The crystal structure of the Sox4 HMG domain–DNA complex suggests a mechanism for positional interdependence in DNA recognition

Ralf JAUCH*1, Calista K. L. NG*†, Kamesh NARASIMHAN*‡ and Prasanna R. KOLATKAR*‡†

*Laboratory for Structural Biochemistry, Genome Institute of Singapore, 60 Biopolis St, Singapore 138672, †School of Biological Sciences, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, and ‡Department of Biological Sciences, National University of Singapore, 14 Science Drive 4, Singapore 117543

It has recently been proposed that the sequence preferences of DNA-binding TFs (transcription factors) can be well described by models that include the positional interdependence of the nucleotides of the target sites. Such binding models allow for multiple motifs to be invoked, such as principal and secondary motifs differing at two or more nucleotide positions. However, the structural mechanisms underlying the accommodation of such variant motifs by TFs remain elusive. In the present study we examine the crystal structure of the HMG (high-mobility group) domain of Sox4 [Sry (sex-determining region on the Y chromosome)-related HMG box 4] bound to DNA. By comparing this structure with previously solved structures of Sox17 and Sox2, we observed subtle conformational differences at the DNA-binding interface. Furthermore, using quantitative electrophoretic mobility-shift assays we validated the positional interdependence of two nucleotides and the presence of a secondary Sox motif in the affinity landscape of Sox4. These results suggest that a concerted rearrangement of two interface amino acids enables Sox4 to accommodate primary and secondary motifs. The structural adaptations lead to altered dinucleotide preferences that mutually reinforce each other. These analyses underline the complexity of the DNA recognition by TFs and provide an experimental validation for the conceptual framework of positional interdependence and secondary binding motifs.

Key words: DNA binding, high-mobility group (HMG), positional interdependence, protein–DNA recognition, Sox, transcription factor.

INTRODUCTION

The evolutionarily ancient [1] Sry(sex-determining region on the Y chromosome)-box-containing TFs (transcription factors) {Sox [Sry-related HMG (high-mobility group) box]} were named after a highly conserved ∼ 80 amino acid HMG (high-mobility group) DNA-binding domain first identified in the testis-determining factor Sry [2]. The human genome encodes 20 Sox proteins, most of which are critical regulators of gene expression programmes in a diverse set of tissues and at various stages of embryonic as well as adult development (reviewed in [3–6]).

The DNA-binding specificities of Sox proteins have been comprehensively interrogated using high-throughput PBM (protein-binding microarray) technologies [7]. These studies reaffirmed previous findings on individual family members [8] that all Sox proteins exhibit highly similar binding preferences to a TTGT core sequence. However, the notion that all Sox proteins bind similar and very short DNA sequences seems to contradict the functional versatility of individual family members. Apparently, different Sox proteins are recruited to distinctive genomic regions by means of their affinity to DNA and regulate non-redundant sets of genes to trigger particular developmental programmes. Three mechanisms can be conceived to underlie the functional uniqueness of individual Sox proteins. First, Sox proteins were found to associate with a range of partner factors, and such interactions may influence the selection of DNA target sites and could modulate the cellular consequences of binding events [9–13]. Secondly, Sox proteins have the peculiar property of binding to the minor groove of DNA and cause a pronounced kink to the double helix [15]. Such Sox-induced structural changes to the DNA may vary for different Sox family members and could thus modulate their regulatory output [16]. Lastly, although the core of the DNA-binding site is shared for all Sox family members, minor variations in nucleotide preference at motif edges could be detected [7]. These differences enabled a hierarchical clustering of the Sox family by virtue of their sequence preferences. Yet, whether these subtle variations within the DNA target sequences alter the binding energy in a substantial manner and influence the genomic occupancy profile remains to be explored.

Structural studies have greatly contributed to unravelling the architecture and dynamics of the HMG domain and mechanisms of DNA recognition by Sox proteins [15,17–24]. Importantly, the kinked DNA conformation seems to depend crucially on highly conserved amino acids and features of the C-terminal tail of the domain [16,20]. Yet, features enabling individual Sox proteins to discriminate between DNA sequences or to induce distinctive DNA conformations have not been discovered. If such features exist, the structural analysis of a larger set of Sox–DNA complexes should reveal their existence.

In the present study, we addressed two fundamental questions: (i) do different Sox proteins kink the DNA in a differential manner; and (ii) are the subtly different DNA motifs suggested by PBM studies supported by different DNA-interaction chemistries? To tackle these problems we chose the HMG domain of Sox4 and Lama1 (laminin, subunit α1) enhancer DNA that was first crystallized in complex with the HMG domain of Sox17 [21], for several reasons. First, Sox4 belongs to the C-group of Sox proteins [25] whose DNA recognition has not yet been structurally investigated. Sox4 and other C-group proteins, Sox11 and Sox12, are critical, partially redundant, regulators in important processes

Abbreviations used: Cy5, indodicarbocyanine; EMSA, electrophoretic mobility-shift assay; HMG, high-mobility group; Lama1, laminin, subunit α1; NSLS, National Synchrotron Light Source; PBM, protein-binding microarray; PWM, position weight matrix; Sry, sex-determining region on the Y chromosome; Sox, Sry-related HMG box; TF, transcription factor.

† Correspondence may be addressed to either of these authors (email jauchr@gis.a-star.edu.sg or kolatkarp@gis.a-star.edu.sg).

The structural co-ordinates reported for the Sox4 HMG domain bound to DNA will appear in the PDB under accession code 3U2B.
such as cardiac and neuronal development [8,26,27] and are thought to be transcriptional triggers of malignancies [28]. Secondly, PBM studies suggest a slightly altered DNA specificity of the C-group as compared with the F-group represented by Sox17 [7]. Thirdly, Sox17 and Sox4 were found to have converse effects on β-catenin/TCF (T-cell factor) signalling and tumour progression in colon carcinomas while sharing cellular interaction partners [29]. This raises the possibility of an antagonistic action of Sox4 and Sox17, possibly through eliciting different effects on jointly targeted DNA sites. To be able to conduct an unbiased comparison of DNA conformation and recognition and to minimize artefacts introduced by altered DNA sequences, we decided to use Lama1 DNA that was first crystallized in complex with Sox17. Indeed, as a result of the present study we discovered subtly different DNA-binding modes between Sox4 and Sox17. On the basis of these structural models, we propose a general mechanism for the accommodation and discrimination between primary and secondary binding sites by Sox proteins by a concerted structural rearrangement at the DNA-contact interface.

MATERIALS AND METHODS

Cloning, protein expression and purification

Sox17 protein was purified as described previously [43]. The region encoding 79 amino acids of the HMG domain of mouse Sox4 was BP cloned (Invitrogen) from a cDNA clone (UniProt code Q06831; IMAGE Consortium code 6822618) into the pDONR221 vector. The resulting pENTR construct was first verified by sequencing and recombined into the pETG40 expression plasmid using Gateway® LR technology (Invitrogen). The destination plasmid was transformed into Escherichia coli BL21(DE3) cells and plated on LB (Luria–Bertani) plates supplemented with 100 μg/ml ampicillin. An overnight innoculum was transferred into 6 litres of 1× Terrific Broth and grown until the $D_{600}$ reached $0.5–0.8$ before inducing with 0.5 mM IPTG (isopropyl-β-D-thiogalactopyranoside) at $18^\circ$C for $18$ h. Cells were harvested, resuspended in lysis buffer (20 mM Hepes, pH 7.0, 1 mM EDTA, 200 mM NaCl and 10 mM 2-mercaptoethanol) and sonicated for 15 min at $4^\circ$C. Cell lysate was cleared by centrifugation at 48000 g for 30 min at $4^\circ$C. Pre-equilibrated amylose beads were added to the filtered lysate and incubated at $4^\circ$C for $1–2$ h. Fusion proteins were eluted with lysis buffer containing 10 mM maltose. The MBP (maltose-binding protein) tag was separated from Sox4 protein by digestion with TEV (tobacco etch virus) protease (1:100, w/w) overnight at $4^\circ$C. The cleaved protein sample was injected into a 6 ml Resource S column (GE Healthcare) connected to an AKTA Express system equilibrated with buffer A (20 mM Hepes, pH 7.0, and 100 mM NaCl) and was eluted by increasing the salt concentration gradually to 1 M over a total of 25 column volumes. Peak fractions were pooled and subjected to gel filtration chromatography using a HiLoad™ 16/60 Superdex™ 75 prep grad column (GE Healthcare) equilibrated with buffer A. Peak fractions were analysed by SDS/13.5% PAGE, pooled and concentrated to approximately 12.7 mg/ml and finally stored at $-80^\circ$C until use. The protein concentration was measured with a Thermo Scientific NanoDrop® ND-1000 spectrophotometer.

Crystallization, data collection and structure solution

PAGE-purified DNA purchased in liquid form from Sigma-Proligo at 1 mM was adjusted to pH 8.0 using Tris/HCl, mixed and annealed by heating to $95^\circ$C and gradual cooling to ambient temperature. The double-stranded DNA and the Sox4 HMG domain were mixed at a 1.2:1 molar ratio and the complex at a concentration of $\sim 320 \mu$M was subjected to crystallization trials. The chromosome-shaped crystal used for data collection was grown using the hanging-drop vapour diffusion technique after combining equal volumes of protein solution and a reservoir buffer containing 100 mM Hepes, pH 7.2, 20 % PEG (poly(ethylene glycol)) 3350, and 50 mM MgCl₂. Crystals were flash-frozen in liquid nitrogen. Data were collected on beamline X29 of the NSLS (National Synchrotron Light Source) and processed using HKL2000 [44] (Table 1). Molecular replacement was performed using PHASER [45] and a poly(alanine) model derived from the Sox17–HMG DNA complex (PDB code 3F27). Model building was initiated using Buccaneer [46] and finalized manually using Coot [47]. Refinement was carried out using phenix.refine [48]. The first refinement round using the Buccaneer model included a simulated annealing with a starting temperature of 10000 K to reduce model bias. TLS (translation/libration/screw) groups were defined using the TLSMD server [49]. The structure of the Sox4 HMG domain bound to DNA has been deposited in the PDB under accession code 3U2B.

EMSAs (electrophoretic mobility-shift assays)

EMSAs were essentially performed as described previously [50] with the following modifications. Binding reactions were set up containing a final concentration of 1 mM Cy5 (indodicarbocyanine)-labelled 16-mer Lama1 DNA with various amounts of unlabelled DNA competitors and 50 nM Sox protein. Reactions tubes were incubated at $4^\circ$C in the dark for 1 h and separated on a 12 % native 1× Tris/glycine (25 mM Tris, pH 8.3, and 192 mM glycine) polyacrylamide gel. Bands were detected using a Typhoon 9140 PhosphorImager (GE

<table>
<thead>
<tr>
<th>Table 1 Crystallographic data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data collection</strong></td>
</tr>
<tr>
<td>Space group</td>
</tr>
<tr>
<td>Cell dimensions (Å)</td>
</tr>
<tr>
<td>c</td>
</tr>
<tr>
<td>Resolution (Å)</td>
</tr>
<tr>
<td>$R_{	ext{free}}$ (%)</td>
</tr>
<tr>
<td>I/σ (I)</td>
</tr>
<tr>
<td>Completeness (%)</td>
</tr>
<tr>
<td>Multiplicity</td>
</tr>
<tr>
<td>Refinement</td>
</tr>
<tr>
<td>No. reflections</td>
</tr>
<tr>
<td>$R_{	ext{work}}/R_{	ext{free}}$ (%)</td>
</tr>
<tr>
<td>No. atoms (without hydrogens)</td>
</tr>
<tr>
<td>Protein</td>
</tr>
<tr>
<td>DNA</td>
</tr>
<tr>
<td>Water</td>
</tr>
<tr>
<td>Mean B-factors (isotopic equivalent)</td>
</tr>
<tr>
<td>Macromolecules</td>
</tr>
<tr>
<td>Water</td>
</tr>
<tr>
<td>Root mean square deviations from ideal</td>
</tr>
<tr>
<td>Bond lengths (Å)</td>
</tr>
<tr>
<td>Bond angles (°)</td>
</tr>
<tr>
<td>Ramachandran analysis</td>
</tr>
<tr>
<td>Favoured (%)</td>
</tr>
<tr>
<td>Additionally allowed (%)</td>
</tr>
<tr>
<td>Disallowed (%)</td>
</tr>
</tbody>
</table>

*Values for the highest-resolution shell are in parentheses. 1 Å = 0.1 nm.*

© The Authors. Journal compilation © 2012 Biochemical Society
Structure of the Sox4 HMG domain

41

Healthcare) and quantified using ImageQuant TL software (GE Healthcare). The intensities of gel bands corresponding to free or TF-bound labelled DNA can be used to calculate the fraction of labelled DNA bound by the TF. Since the total concentration of labelled DNA is known (1 nM), these fractional concentrations can be converted into absolute estimates of the amount of bound and free labelled DNA. The concentration of unlabelled competitor DNA is also known, as is the total TF concentration. We can therefore substitute these known values into the equations that define the dissociation constants in terms of reactant and product concentrations to calculate the $K_d$ values, which can be converted into PWMs (position weight matrices) ([51], and W. Sun, C. Ng, R. Jauch and S. Prabhakar, unpublished work). $K_d$ denotes the dissociation constant for the competitor DNA containing systematic mutations and $K'_d$ the dissociation constant for the labelled DNA. EMSA titrations to estimate absolute $K_d$ values were conducted and the results analysed using equations, as described previously [50].

RESULTS AND DISCUSSION

Overall structure and DNA recognition

The Sox4 HMG domain contains characteristic sequence signatures found in all 20 mammalian Sox paralogues and, in particular, DNA contact residues are highly conserved (Figures 1A and 1B). The protein resembles a prototypical HMG scaffold with three helices forming an L-shaped fold (Figure 1C). Previous biophysical studies divided the HMG domain into two subdomains comprising a minor wing encompassing helices H1 and H2 and a major wing formed by the C-terminal part of helix H3 and the extended N-terminus that runs antiparallel to this helix [30]. In accordance with NMR studies of the Sox4 HMG in the absence of DNA [23], the C-terminal end of the Sox4 HMG exhibits a high degree of structural flexibility as reflected by high temperature factors and noisy electron density. Owing to this structural disorder, the last three amino acids of the 79 amino acid Sox4 HMG domain under study could not be traced and were excluded from the model. Nevertheless, the C-terminal part of the protein is crucial for DNA binding and the formation of the kink [16,21].

A key question of this project was to compare DNA architectures induced by variant members of the Sox family. To this end, we crystallized the Sox4 HMG in complex with Lama1 DNA that we had earlier crystallized in complex with Sox17. Lama1 is a cognate binding site of Sox17 during endodermal differentiation [31], whereas Sox4 is not known to bind and regulate this gene in vivo. It is possible that Sox proteins can only deform their cognate binding sites in a manner compatible with the assembly of regulation-competent complexes. We therefore

Figure 1  Overall structure and DNA binding by the Sox4 HMG

(A) Conservation score plotted against the sequence of the mouse Sox4 HMG domain used in the present study. The score was calculated using the ScoreCons server (http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/valdar/scorecons_server.pl) [52] using an alignment of all 20 mouse Sox proteins. Residues were numbered according to HMG conventions for easier comparison [3]. DNA-contact residues are marked with open circles, the three helices are marked with red bars and the FM wedge with an orange bar. (B) Schematic drawing of the Lama1 DNA sequence used in the present study. Prominent protein–DNA interactions are indicated. Dashed lines indicate base-specific interactions and dotted lines indicate phosphate or deoxyribose interactions. Red fonts highlight alternative interactions seen for Arg18 and Asn30 in the Sox17 structure [21]. Red numbers label the base-pairs of the core of the Sox17 motif (1–7) discussed in the text. (C) Cartoon drawing depicting the overall fold of the Sox4 HMG–DNA complex highlighting the strong overall bend of the helix. Positions of N- and C-termini are indicated, helices are numbered H1–H3, and 5′ and 3′ DNA ends are labelled for both strand A and strand B. Arg18 and Asn30 are shown as sticks. The inset depicts a magnification of Arg18 and Asn30 rotated by 90° around the x-axis, and the 2Fo −Fc electron density is shown contoured at the 1.0σ level.
carefully analysed the architecture of Sox4- and Sox17-bound Lama1 DNA using Curves+ [32]. Both Sox4 and Sox17 were found to deform the Lama1 DNA by binding to the minor groove in a virtually identical manner with overall bends of the helical axis of 65.8° for Sox4 and 68.9° for Sox17. Likewise, intra- and inter-base-pair parameters were also highly similar. We therefore conclude that the differential deformation of DNA makes no major contribution when Sox4 and Sox17 proteins select their target genes. Accordingly, the majority of DNA contacts are identical for both proteins (Figures 1A and 1B). As an exception, Arg18 and Asn30 show a different pattern of hydrogen bonding in Sox4 as compared with Sox17. The side-chain conformations of Arg18 and Asn30 are well supported by the 2Fo − Fo electron density and were determined after automated model building using Buccaneer and simulated annealing refinement (Figure 1C). Model bias is unlikely since the poly(alanine) starting model used for molecular replacement was derived from Sox17 co-ordinates, which exhibit a different side-chain conformation in this region. In Sox4, Arg18 and Asn30 position their side chains to contact base-pairs 1 and 2 of the core motif. Since those nucleotides are positioned at the edge of the Sox motif that was found to exhibit some variability among Sox paralogues, we decided to study the DNA-binding specificity of Sox4 in greater detail.

Sox4 binds primary and secondary DNA motifs in a non-additive manner

We have recently developed a method to measure relative dissociation constants (Kd) in a competitive EMSA (see the Materials and methods section). This method enables the quantification of the affinity ratio between a fluorescently labelled reporter DNA as compared with an unlabelled competitor. By measuring Kd ratios for a DNA library where all bound nucleotides are systematically mutated into the three other possible nucleotides, a position-specific affinity matrix can be constructed (Figures 2A–2C). We conducted this experiment using a DNA library derived from the core of the Lama1 DNA. As expected, we found that the central sequence TCG′G′TT is critical for high-affinity binding. However, positions 1 and 2 of the Sox4-binding site exhibit surprisingly low information contents after converting the data into a PWM (Figure 2C). In contrast, PBM studies reported a rather strong preference for CT at these positions (Figure 2D). However, the PBM method measures a large number of sites and is capable of detecting less preferred, but significantly enriched, secondary motifs (Figure 2E). For Sox4 and the other SoxC group proteins Sox11 and Sox12, the secondary motif contains an AA at the beginning of the motif.

Two considerations led us to wonder whether our EMSA method captures the true sequence preference of Sox4. First, the affinity matrix was recorded on the basis of the Lama1 element containing a starting TA which constitutes neither an ideal primary nor an ideal secondary motif. Secondly, this approach assumes positional independence. That is, nucleotides are changed one by one, assuming that mutations at site 1 have no effect on the nucleotide preference at site 2. As a consequence, positional interdependence, in this case between site 1 and site 2, could explain the discrepancy between the EMSA and the PBM-derived PWMs. To test this hypothesis, we designed an additional series of mutated Lama1 DNA elements and tested for the affinity ratio compared with wild-type Lama1 DNA (Figure 2G). Indeed, the CT element bound Sox4 more strongly than the TA-containing Lama1 DNA. Overall, the T2 was preferred in instances when C1 was present, underlining the reliability of the primary PBM motif. As expected, the secondary A1A2 motif has an affinity lower than the primary C1T2 motif. However, when an A1 is present the preference at position 2 changes from a T to an A. Taken together, these experiments support the presence of primary and secondary motifs in the affinity landscape of Sox4. Furthermore, there seems to be a significant positional interdependence at binding positions 1 and 2: a C1 favours a T at position 2, whereas A1 enforces an A at position 2.

Sox2, Sox17 and Sox4 adopt different conformations at the DNA-interaction interface

Next we carefully inspected the environment in the structures of Sox4 and Sox17 around base-pairs 1 and 2 exhibiting positional interdependence in EMSA experiments. Intriguingly, the only notable conformational differences at the DNA-contact interface between Sox4 and Sox17 is seen for Arg18 and Asn30, which contact base-pairs 1 and 2 (Figures 1B and 1D and Figure 3A). Specifically, in Sox4, Arg18 contacts the minor groove O2 carbonyl of T2 and Asn30 contacts the minor groove N3 of A1′ (numbering according to base-pairs 1–7 of the core Sox4 motif shown in Figures 2C–2E; N′ refers to the complementary nucleotide). This conformational arrangement is further stabilized through direct side chain contacts between Arg18 and Asn30. In contrast, in Sox17, the side chains of both amino acids are slightly re-oriented, causing Arg18 to contact A3′ and Asn30′ to contact T2′, whereas the intramolecular Arg18′–Asn30′ interaction is lost.

To further investigate these subtle, but potentially consequential, conformational differences, we inspected the crystal structure of Sox2 (PDB code 1GTO), which is bound to a sequence closely resembling a primary motif beginning with C-G1′;T-A2 base-pairs (Figure 3B). Importantly, the binding pattern of the Arg18′–Asn30′ dipeptide in Sox2 is reminiscent to the situation seen in Sox4. Given that Sox2 is bound to a perfect primary sequence (C1T2), it seems likely that the Arg18′–Asn30′ conformation seen for Sox2 and Sox4 represents the interaction mode characteristic for the higher-affinity primary Sox motifs. Apparently, the primary scenario is facilitated by two consecutive purine N3 acceptors at motif positions 1 and 2 within the reverse strand. This poses the question of whether the conformation modelled for Sox17 represents a conformational switch of Arg18′/Asn30′ reminiscent of Sox4 bound to a secondary motif (Figure 3C). If so, Arg18′ and Asn30′ may undergo an inter-connected rearrangement to accommodate the different chemical environments encountered on primary and secondary motifs (Figure 4).

To further assess this possibility, we converted the Sox17-bound Lama1 sequence into a secondary motif and inspected the potential binding pattern occurring on such an element. In this modelled secondary motif a carbonyl is introduced by T1′ at position 1, whereas the primary motif as well as the Lama1 site would provide a purine N3 as an H-bond acceptor. Both purine N3 as well as the pyrimidine O2 carbonyl can act as H-bond acceptors. However, a comprehensive analysis of protein–DNA complexes revealed that the H-bonding geometries of amino acids interacting with purines and pyrimidines are markedly different [33]. For example, H-bonds around purine N3 were found to be more compact and shifted by a different amplitude away from the base plane and exhibit different binding angles. Although subtle, such differences have enabled the design of sequence-specific drugs that target the minor groove [34]. Thus replacing the purine N3 with a pyrimidine O2 at position 1 would probably disturb the H-bonding geometry and would be incompatible with the Asn30′ conformation seen for Sox4 and Sox2. As a consequence, a conformation seen in Sox17 where the side chain flips and the T2′ carbonyl is contacted instead would be more favourable. Indeed, we observed favourable H-bond distances when Sox17 was modelled on the secondary motif (Figure 3D and Figure 4).
Figure 2  Validation of positional interdependence in DNA-motif recognition by Sox4

Representative competition EMSA experiments performed using the Sox4 HMG (A) and Sox17 HMG (B). Proteins were mixed with 1 nM Cy5–DNA_inf, and systematically mutated unlabelled competitor DNA elements. Every position of the 7 bp core of the Lama1 core DNA element was mutated into all possible alternative base-pairs, producing a library of 21 elements. Lanes 1–3 contain no competitor and the indicated concentration of protein. WT indicates wild-type competitor DNA, and the letters in lane 5–25 (A) or lanes 4–24 (B) and subscripts denote the mutant nucleotide and its position within the core motif. The band intensity of the TF complexed with labelled DNA is a monotonic function of the dissociation constant $K_{d\text{mut}}$ of the unlabelled competitor DNA fragment. A weak intensity indicates a low $K_{d\text{mut}}$ and a high affinity. This intensity can therefore be quantified and used to estimate $K_{\text{mut}}/K_{d}'$, where $K_{d}'$ is the $K_{d}$ for the labelled reporter DNA (see the Materials and methods section). (C) PWM derived from $K_{\text{mut}}/K_{d}'$ ratios estimated from the competitive EMSA in (A) were visualized using the R tool seqLogo. Primary (D) and secondary (E) Sox4 motifs as determined using the PBM method [7]. PBM PWMs were downloaded from the UniProbe database (http://the_brain.bwh.harvard.edu/uniprobe/) and visualized using seqLogo. (F) PWM derived for the Sox17 HMG by EMSAs. To test for positional interdependence, double-mutant competitor DNA elements were analysed using Sox4 HMG (G) and Sox17 HMG (H). Positions 1 and 2 were modified to resemble primary (CT) and secondary (AA) motifs. Note that the Lama1 element has a TA at these positions constituting neither a perfect primary nor secondary element. Base-pairs at position 2 were changed into all of the three possible nucleotides. The y-axis shows the log2-transformed $K_{d\text{mut}}/K_{d}'$ of the mutant DNA and the Lama1 DNA as means ± S.D. Values above zero correspond to an increased $K_{d\text{mut}}$ (decreased affinity) and values below zero to a decreased $K_{d\text{mut}}$ (increased affinity). The primary CT element has an even higher affinity than the WT element to the Sox4 HMG, whereas the drop in affinity by an A at position 1 is partially rescued by a second A at position 2. However, the single-nucleotide exchange from CT to AT results in a pronounced drop in affinity, suggesting positional interdependence. The positional interdependence is not observed for Sox17 (H), and the TA dinucleotide of the Lama1 DNA, which is a cognate Sox17 target site, is indeed the highest-affinity binding site.

© The Authors Journal compilation © 2012 Biochemical Society
Figure 3  Structural differences at the DNA-binding interface in Sox2, Sox4 and Sox17

Comparison of the binding to base-pair positions 1 and 2 as seen in (A) Sox4 (blue), (B) Sox2 (PDB code 1GT0; green), (C) Sox17 (PDB code 3F27; light pink) and (D) Sox17 on a model of a secondary motif (light blue). The models correspond to a view rotated around the x-axis as compared with Figure 1(B) to visualize the complex from the top and are shown in wall-eyed stereo. The side chains of Arg18, His29, Asn30 and Ser34 and the base-pairs −1 to 3 (see Figures 1B and 2C) are shown as sticks. Note that Ser34 occupies the same position and binding characteristics in all structures, whereas Arg18 and Asn30 show a different H-bonding geometry in Sox2 and Sox4 when compared with Sox17. In addition, His29 exhibits a flipped-out conformation in Sox2.

To facilitate the rearrangement from a primary to a secondary binding mode, Arg18 has to give way for Asn30 to now contact the N3 of A3′ instead of the position 2′. Are these structural rearrangements consistent with the positional interdependence? Apparently, Sox4 favours a T2′ when a T1′ is present (Figure 2G). It is conceivable that a purine with an N3 instead of a pyrimidine’s O3 as at the 2′ position would not install the energetically most favourable H-bond geometry for the conformation seen in Sox17. Likewise, if the pyrimidine O3 was provided by a C2′ instead of a T2′, the complementary G2 could disturb the H-bonding geometry by inserting an additional amino group into the minor groove which could, for example, interfere with the binding of
The side chains of Arg₁₈ and Asn₃₀ is triggered to accommodate the variant chemical environments. Although subtle, this rearrangement could rationalize the positional interdependence for changes from an N₃ to an O₂ as the H-bond acceptor at position 1 of binding both primary and secondary motifs with high affinity.

energies can influence sequence preferences and shape the binding landscape of TFs [35].

addition, although the present model emphasizes the H-bonding geometry as a key contributor to the Sox4 interaction with primary and secondary DNA motifs, other factors such as base stacking geometries, explaining the positional interdependence.

Interestingly, Sox17 appears to have a stronger preference for an A-T₂ basepair using additive binding models (Figure 2F). Thus Sox17, unlike Sox4, may be unable to discriminate between primary and secondary motifs. Indeed, a C1T₂-type motif has not been found for Sox17 in PBM studies (http://the_brain.bwh.harvard.edu/uniprobe/). In fact, according to this dataset, the distinction between C1T₂- and A1A₂-type motifs appears to be characteristic for SoxC-type HMG domains and may not apply to the whole Sox family. Additional structures of Sox HMG domains in complex with motif variants, in conjunction with molecular dynamics simulations, are needed to further explore differences between Sox family members and the discrimination between motif variants by HMG domains. In addition, although the present model emphasizes the H-bonding geometry as a key contributor to the Sox4 interaction with primary and secondary DNA motifs, other factors such as base stacking energies can influence sequence preferences and shape the binding landscape of TFs [35].

Conclusion

It has long been recognized that protein–DNA recognition is a highly complex process and does not follow a simple recognition code [36,37]. That is, for most TF families the DNA target sequence cannot be straightforwardly predicted from the amino acid sequence. The goal of many high-throughput TF–DNA-interaction studies, such as HITS-FLIP (high-throughput sequencing fluorescent ligand interaction profile), PBM and HTSELEX (high-throughput systematic evolution of ligands by exponential enrichment), and microfluidic experiments, such as MITOMI (mechanically induced trapping of molecular interactions), is to arrive at a minimal model that would adequately describe the DNA-interaction landscape for TFs [7,38–40]. Accurate models for the TF–DNA interactions are critical to predict the genomic binding of a TF in cells and to understand gene regulatory networks. In addition, these insights can help to re-engineer TF proteins with novel specificities. The issue has been complicated by high-throughput studies which contend that simple additive PWM-based models do no suffice to comprehensively capture the affinity landscape of TFs [7,40]. Rather, individual TFs seem capable of recognizing multiple motifs with positionally interdependent base substitutions. However, the prevalence of such secondary motifs in available datasets is under some debate and method development to analyse TF binding data is far from complete [41,42]. Nevertheless, the structural pre-requisite for binding to disparate DNA motifs is that a TF must be sufficiently flexible to adapt to alternative chemical environments and to engage in new intermolecular contacts.

In the present study we explored the positional interdependence and the structural basis for the discrimination between primary and secondary motifs by SoxC group TFs. By definition, a primary motif describes the highest-affinity consensus motif of a TF. In contrast, a secondary motif describes a population of lower-affinity binding sites that significantly deviates by more than 1 bp from the primary consensus. Additive PWM-based binding models are not expected to be able to fully describe the affinity landscape of a protein if a noticeable secondary binding mode is present. We propose a model that links recognition of primary versus secondary DNA motifs with the structural
rearrangements at the DNA-binding interface of SoxC proteins. In the light of the present study, we propose that the SoxC proteins are capable of high-affinity binding to primary as well as to secondary binding sites by the concerted re-adjustment of the H-bonding pattern of two amino acid side chains. High-resolution structures of primary and secondary binding sites are needed to put this model to a rigorous test. Although the positional interdependence appears to be significant, the resultant affinity gain is rather moderate, and it remains to be investigated how these differences affect the targeting of genomic loci. It is possible that primary and secondary motifs represent populations of high- and medium-affinity binding that are sensitive to even slightly different concentration thresholds of a TF. A TF’s ability to be inherently flexible and to structurally adapt in order to bind primary and secondary motifs with graduated affinity windows may therefore expand its repertoire of binding sites. It will be interesting to explore whether primary and secondary motifs of Sox4 can be functionally set apart for carrying out distinctive regulatory roles.

AUTHOR CONTRIBUTION
Ralf Jauch conceived and designed the study, collected and assembled data, wrote the paper and gave final approval of the paper. Calista Ng collected, analysed and interpreted data and gave final approval of the paper. Prasanna Kolatkar obtained financial support, provided administrative support and gave final approval of the paper. Kamesh Narasimhan analysed and interpreted data and wrote the paper. Ralf Jauch conceived and designed the study, collected and assembled data, wrote the paper and gave final approval of the paper. Ng, C. K., Li, N., Chee, S., Prabhakar, S., Kolatkar, P. R. and Jauch, R. (2012) Deciphering the Sox-Oct partner code by quantitative cooperativity measurements. Nucleic Acids Res., in the press.

ACKNOWLEDGEMENTS
Bob Robinson [Institute of Molecular and Cell Biology, Singapore Agency for Science, Technology and Research (I*STAR, Singapore) generously provided access to crystallization and X-ray diffraction equipment. We thank Howard Robinson for data collection and processing at beamline X29 of the NSLS, and Shyam Prabhakar and Wenjie Sun for sharing analysis methods and critical comments on the paper prior to publication.

FUNDING
This work was supported by the I*STAR. The NSLS is supported by the Office of Basic Energy Sciences, Office of Science, U.S. Department of Energy.

REFERENCES
17 R. Jauch and others


