-4 HMG box. Residues with high and intermediate backbone mobility are indicated by filled and open triangles, respectively. Those residues with low backbone mobility are indicated with plus signs. Filled and open circles indicate residues with fast and intermediate exchanging NH backbone protons, respectively. Slowly exchanging NH backbone protons are indicated by $\times$ signs. Boldface amino acids are not assigned (see "Assignment").
A similar pattern is observed when the RMSD value of the C-terminal backbone atoms is plotted against the residue number (Fig. 9). Those structures with D-amino acids and/or residues shorter than helix III of HMG1B (26, 27) and HMG-D (28) were excluded. The stereochemical quality of the structures was evaluated with the program PROCHECK (47). Those structures with cis-amino acids and/or cis peptide bonds were also discarded. A final set of 15 structures is presented in Fig. 8. The overall structure of the Sox-4 HMG box is L-shaped (Fig. 8). Helix I (Val\(^{10}\)-Gln\(^{22}\)) and II (Glu\(^{30}\)-Leu\(^{41}\)) are positioned in an antiparallel mode and form one arm of the molecule. The position of helix III (Phe\(^{50}\)-Tyr\(^{65}\)) varies, makes an average angle of \(-90^\circ\) with helices I and II, and constitutes the other arm of the L-shaped HMG box. The average pairwise RMSD value of the backbone atoms (N, C\(_\alpha\), C\(^\prime\)) of helix I (Val\(^{10}\)-Gln\(^{22}\)) and II (Glu\(^{30}\)-Leu\(^{41}\)) is 0.84 \(\pm\) 0.29 Å. As a result of the variable position of helix III this value goes up to 1.97 \(\pm\) 0.77 Å when helix III is included. However, the internal average pairwise RMSD of helix III (Phe\(^{50}\)-Tyr\(^{65}\)) is 0.76 \(\pm\) 0.26 Å. A similar pattern is observed when the RMSD value of the C\(_\alpha\) backbone atoms is plotted against the residue number (Fig. 9). In accordance with the T\(_{1\rho}\) relaxation times (Fig. 6B), these data indicate that in these computations helix III forms a helical element whose position varies relative to helix I and II. This variation is caused by the absence of long range NOE contacts between helix III and the other parts of the Sox-4 HMG box. Residues Ala\(^{7}\), Phe\(^{7}\), Met\(^{9}\), Val\(^{10}\), and Trp\(^{31}\), Trp\(^{39}\), Leu\(^{41}\), and Phe\(^{60}\) form a hydrophobic core and stabilize the structure of Sox-4 HMG box. Note that these residues with the exception of Lys\(^{41}\) are conserved within the HMG box family (2).

### DISCUSSION

Here, the NMR solution structure of the sequence-specific HMG box of Sox-4 is presented. The overall L-shape structure compares well with that reported for the non-sequence-specific HMG boxes of HMG1B (26, 27) and HMG-D (28), which recognize structural features of DNA (25). As in the HMG1B and HMG-D, three \(\alpha\)-helical regions dominate the HMG box structure of Sox-4. The sequential positions of helix I and II coincide with the corresponding helices in HMG1B (26, 27) and HMG-D (28). Helix III is positioned between proline 49 and 66 and is 4 residues shorter than helix III of HMG1B (26, 27) and HMG-D (28). Apparently, this results from the helix-breaking Pro\(^{66}\), which is unique to the sequence-specific HMG boxes (2) but is replaced by a structurally neutral alanine in HMG1B (26, 27) and by lysine in HMG-D (28). The helices I and II are followed by loops that start with type I turns (Ser\(^{23}\)-Met\(^{26}\) after helix I and Leu\(^{39}\)-Ser\(^{46}\) after helix II). The presence of such turns was not reported for HMG1B (26, 27) and HMG-D (28).

The overall HMG box fold is stabilized by a hydrophobic core involving residues Ala\(^{7}\), Phe\(^{7}\), Val\(^{10}\), and Trp\(^{31}\), Trp\(^{39}\), Leu\(^{41}\), and Phe\(^{60}\). With the exception of Leu\(^{41}\), these residues are conserved within the HMG box family, irrespective of their binding specificity (2). The structure of this hydrophobic core should be considered as the HMG box “signature.”

The mechanism of binding to DNA is fundamentally different for the two types of HMG boxes. The non-sequence-specific HMG1B box binds to preexisting structures (25), such as cruciform DNA (23, 24) and DNA bent by the cis-platinum-GG adduct (21, 22). The binding of the HMG box proteins to cruciform DNA has not been reported to induce conformational changes in the DNA. Therefore, it is likely that the rigid HMG1B-type box fits directly onto these unusual DNA structures. This is in contrast with the sequence-specific HMG box proteins, that alter the DNA conformation significantly. The binding of a monomeric sequence-specific HMG box to the minor groove of a straight DNA helix (13–16) introduces a sharp bend (on the order of 90\(^\circ\)) in the DNA helix as determined in circular permutation assays (14, 16–18). This is supported by the dispersion of the \(^{31}\)P resonances in the SRY-DNA complex (31).

Exchange of the N- and C-terminal regions of the sequence-specific HMG box of HLEF-1 with those of non-sequence-specific HMG1B showed that the sequence specificity of HLEF-1 is maintained by the N- and C-terminal residues (51). Mutation of
The sequence-specific HMG box of SRY at position V60L (Ile1), M64I (Met5), I68T (Met9), I90M (Ile31), G95R (Gly36), K106I (Lys47) (41, 52–55) as well as the double mutation K298E, K299E (K2R3) and the point mutation L301T (Met5) in the sequence-specific HMG box of LEF-1 (56) affect the DNA-binding. The corresponding residue positions in Sox-4 are given in parentheses. These mutations are mainly located in the N-terminal part of the HMG box. (Fig. 9). Gly36 is located in helix II, and Lys47 is positioned in the loop region between helix II and III. (Fig. 9). Mutations in other parts of the HMG box such as F109S (Phe50) in SRY (55) and V316L (Met20) and Y346S (Phe50) in LEF-1 (56) do not influence the binding properties. However, they can still disrupt the biological function of the protein as is demonstrated by the presence of the F109S (Phe50) mutation in SRY of sex-reversed XY female (55). Of special interest is mutation M64I (Met5) in SRY, which shows an almost normal DNA-binding affinity, but decreases the DNA bending with 20° (55).

The side chain of Ile68 (Met9) in the N-terminal region of the sequence-specific HMG box of SRY intercalates partially from the minor groove side between the two central AT base pairs of its d(AACAATCA)-d(TGATTGTT) heptamer motif (30, 31). Note that in murine SRY and Sox-4 this interacting Ile is replaced by Met. With this information a model for the SRY-DNA complex was constructed (31). Here, the concave surface of the HMG box of SRY, whose structure was based on the NMR solution structure of the non-sequence-specific HMG1B box, faces the bent DNA with helix I, which is docked in a widened minor groove.

The effect of the mutations located in the N-terminal region
of the HMG box of SRY (41, 52–55) and LEF-1 (56) on the DNA binding (see also above and Fig. 10) indicate that the N-terminal residues of the HMG box interact with the DNA. On the other hand the results of methylolation- and diethyl-pyrocarbonate carboxylation interference footprinting and TIC(A1) nucleotide substitutions (13–16) show that the HMG box interacts in the minor groove with the first 6 base pairs of the d(AACAAAG)d(TCTTGGT) consensus sequence.

Based on the notion that the N terminus of the HMG box interacts with the first 6 base pairs of the d(AACAAAG)-d(TCTTGGT) binding sequence and the finding that 11e88 (Met88) of SRY intercalates between the two central AT base pairs of the d(AACATCA)-d(TGTATGG) heptamer motif (30, 31), we add the proposal that in the HMG box-DNA complex the N terminus of the HMG box points in the direction of the 5’ AT base pair of the d(AACAAAG)-d(TCTTGGT) consensus binding sequence. Considering the sequence homology between the HMG boxes of SRY and Sox-4 a similar model for the Sox-4 HMG box-DNA complex might be proposed. However, a definitive model awaits the experimental determination of the structure of the complex of Sox-4 and DNA.

Note Added in Proof—After this manuscript was submitted, the structure of the DNA complex of SRY (57) and LEF-1 (58) was reported.
