doi:10.1016/j.jmb.2006.04.059 J. Mol. Biol. (2006) 360, 90–104

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Structure of a Complex of Tandem HMG Boxes and DNA

Katherine Stott, George S. F. Tang, Keng-Boon Lee and Jean O. Thomas*

Department of Biochemistry
University of Cambridge
80 Tennis Court Road
Cambridge CB2 1GA, UK

*Corresponding author

The high-mobility group protein HMGB1 contains two tandem DNA-binding HMG box domains, A and B, linked by a short flexible linker that allows the two domains to behave independently in the free protein. There is no structural information on how the linked domains and linker behave when bound to DNA, mainly due to the lack of any DNA-sequence preference of HMGB1. We report the structure determination, by NMR spectroscopy, of a well-defined complex of two tandem HMG boxes bound to a 16 bp oligonucleotide. The protein is an engineered version of the AB di-domain of HMGB1, in which the A box has been replaced by the HMG box of the sequence-specific transcription factor SRY, to give SRY.B. In the SRY.B/DNA complex, both HMG boxes bind in the minor groove and contribute to the overall DNA bending by intercalation of bulky hydrophobic residues between base-pairs; the bends reinforce each other, and the basic linker lies partly in the minor groove. As well as being the first structure of an HMG-box di-domain bound to DNA, this provides the first structure of the B domain of HMGB1 bound to DNA.

Keywords: HMG box; HMGB1; NMR spectroscopy; protein–DNA complex; SRY

Introduction

HMGB1 is an abundant vertebrate nuclear protein that binds to DNA in the minor groove without sequence specificity, resulting in DNA bending, and which binds preferentially to distorted DNA. These properties contribute to the ''architectural'' role of HMGB1 in the formation of nucleoprotein complexes.[1] The structural feature responsible for the DNA-binding properties is the HMG box domain, of which there are two in tandem in HMGB1; the boxes are linked through a basic region to a long acidic C-terminal tail (30 consecutive glutamic/aspartic acid residues). Although the binding of an HMG box to DNA is now reasonably well understood (see below), how two boxes and the intervening linker are accommodated, and the role and disposition of the acidic tail, are unclear.

Many nuclear HMG-box-containing DNA-binding proteins have been identified. Some, like HMGB1, are relatively abundant and bind without sequence specificity (e.g. the single-HMG-box proteins yeast Nhp6a,b and *Drosophila* HMG-D). Others are much less abundant sequence-specific transcription factors also with a single HMG box, such as SRY (sex-determining region on the Y chromosome),[2] LEF-1 (lymphocyte enhancer factor 1)[3] and the related TCF-1 (T-cell receptor-α factor 1).[4] Yet others are more recently discovered components of chromatin remodeling complexes.[1,5]

Many proteins in addition to HMGB1 and HMGB2 contain more than one copy of the HMG-box motif, e.g. the mitochondrial factors mtTF1[6] and ABF2,[7] and *Drosophila* DSP1.[8] UBF (upstream binding factor) involved in RNA polymerase I transcription contains four to six boxes, depending on species.[9,10] In general, proteins with more than one HMG box bind with no apparent sequence specificity or, at best, loose specificity (e.g. in the case of UBF[9]).

The solution structures of the free A and B HMG boxes of HMGB1 have been determined by NMR spectroscopy.[11–13] The HMG box motif is L-shaped

K.S. and G.S.F.T. contributed equally to this work.

Abbreviations used: DQF, double quantum filtered; HMG, high-mobility group; HSQC, heteronuclear single quantum coherence; NOE, nuclear Overhauser effect; NOESY, NOE spectroscopy; TOCSY, total correlation spectroscopy; r.m.s.d., root-mean-square deviation; UBF, upstream binding factor.

E-mail address of the corresponding author: jot1@bioc.cam.ac.uk

and contains a short arm comprising helices I and II, and a long arm comprising helix III and the N-terminal extended strand. Although the A and B boxes possess the same global fold, they differ in detail, predominantly in the orientation of helices I and II with respect to the rest of the molecule, and in the length of the loop between these helices, which is longer in the A box.[1,14] The structures of the HMG boxes of HMG-D,[15] Sox-4[16] and Nhp6a[17] have been determined by NMR spectroscopy and all are very similar to the structures of the A and B boxes from HMGB1, resembling most closely the B box.[1]

Structures of single HMG boxes bound to DNA have been determined for a number of proteins, both sequence-specific and non-sequence-specific. The structures of the LEF-1 HMG box bound to a 15 bp oligonucleotide,[18] of the SRY box (in shorter and longer forms) in complex with 8 bp[19] and 14 bp DNA,[20] and of Nhp6a in complex with a 15-mer,[17,21] were determined by NMR spectroscopy; the structures of the A box from rat HMGB1 bound to a cisplatin-modified 16-mer[22] and of the HMG-D box bound to a decamer[23] were solved by X-ray crystallography. The complexes have revealed that the basis of the DNA distortion induced by binding of HMG boxes in the minor groove is intercalation of bulky hydrophobic amino acid side-chains between successive base-pairs (first shown for the SRY HMG box),[24] which partly unwinds the DNA and bends it towards the major groove.[1]

As yet, no structure has been determined of a complex with DNA of two tandem HMG boxes, from HMGB1 or any other multiple-HMG-box protein (although models have been proposed for HMGB1[25] and UBF[26]), or indeed of a complex of box B of HMGB1 with DNA. The AB di-domain from HMGB1 free in solution has independent (non-interacting) A and B boxes; evidently the connecting linker allows free rotation.[27] When the di-domain binds to DNA, whether one bound HMG box acts to reinforce or counter the bending of the other is not known. The former is likely to be the case, since the AB di-domain bends and supercoils DNA more effectively than the single domains;[27] however, the effect is likely to be due, at least partly, to increased DNA affinity, through additional contacts in the di-domain. Also unknown for tandem HMG boxes bound to DNA is the path of the intervening basic linker or its contribution to binding, although some information on the path of basic extensions to the HMG box has come from the structures of the LEF-1,[18] SRY[20] and Nhp6a[17,21] HMG boxes bound to DNA. These show that the C-terminal extension to the LEF-1 box and the N-terminal extension in Nhp6a bind in the major groove, whereas the C-terminal extension to the SRY box binds in the minor groove. (The N and C termini of HMG boxes are spatially close in the molecule, exiting the boxes in roughly the same place.) In the case of LEF-1, the interaction of the basic extension stabilizes the DNA bend introduced by intercalation, by facilitating compression of the major groove through charge neutralization.[28]

The major obstacle to structure determination of a complex of tandem HMG boxes with DNA is the lack of sequence specificity of nearly all proteins with multiple HMG boxes, which precludes the generation of a well-defined, stable complex. We have circumvented this problem by replacing the A box of the AB di-domain of HMGB1 with the sequence-specific HMG box of SRY, giving a hybrid protein, SRY.B, that gives a well-defined complex with DNA. We describe here the NMR solution structure of the two tandem HMG boxes of SRY.B bound to a linear 16 bp oligonucleotide containing the consensus SRY binding site. This shows intercalation from both of the boxes, and in-phase bending, and reveals the interaction of the linker with the minor groove.

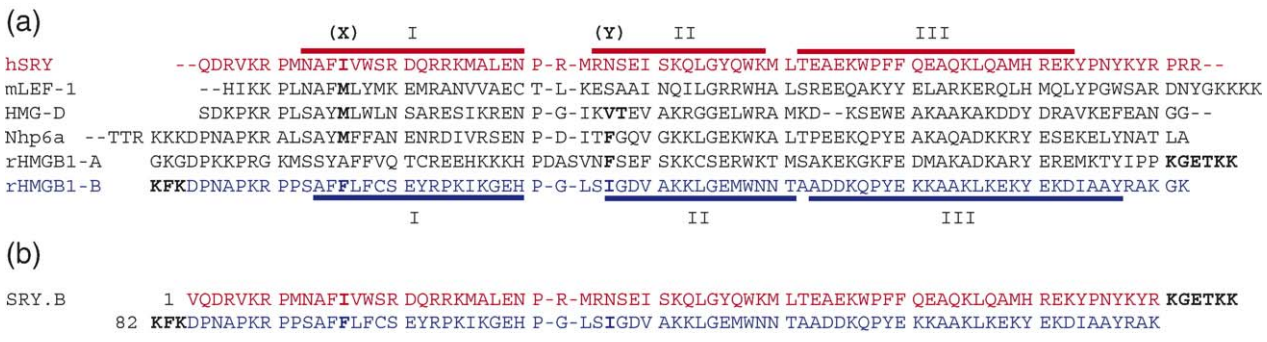

Figure 1. (a) Sequence alignment of HMG-box proteins. Prefixes r, h and m signify rat, human and mouse, respectively. Residues that intercalate at positions X and Y are shown in boldface. SRY, LEF-1 and the A box from HMGB1 have only one intercalating residue (the residue in the other intercalating position (Y, Y and X, respectively) is in each case too small for this purpose), and HMG-D has two at site Y. The α helices (I, II and III) in human SRY and the B-domain of rat HMGB1 are shown by red and blue bars, above and below the sequences, respectively. Residues 81–89 in HMGB1 (bottom two lines) correspond to the basic linker and are shown in bold. (b) Sequence of the SRY.B hybrid protein. SRY.B comprises residues 55–129 of human SRY (shown in red), followed by residues 81–164 of rat HMGB1, in which residues 81–89 (black) and 90–164 (blue) correspond to the basic linker and HMG box B, respectively.