Crystal Structure of the Msx-1 Homeodomain/DNA Complex[†,‡]

Stacy Hovde,[§] Cory Abate-Shen,[||] and James H. Geiger*[,§]

Michigan State University Chemistry Department, East Lansing Michigan 48824, Center for Biotechnology and Medicine, Piscataway New Jersey 08854, and Department of Neuroscience and Cell Biology, University of Medicine and Dentistry of New Jersey-Robert Wood Johnson Medical School, Piscataway, New Jersey 08854

Received April 20, 2001; Revised Manuscript Received August 14, 2001

ABSTRACT: The Msx-1 homeodomain protein plays a crucial role in craniofacial, limb, and nervous system development. Homeodomain DNA-binding domains are comprised of 60 amino acids that show a high degree of evolutionary conservation. We have determined the structure of the Msx-1 homeodomain complexed to DNA at 2.2 Å resolution. The structure has an unusually well-ordered N-terminal arm with a unique trajectory across the minor groove of the DNA. DNA specificity conferred by bases flanking the core TAAT sequence is explained by well ordered water-mediated interactions at Q50. Most interactions seen at the TAAT sequence are typical of the interactions seen in other homeodomain structures. Comparison of the Msx-1−HD structure to all other high resolution HD−DNA complex structures indicate a remarkably well-conserved sphere of hydration between the DNA and protein in these complexes.

Homeobox genes code for a large superfamily of regulatory proteins. The two defining characteristics of these proteins are that they contain a 60 amino acid homeodomain that apparently confers sequence specific DNA binding activity to all of these genes, and that they all appear to be involved in transcriptional regulation (*1*−*3*). Initially identified as the master control genes of *Drosophila* development, they have subsequently been identified in all animal species as well as plants and fungi (*4*, *5*). Several types of deformities and cancers can be attributed to defects in homeoproteins (*1*, *6*−*15*).

The homeobox itself codes for a sequence specific DNA binding domain distantly related to prokaryotic helix-turn-helix domains. It consists of three helices that form a core structure and an N-terminal arm that extends outward from this core. Several homeodomain structures have been solved both by NMR and by X-ray diffraction and most of these structures are in complex with DNA (*16*−*36*). The overall fold of the homeodomain is very similar in these structures, as is the way in which they interact with the DNA. Though most homeodomains bind DNA exclusively as monomers, some also bind as heterodimers, the MATa1−MAT α2 heterodimer (*17*, *34*) and the HoxB1−Pbx-1 heterodimer (*24*) being examples. Other families of homeodomain proteins contain a second DNA binding domain within their sequence. These include the *POU* and *Paired* classes of homeobox genes.

The three *Msx* genes, *Msx-1*, *Msx-2*, and *Msx-3*, form a subfamily of homeobox genes and play essential roles in vertebrate development (*37*, *38*). They are mainly expressed and appear to play important roles during early, middle, and late stages of craniofacial, limb and nervous system development (*39*). Interestingly, in general *Msx* gene expression is restricted to cells that are either proliferating or dying as opposed to cells that are differentiating, implicating them in the regulation of these processes (*1*).

Though very few genes have been identified that are regulated by the Msx proteins in organisms, both in vivo and in vitro studies show that these proteins act as potent transcriptional repressors (*37*, *40*−*44*). Surprisingly, this transcriptional repression activity is independent of Msx DNA binding sites. Instead, repression activity appears to be transduced via protein−protein interactions with basal machinery factors and with other homeodomain proteins. In fact Msx-1 binds tightly and specifically to the TATA[1] binding protein (TBP) while Msx-2 interacts specifically with TFIIF (*42*, *44*). Mutations that abrogate these interactions correlate with loss of repression activity and are localized to the N-terminal arm of the homeodomain. Msx-1 also directly interacts with members of three other homeodomain protein families, Dlx (Dlx 2 and Dlx 5), LIM(Lhx2), and Pax (Pax3) (*43*, *45*, *46*). In each case, the interactions are localized to the homeodomain and in each case the interaction abrogates DNA binding of both proteins, and neutralizes the transcriptional effects of each. It thus appears that these interactions may lead to functional antagonism in vivo in tissues where there is coexpression of Msx-1 and these other proteins.

To further understand the details of Msx-1 function, we have determined the three-dimensional structure of the Msx-1

[†] This work was supported by MCB/NIH (Grant 9982536).

[‡] Coordinates have been deposited in the Protein Data Bank (accession number 1IG7).

* To whom correspondence should be addressed. E-mail: geiger@cem.msu.edu.

[§] Michigan State University Chemistry Department.

[||] Center for Biotechnology and Medicine and Department of Neuroscience and Cell Biology.

[1] Abbreviations: HD, homeodomain; TBP, TATA binding protein; TFIIF, transcription factor IIF; SDS, sodium dodecyl sulfate; PEG, polyethylene glycol; Eve, even-skipped, Cyt (C), cytosine; Ade (A), adenine; Gua (G), guanine; Thy (T), thymine; Wat, water; standard one letter abbreviation codes were used for the amino acids.

10.1021/bi0108148 CCC: $20.00
Published on Web 09/14/2001

Table 1: Data Collection Statistics[a]

wavelength (Å)	1.54
resolution range (Å)	40.0−2.15
reflections recorded	11219
completeness (last shell)	98.9%(99.8)
$R\ (I)_{merge}$* (last shell)	6.2%(29.8)

[a] $R_{merge} = \sum_i |I_I - \langle I \rangle| / \sum \langle I \rangle$, where I_I is an individual intensity measurement, and $\langle I \rangle$ is the average intensity for this reflection with summation over all the data.

homeodomain bound to its cognate DNA site. This structure elucidates the details of Msx-1−DNA binding specificity and allows us to make some further conclusions regarding homeodomain−DNA recognition.

MATERIALS AND METHODS

Crystallization. Purification of the Msx-1 (Hox 7.1) homeodomain protein has been described previously (*47*). Msx-1 was overexpressed in *Escherichia coli* and purified to homogeneity by nickel-affinity chromatography. The purity of the protein was determined by SDS−polyacrylamide gel electrophoresis. Purification of the DNA has been described previously (*48*). The final DNA concentration was approximately 2 mM. The DNA and protein were combined in a 1.2:1 ratio and final protein concentration was 10 mg/mL.

The Msx-1−DNA complex was crystallized using the hanging drop vapor diffusion method. Crystals were grown by mixing equal volumes of the protein−DNA solution and a precipitating solution containing 12% PEG 4000 and 0.1 M sodium acetate, pH 4.6. The DNA that yielded the largest crystals was 5′-TGTCACTAATTGAAGG-3′ and contained an overhang T on each end that promotes stacking of the complex in the crystal. The crystals grew in 4 weeks to the maximum dimensions of 0.4 × 0.2 × 0.2 mm. The crystals were cryoprotected with 30% glycerol in the precipitating agent solution. The Msx-1−DNA complex crystallizes in the $P2_12_12_1$ space group with unit cell dimensions of $a = 33.66$, $b = 60.96$, $c = 83.37$ Å with one complex in the asymmetric unit (55% solvent). A flash-frozen crystal maintained at −150 °C diffracts X-rays to 2.1 Å.

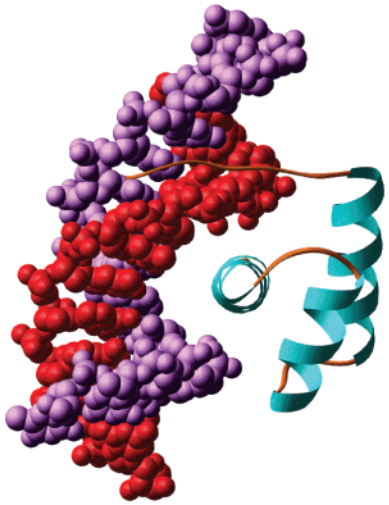
Structure Determination and Refinement. Data were collected on a single crystal at −150 °C using a MSC Raxis II imaging plate detector (Molecular Structure Corp., TX). CuKα X-rays were produced by a Rigaku RU 200 rotating anode source operating at 5 kW (50 kV x 100 mA). Data collection parameters are listed in Table 1. The data were processed with DENZO and scaled with SCALEPACK (*49*). AMoRe was used for the molecular replacement solution of the Msx-1−DNA complex (*50*). Because of high conservation among homeodomains in general, many of the known structures were used as search models. A stripped-down version of the even-skipped (Eve) homeodomain−DNA complex gave the only interpretable molecular replacement solution (*19*). All surface amino acid side chains not homologous to Msx-1 were mutated to Ala in this search model, and in addition the DNA was truncated to the three central base pairs at the binding site. Though this model contains only about one-third of the total asymmetric unit, a weak but correct molecular replacement solution was obtained with a correlation coefficient of 19.3 and an *R*-factor of 46.5%. The model was built using FRODO (*51*, *52*). Using $2F_o - F_c$ and $F_o - F_c$ maps, 153 water molecules were added and the model refined with CNS (*53*−*56*). The final *R*-factor stands at 19.8% and $R_{free} = 26.8\%$ with good stereochemistry as determined by PROCHECK (*57*). The R factors were calculated for data within 8−2.2 Å and $F_o > 2\sigma(F_o)$. The rms deviations from ideality for bond lengths and bond angles are 0.0192 Å and 2.174° respectively. The averages of the final *B*-factors for the HD and the DNA are 32.4 and 42.8 Å^2, respectively.

RESULTS

Structure of the Overall Complex. The overall structure of the Msx-1 HD−DNA complex is very similar to that seen for all other known high resolution homeodomain−DNA complexes, with the recognition helix of the HD running across the major groove of the DNA, roughly parallel to the planes of the bases (Figure 1A). It is this helix which makes the majority of the protein−DNA interactions. Additionally and analogously, the N-terminal arm of the Msx-1 HD tracks the minor groove of the DNA in an extended conformation and also makes several direct interactions with the DNA. As expected, the Msx-1 homeodomain binds the recognition helix as a monomer. Due to the propensity for homeodomains to make what are thought to be spurious pseudo-specific DNA interactions (*16*, *19*−*21*, *32*) resulting in two HD's bound to the same DNA in the context of crystallography, the oligonucleotide used here was carefully designed to contain only a single binding site that could conceivably be recognized by the Msx-1 HD. The sequence of the recognition element was designed to be optimal for Msx-1 HD binding as determined by selection assay (*47*) and is shown in Figure 1B. The most significant differences between the Msx-1 HD−DNA structure and the structures of most other HD−DNA complexes include an extension by several amino acids of the interaction of the N-terminal arm with the minor groove, and the degree of DNA bending seen in this complex. Both of these issues will be elaborated in detail below. The overall model includes 58 of the 60 amino acids that make up the homeodomain, 29 of the 32 bases of DNA and 153 water molecules. Additionally, the side chain atoms of residues R21, Q22 and R57 are disordered in our model.

Structure of the Msx-1 Homeodomain. As predicted from the high degree of sequence homology (Figure 1C), the structure of the Msx-1 homeodomain takes the form seen in those HDs whose structures are known. This structural homology is maintained due to the dispensation of several highly conserved hydrophobic residues that make up the small hydrophobic core of the domain. Mutation of these residues severely compromises both the transcriptional repression activity as well as the DNA-binding affinity of the Msx-1 HD, indicating the importance of these residues for structural integrity (*42*). Superposition of all known HD structures to the Msx-1 HD results in root-mean-square deviations between 0.54 and 0.85 Å. The most significant structural differences between HDs are seen in the N-terminal arm conformation. In fact much of this region is disordered in many of the known HD structures. The Msx-1 HD N-terminal arm on the other hand is unusually well ordered with only the first residue of the HD not accounted for in the final model. The Msx-1 HD is additionally stabilized by three salt bridges, which serve as "electrostatic cross-links" between each of the three helices (Table 2). The salt bridge,

A.



B.

```
          1  2  3  4  5   6  7  8  9  10 11 12 13 14 15 16
5'-T  G  T  C  A  C  T  A  A  T  T  G  A  A  G  G
      C  A  G  T  G  A  T  T  A  A  C  T  T  C  C  T-5'
     32 31 30 29 28 27 26  25 24 23 22 21 20 19 18 17
```

C.

```
                  *  *  ^       ^^  ^  ^    *    ^^          ^    *^^**^     *
Msx-1             NRKPRTPFTTAQLLALERKFRQKQ---YLSIAERAEFSSSLSLTETQVKIWFQNRRAKAKRLQ
MATa1             SPKGKSSISPQARAFLEQVF--RRKQSLNSKE-KEEVAKKCGITPLQVRVWFINKRMRSK
MATalpha2         KPYRGHRFTKENVRILESWFAKNIENPYLDTKGLENLMKNTSLSRIQIKNWVSNRRRKEKTIT
pbx               ARRKRRNFNKQATEILNEYFYSHLSNPYPSEEAKEELAKKCGITVSQVSNWFGNKRIRYKKNI
hoxb1             PSGLRTNFTTRQLTELEKEFHFNK---YLSRARRVEIAATLELNETQVKIWFQNRRMKQKKRE
ubx               RRRGRQTYTRYQTLELEKEFHTNH---YLTRRRRIEMAHALCLTERQIKIWFQNRRMKLKKEI
exd               ARRKRRNFSKQASEILNEYFYSHLSNPYPSEEAKEELARKCGITVSQVSNWFGNKRIRYKKNI
engrailed         EKRPRTAFSSEQLARLKREFNENR---YLTERRRQQLSSELGLNEAQIKIWFQNKRAKIKKST
antennapedia      RKRGRQTYTRYQTLELEKEFHFNR---YLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKEN
even-skipped      VRRYRTAFTRDQLGRLEKEFYKEN---YVSRPRRCELAAQLNLPESTIKVWFQNRRMKDKRQR
paired            QRRSRTTFSASQLDELERAFERTQ---YPDIYTREELAQRTNLTEARIQVWFQNRRARLRKQH

Majority          ARRGRTTFTKQQLLELEKEFHSNR---YLSRERREELAKKLGLTESQIKIWFQNRRMKLKKEI
```

FIGURE 1: (A) Ribbon (*66*) drawing of the Msx-1HD−DNA complex. The protein is shown in blue while the DNA is shown in red and purple. This view is looking down the recognition helix. (B) Sequence of the 16-bp DNA fragment used to crystallize the complex. The bases in bold (**TAAT**) are the core binding site for Msx-1. (C) The sequence of the Msx-1 construct aligned with other homeodomains found in crystal structures. In the majority sequence every 10th residue is underlined. The Msx-1 residues involved in DNA recognition are denoted by an asterisk (*), and those involved in HD core stabilization are marked with a carat (∧).

Table 2: Salt Bridges in Msx-1

anion	cation	distance (Å)
Glu 30 COO^-	Lys 23 NH_3	2.82
Glu 42 COO^-	Arg 31 NH_2	4.57
Glu 17 COO^-	Arg 52 NH_2	3.69

between E30 and K23 links helix one with helix two; the salt bridge between E42 and R31 links helix two with helix three; the salt bridge between E17 and R52 links helix three with helix one. Similar, though not identical salt bridges are observed in most of the other HD structures, indicating these electrostatic interactions to be important for domain stability. For example, in most HD's residue 30 is basic and interacts with the highly conserved E19 residue, which still represents an electrostatic interaction between helix one and two.

Protein−DNA Recognition. Most of the protein−DNA interactions between the HD recognition helix and the DNA major groove are common among other HD−DNA complexes whose structures are known, especially those HDs that recognize the canonical TAAT sequence. The I47 in Msx-1 makes a hydrophobic contact to Ade9 and Thy10 but this contact is seen no matter what residue is there, with valine and asparagine also occurring at position 47. Figure 2 schematically summarizes all of the interactions seen in the Msx-1 HD−DNA structure.

Residue Q50. Residue 50 deserves special mention because it is one of the critical residues involved in discriminative DNA recognition for distinct classes of homeodomains. When residue 50 is K, there is a clear preference for 5′-TAATCC-3′ sequences over other sequences such as GG and TA. The structure of the Q50K mutant *engrailed* homeodomain bound to a GG containing DNA sequence clearly explains this preference by showing a direct hydrogen bond between K50 and the GG of the DNA (*32*). What is more difficult to explain, however, is why Q50 variants *disfavor* the GG sequence relative to other sequences such as TG, which is part of the core binding sequence for the Msx-1 homeodomain, or TA, which is the preferred sequence

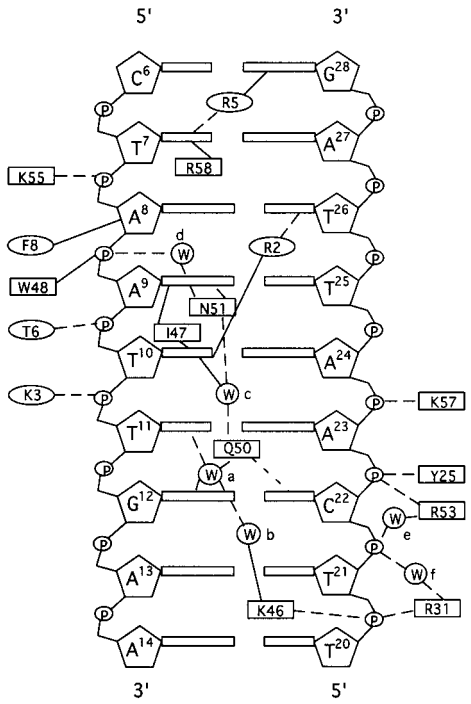


FIGURE 2: Contacts between the DNA and the protein in the complex. The major and minor groove DNA contacts are shown by squares and circles, respectively. Dotted lines indicate hydrogen bonding, and solid lines indicate hydrophobic interactions.

for the wild-type *engrailed* homeodomain. We believe that GG sequences are disfavored both by the Msx-1 and *engrailed* homeodomains because of the water-mediated interaction between Q50 and DNA. In the Msx-1 homeodomain−DNA complex there is a water-mediated interaction between Q50, Thy11, Gua12, and Wat a (Figure 3A). The 26.9 *B*-factor of Wat a is low compared to the average *B*-factors for all of the waters in the structure (43.6), indicating this water to be quite well ordered in our structure. The coordination sphere of Wat a interacts with Q50, Thy11, Gua12, and Wat b serving to completely define its hydrogen bonding pattern. Substitution of either Thy11 or Gua12 with a Cyt nucleotide would cause a hydrogen-bonding clash with Wat a necessitating some structural reorganization in this region. An almost identical duplication of the water-mediated interaction seen in the Msx-1−DNA complex can be seen in the *engrailed*−DNA complex structure (*20*). Binding studies for both the Msx-1 and engrailed homeodomain−DNA complexes have shown that the presence of a Cyt nucleotide at either of these positions causes a significant loss of binding affinity for each of these complexes (*32*, *47*). This interaction interface defines most of the Msx-1 base specificity flanking the core TAAT sequence. The conclusion to be drawn from this is that the fully coordinated Wat a is *the* critical component in the Thy11−Gua12 base specificity for Msx-1 and other homeodomains which favor this sequence.

An example of a structural reorganization in this region can be seen in the HoxB1−Pbx1−DNA complex structure (*24*). Though a water is present very near to the location of Wat a, the presence of a Cyt nucleotide has forced Q50 to move completely out of the region. Q50 actually makes interactions with the phosphate backbone in this structure. The paired HD structure also shows an important water mediated interaction for Q50 but the DNA in this region is a bit different from the engrailed and Msx-1 HDs and therefore we do not see this exact pattern in the paired structure (*16*). The paired HD also has a mutation of the DNA from TAATCA to TAACGA which encompasses part of the core sequence. There is a reduction in binding from this mutation but how much is due to the flanking sequence is hard to infer (*16*).

Residue A54. In most homeodomains, residue 54 is a reasonably large hydrophobic residue (I or M) that fills a cavity that exists between the recognition helix and the surface of the DNA, often making direct interactions with the DNA (*17*, *19*, *21*, *23*, *24*). In contrast, the Msx-1, *engrailed*, and paired homeodomains have alanines at this position (*16*, *20*). In Msx-1 and *engrailed* the lack of a large side chain produces an identical ring of ordered water molecules that surround the residue, some of them making interactions with the DNA (Figure 3B). These water-mediated interactions serve to replace the direct hydrophobic interaction that was lost. The paired HD has four of these waters but also has an R at position 57, which swings into this area to fill the gap. In the other two structures with A54, K57 has a different conformation that does not interfere with the water ring.

DNA Minor Groove and N-Terminal Arm Interactions. In addition to the interface between the HD recognition helix and the DNA major groove, which is reminiscent of bacterial helix-turn-helix−DNA interactions, homeodomains have additionally an N-terminal arm region lacking secondary structure that snakes across the minor groove, making critical interactions with the DNA. The Msx-1 HD N-terminal arm makes three direct interactions with DNA (Figure 3C). T6 makes a hydrogen bond to the DNA phosphate backbone, which is commonly found in other HD structures containing this residue (*19*, *20*, *24*). R5 makes a direct base contact to Thy7, which is virtually identical to the interaction seen in all other HD−DNA structures that have an R at position 5. In contrast to most other HD−DNA structures, however, the N-terminal arm of Msx-1 is well ordered all the way to R2. In fact, R2 forms a tight hydrogen bonding−salt bridge interaction with base Thy26, this can also be seen in the paired structure (*16*). The average *B*-factor of the N-terminal arm residues in Msx-1 is 40 Å^2, which compares well with the average *B*-factor of the entire protein (32.4 Å^2), indicating the N-terminal arm to be well ordered relative to the rest of the structure. We believe that the structural integrity of the Msx-1 N-terminal arm is partially preserved by the presence of two unconserved prolines in the sequence. The paired structure does not contain these prolines but is involved in a dimerization interaction with helix II and helix III of its second HD.

Hydration of the HD−DNA Interface. The protein−DNA interface not only includes protein−DNA interactions, but protein−water and DNA−water interactions as well. Though the complex buries 1832 Å^2 of surface area between protein and DNA, there is also significant hydration within the interface. It turns out that this hydration is remarkably well conserved between the homeodomain−DNA structures. By overlaying all known HD−DNA complex structures and looking for conserved waters, i.e., waters in different structures that were no more than 1.5 Å away from corresponding Msx-1 waters, we observed a remarkable conservation of water structure in these complexes (Table

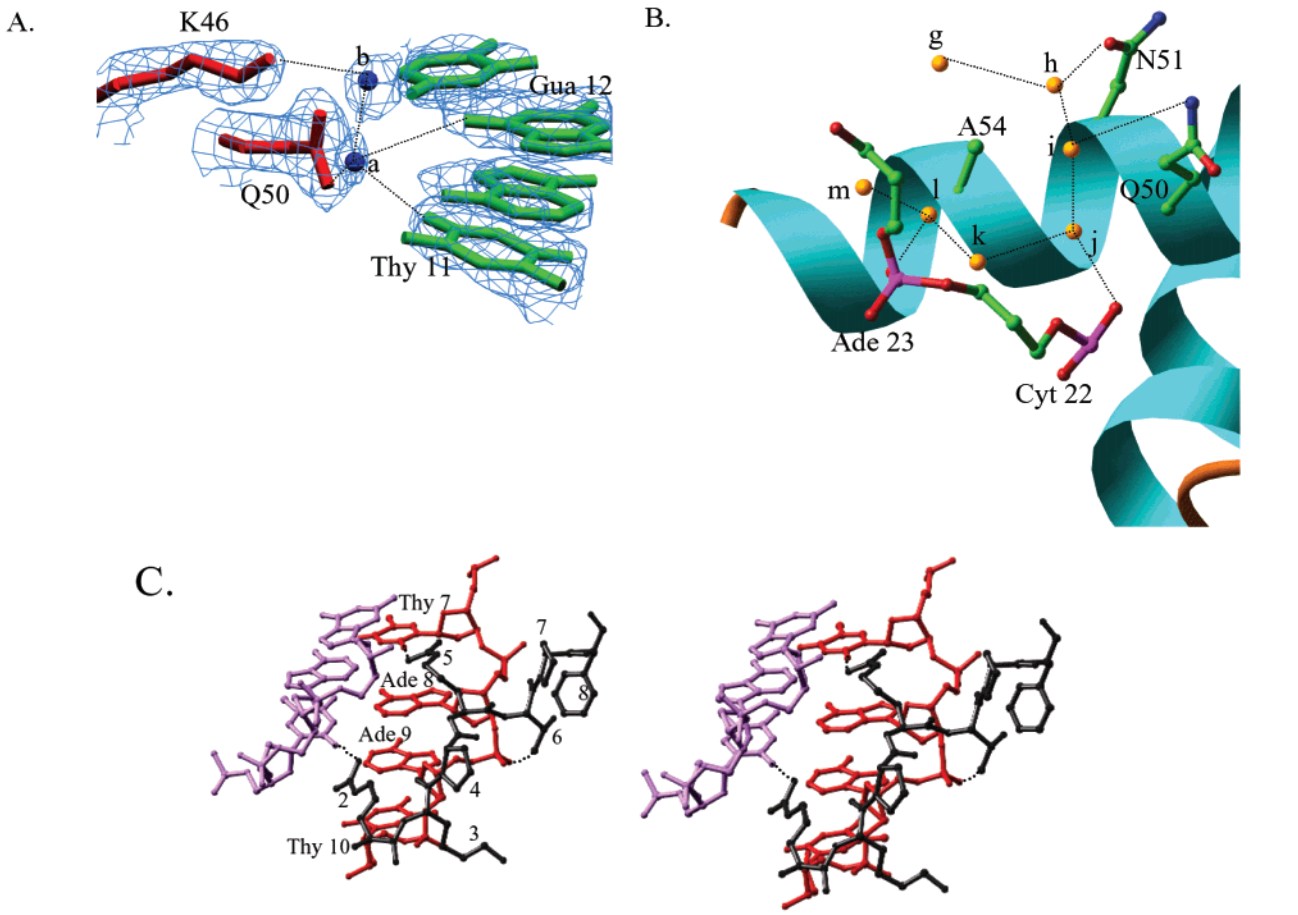


FIGURE 3: Structure of the protein−DNA interface. (A) Simulated annealing omit map of the Q50−water−DNA interaction, contoured at 1.5σ. Picture was made with Setor (*67*). (B) The conserved water ring that surrounds A54 and fills the cavity present between the protein and the DNA backbone in this region. (C) Stereoview of the trajectory of the N-terminal arm of Msx-1. Hydrogen bonds are represented by dotted lines in all parts of this figure.

Table 3: Conserved Water Table

HD/DNA complex	res. limit (Å)	no. conserved waters	data collection condition	total no. of waters	ref
Antennapedia	2.4	8	room temp	38	*21*
MATa1/MATα2	2.5	a1−7, α2−10	frozen	58	*17, 18*
even-skipped	2.0	12	frozen	68	*19*
HoxB1/Pbx1	2.35	hox-10,pbx-13	frozen	61	*24*
Ubx/Exd	2.4	ubx-15,exd-12	frozen	110	*23*
engrailed	2.2	13	room temp	53	*20*
engrailed mutant	1.9	14	frozen	183	*32*
paired	2.0	12	frozen	242	*16*
Msx-1	2.2	16	frozen	153	

3). Table 3 contains those HD complexes that diffract to better than 2.5 Å. Figure 4 depicts the Msx-1−DNA structure with all 16 of the most conserved waters. Note how virtually all of these waters create a hydration sphere *between* the recognition helix and the major groove of the DNA. A total of 130 of the possible 180 conserved waters are observed (72%). A total of 23 of the 50 missing waters are in the three structures whose water structure is least conserved (MATa1, MATalpha2, and *Antennapedia* HD−DNA complexes). Of the 130 conserved waters identified in this analysis, only 16 are more than 1.2 Å away from the corresponding waters of the Msx-1 structure and most are within 1 Å of an Msx-1 water. Thirteen of the sixteen waters make bridging interactions between the protein and DNA

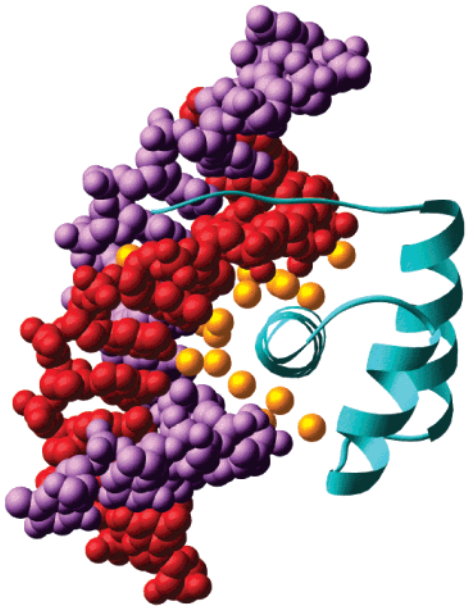


FIGURE 4: Conserved water network present in the homeodomain−DNA complexes studied. The waters are shown in gold.

while three make contacts only to the DNA. The overwhelming majority of these waters are making similar interactions in each of the HD−DNA structures. All of these interactions and distances are tabulated in the Supporting Information to

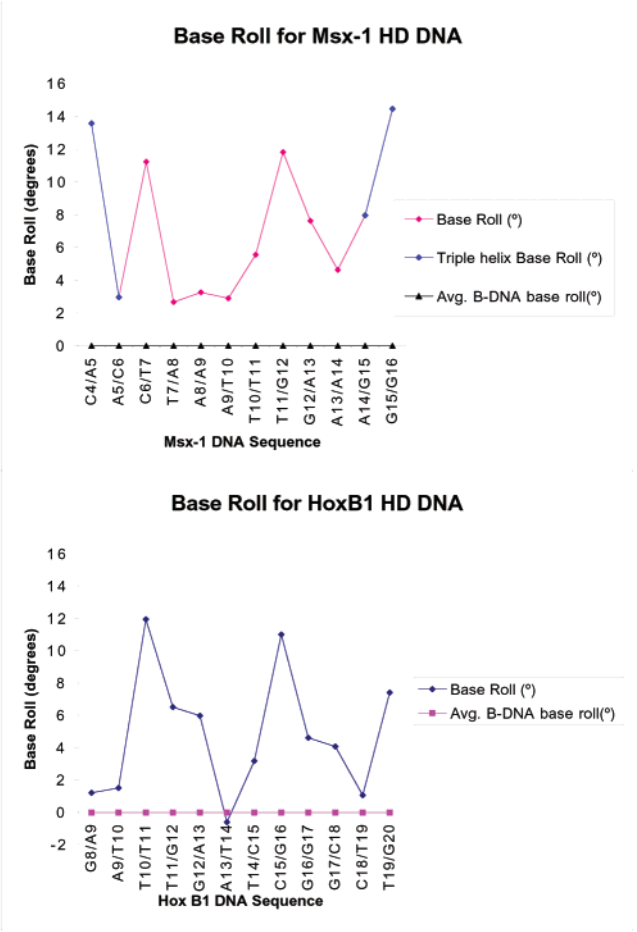


FIGURE 5: Plots of DNA base roll (A) and helical twist (B) as a function of DNA sequence. Horizontal dotted or dashed lines indicate average values for B-DNA. The sequence of the top strand only is shown. All parameters were calculated with the program Curves (*68*).

this paper. It appears that the strong conservation in HD−DNA interactions extends to the sphere of hydration of these complexes. Interestingly, no conserved waters are seen outside the protein−DNA interface. These waters have a low average *B*-factor of 30 Å^2 when compared to the average water *B*-factor in the Msx-1−DNA complex which is 43.6 Å^2.

Structure of the DNA. Most of the DNA−HD complexes have a modest bend in the DNA caused by protein binding. In the MATa1−MATα2−DNA heterodimer complex, there is a significant 60° bend in the DNA (*17*), while in the majority, there is a modest bend of about 10−13° toward the major groove (*17*, *19*−*21*, *23*, *24*, *32*). The paired HD complex shows a 21° bend (*16*). This bend appears to be caused by the protein pulling the DNA toward it on the major groove side. This deformation is very common in major groove binding proteins including helix turn helix proteins and other DNA binding proteins (*36*, *58*−*60*). The DNA in the Msx-1 HD−DNA complex, however, exhibits a much more severe bend of 28° relative to other monomeric HD−DNA complexes. The very close structural homology between the Msx-1 HD−DNA complex and other HD−DNA complexes indicated that the extra bending was not due to protein−DNA interactions. Further evidence of this can be gleaned from Figure 5 where the base roll angle is plotted for each sequence in the DNA. The roll angle is shown for both the Msx-1 HD−DNA complex and for the HoxB1 HD in the HoxB1−Pbx1−DNA complex, a structure that is highly homologous to that of the Msx-1 HD−DNA complex (*24*). The base roll parameter can give an indication of relative bend per base for a DNA sequence, since a DNA bend must be accommodated by some base roll. As shown in Figure 5, the base roll per residue is very similar for the two structures in the core binding region, but the Msx-1 base roll parameters deviate significantly at the end of the sequence where the crystal packing-induced triple helix is formed. It is apparently the formation of this triple helix that causes the larger bend in the DNA of the Msx-1 HD−DNA complex.

This triple helix formed when the first two bases of the double stranded portion of the DNA melted (Gua2 and Thy3). One of the resulting single-stranded segments (Ade31 and Cyt32) then formed a triple helix interaction with the end of a neighboring DNA in the crystal, while the other single-stranded segment (Thy1, Gua2, and Thy3) was not seen in the structure and is likely disordered. The result is that a triple helix is formed between Gua15−Cyt19 (Cyt32) and Gua16−Cyt18 (Ade31) with the base in parentheses being the third base forming the triple helix between the Watson and Crick base pairs. In addition, a third triple helix base step is formed by the interaction of an overhanging Thy17 with Cyt4−Gua30, which is butted against Gua16−Cyt18 in the crystalline lattice. The complete triple helix is depicted in Figure 6, panels A and B. There are a few possible explanations for this phenomenon. Some DNase I footprinting experiments have shown that there is a propensity for triple helixes to form in GA and GT rich oligonucleotides (*61*) such as ours. The Msx-1 DNA has an AGG sequence in the region where the triple helix forms. Also, in all three steps of triple helix seen in this structure base protonation is observed. We believe this protonation to be a result of the comparatively low pH (pH 4.6 in the crystallization). The protonation values for the nucleotides appear to support this theory with the pK_as of the ring nitrogens being around pH 4 (*62*). This low pH served to both destabilize the double helix and to stabilize the triple helix by base protonation. Figure 6C shows the Gua16−Cyt18 (Ade31) triple helix step, a structure not seen previously in DNA or protein−DNA structures.

The source of the 15° bend in the region of the triple helix is more difficult to explain, since triple helix formation has not previously been correlated with DNA bending. Clearly, however, in this unusual case of triple helix formation, modest DNA bending is a result. There are examples in the literature of a triple helix forming at the ends of double-stranded DNA with overhangs (*63*−*65*), and these have been utilized in designing DNAs for crystallization purposes. The triple helix in this case is not a pseudo continuous helix but one that has packed against their opposing strands. So it is a rare occurrence in itself but the way that it has formed makes it even more unusual.

Msx-1 HD Protein Interactions. In addition to its interaction with DNA, the Msx-1 HD makes specific and functionally important interactions with other proteins as well. Specifically, GST pulldown experiments and gel shift analyses have shown that Msx-1 binds TBP and the TBP−TATA box DNA complex. While residues in the N-terminal arm of the homeodomain are required for this interaction

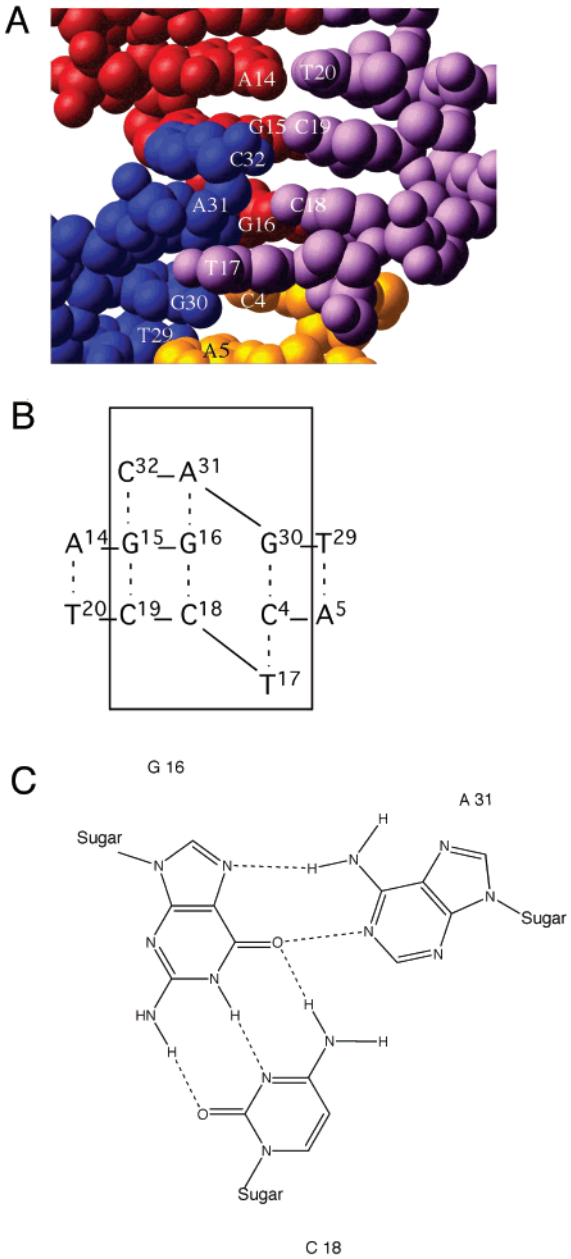


FIGURE 6: Structure of the triple helix. (A) Overall triple helix interaction for the stacking DNA. (B) Triple helix schematic for the unusual helical interaction of the stacked DNAs. (C) The triple helix trio of Gua16:Cyt18 and Ade31, a previously unseen interaction in triple helix combinations.

(*42*), residues in other parts of the HD had little effect on this interaction.

Specifically the F8A, R5A, K3A triple mutation completely abrogated Msx-1 HD−TBP interaction. All three of these residues reside in the N-terminal arm. In contrast, both the double mutant, L16A, F20A, and the triple mutant I47A, E50A, N51A, bind TBP with near wild-type affinity. While L16 and F20 lie in helix I, I47, E50, and N51 are all located in the DNA binding interface of helix III. All of these mutants have lost measurable DNA binding affinity, indicating that the interaction with TBP is independent of the protein's DNA binding interface. Both the helix I mutants and the N-terminal arm mutants were inactive toward transcriptional repression in the context of the full-length protein. Mutants in helix III, on the other hand, were still capable of transcriptional repression, again indicating that the DNA binding interface is not required for this activity. The two residues in helix I are part of the conserved hydrophobic core and when mutated to Alanine probably cause important changes in the structure of the HD. Our structure indicates that the structure of the Msx-1 N-terminal arm is somewhat unique relative to many other HDs and also indicate it to be comparatively less flexible. It is possible that this increased structural rigidity is important in defining a TBP binding interface.

In addition to its interaction with TBP, a member of the transcriptional basal machinery, Msx-1 appears to make interactions with several other members of the homeobox gene family. These interactions appear to play important regulatory roles in transcription. The HD protein Dlx, a transcriptional activator, is one of these proteins. Heterodimerization between Msx-1 and Dlx causes mutual loss of transcriptional activity (*43*). Mutations both in the N-terminal arm and in the recognition helix of Msx-1 severely compromise this interaction, indicating the interaction to encompass the entire DNA binding surface of the protein. Presumably, a similar interaction surface is used in the interactions between Msx-1 and other homeodomain proteins such as Pax3 and Dhx2 (*45*, *46*).

The same mutations studied above for TBP binding were also studied for Dlx binding. The N-terminal arm residues again proved important as the Msx-1−Dlx interaction was demolished when residues in this region were mutated. Helix I and helix III mutations also reduced dimerization, indicating that the interaction is at least partially mediated via the recognition helix. The dimerization of the Msx and Dlx proteins seems to differ from the interaction of Msx-1 with TBP. Gel retardation assays indicate that Msx and Dlx proteins bind to homeodomain sites as monomers. Dimerization excludes DNA binding, and mutual repression of their transcriptional activities is the result.

CONCLUSIONS

Comparison of our 2.2 Å Msx-1 homeodomain−DNA complex structure with all other homeodomain−DNA complexes has illustrated some enlightening aspects of homeodomain structure and DNA recognition. Most notably comparison of ordered water molecules between HD−DNA complex structures has identified a well conserved hydration interface between the recognition helix and major groove. More insight has been gained concerning how residue Q50 determines HD binding site preferences. The coordination sphere of Wat a (a very conserved water) dictates the interaction of Q50 with the DNA flanking sequence. Mutations of this flanking sequence led to a significant loss in binding affinity of the complex, mainly due to the disruption of the critical water coordination sphere of Wat a. Our structure has reaffirmed the importance of water-mediated interactions for HD−DNA binding.

There are also important differences between our structure and those of other structures. The N-terminal arm of Msx-1 is unusually well ordered, exhibiting clear electron density to the second residue. We believe that two proline residues not usually seen in HD N-terminal arms stabilize the structure. A combination of HD−DNA interactions and the formation of an unusual triple helix packing interaction leads

to a significant, 28° bend in the DNA, which is significantly larger than that seen in other monomeric HD−DNA complex structures. Though an artifact of crystal-packing interactions, the triple helix formed between stacked DNA helices provides an interesting example of triple helix structure. It contains an unusual GC (A) step not previously seen in triple helices and leads to a significant DNA bend not typically seen in triple helix structures.

SUPPORTING INFORMATION AVAILABLE

Tables of the interactions and distances of water. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES

1. Cillo, C., Faiella, A., Cantile, M., and Boncinelli, E. (1999) *Exp. Cell Res. 248*, 1−9.
2. Gehring, W. J., and Hiromi, Y. (1986) *Annu. Rev. Genet. 20*, 147−73.
3. McGinnis, W., and Krumlauf, R. (1992) *Cell 68*, 283−302.
4. Lewis, E. B. (1978) *Nature 276*, 565−70.
5. Manak, J. R., and Scott, M. P. (1994) *Dev. Suppl.* 61−77.
6. Koeneman, K. S., Yeung, F., and Chung, L. W. (1999) *Prostate 39*, 246−61.
7. Tucker, A. S., Al Khamis, A., and Sharpe, P. T. (1998) *Dev. Dyn. 212*, 533−9.
8. Nugent, P., and Greene, R. M. (1998) *In Vitro Cell Dev. Biol. Anim. 34*, 831−5.
9. Wang, Y. H., Upholt, W. B., Sharpe, P. T., Kollar, E. J., and Mina, M. (1998) *Dev. Dyn. 213*, 386−97.
10. Koshiba, K., Kuroiwa, A., Yamamoto, H., Tamura, K., and Ide, H. (1998) *J. Exp. Zool. 282*, 703−14.
11. Blanco, R., Jara, L., Villaseca, C., Palomino, H., and Carreno, H. (1998) *Rev. Med. Chil. 126*, 781−7.
12. Ferrari, D., Lichtler, A. C., Pan, Z. Z., Dealy, C. N., Upholt, W. B., and Kosher, R. A. (1998) *Dev. Biol. 197*, 12−24.
13. Stelnicki, E. J., Komuves, L. G., Holmes, D., Clavin, W., Harrison, M. R., Adzick, N. S., and Largman, C. (1997) *Differentiation 62*, 33−41.
14. Tiberio, C., Barba, P., Magli, M. C., Arvelo, F., Le Chevalier, T., Poupon, M. F., and Cillo, C. (1994) *Int. J. Cancer 58*, 608−15.
15. De Vita, G., Barba, P., Odartchenko, N., Givel, J. C., Freschi, G., Bucciarelli, G., Magli, M. C., Boncinelli, E., and Cillo, C. (1993) *Eur. J. Cancer 6*, 887−93.
16. Wilson, D. S., Guenther, B., Desplan, C., and Kuriyan, J. (1995) *Cell 82*, 709−19.
17. Li, T., Stark, M. R., Johnson, A. D., and Wolberger, C. (1995) *Science 270*, 262−9.
18. Li, T., Jin, Y., Vershon, A. K., and Wolberger, C. (1998) *Nucleic Acids Res. 26*, 5707−18.
19. Hirsch, J. A., and Aggarwaal, A. K. (1995) *EMBO J. 14*, 6280−6291.
20. Fraenkel, E., Rould, M. A., Chambers, K. A., and Pabo, C. O. (1998) *J. Mol. Biol. 284*, 351−61.
21. Fraenkel, E., and Pabo, C. O. (1998) *Nat. Struct. Biol. 5*, 692−7.
22. Wolberger, C., Vershon, A. K., Liu, B., Johnson, A. D., and Pabo, C. O. (1991) *Cell 67*, 517−28.
23. Passner, J. M., Ryoo, H. D., Shen, L., Mann, R. S., and Aggarwal, A. K. (1999) *Nature 397*, 714−9.
24. Piper, D. E., Batchelor, A. H., Chang, C. P., Cleary, M. L., and Wolberger, C. (1999) *Cell 96*, 587−97.
25. Kissinger, C. R., Liu, B. S., Martin Blanco, E., Kornberg, T. B., and Pabo, C. O. (1990) *Cel. 63*, 579−90.
26. Gruschus, J. M., Tsao, D. H., Wang, L. H., Nirenberg, M., and Ferretti, J. A. (1999) *J. Mol. Biol. 289*, 529−45.
27. Guntert, P., Qian, Y. Q., Otting, G., Muller, M., Gehring, W., and Wuthrich, K. (1991) *J. Mol. Biol. 217*, 531−40.
28. Ippel, H., Larsson, G., Behravan, G., Zdunek, J., Lundqvist, M., Schleucher, J., Lycksell, P. O., and Wijmenga, S. (1999) *J. Mol. Biol. 288*, 689−703.
29. Schott, O., Billeter, M., Leiting, B., Wider, G., and Wuthrich, K. (1997) *J. Mol. Biol. 267*, 673−83.
30. Qian, Y. Q., Billeter, M., Otting, G., Muller, M., Gehring, W. J., and Wuthrich, K. (1989) *Cell 59*, 573−80.
31. Qian, Y. Q., Furukubo Tokunaga, K., Resendez Perez, D., Muller, M., Gehring, W. J., and Wuthrich, K. (1994) *J. Mol. Biol. 238*, 333−45.
32. Tucker-Kellogg, L., Rould, M. A., Chambers, K. A., Ades, S. E., Sauer, R. T., and Pabo, C. O. (1997) *Structure 5*, 1047−54.
33. Clarke, N. D., Kissinger, C. R., Desjarlais, J., Gilliland, G. L., and Pabo, C. O. (1994) *Protein Sci. 3*, 1779−1787.
34. Tan, S., and Richmond, T. J. (1998) *Nature 391*, 660−6.
35. Chasman, D., Cepek, K., Sharp, P. A., and Pabo, C. O. (1999) *Genes Dev. 13*, 2650−7.
36. Klemm, J. D., Rould, M. A., Aurora, R., Herr, W., and Pabo, C. O. (1994) *Cell 77*, 21−32.
37. Catron, K. M., Wang, H., Hu, G., Shen, M. M., and Abate Shen, C. (1996) *Mech. Dev. 55*, 185−99.
38. Shimeld, S. M., McKay, I. J., and Sharpe, P. T. (1996) *Mech. Dev. 55*, 201−10.
39. Bendall, A. J., and Abate-Shen, C. (2000) *Gene 247*, 17−31.
40. Catron, K. M., Zhang, H., Marshall, S. C., Inostroza, J. A., Wilson, J. M., and Abate, C. (1995) *Mol. Cell Biol. 15*, 861−71.
41. Semenza, G. L., Wang, G. L., and Kundu, R. (1995) *Biochem. Biophys. Res. Commun. 209*, 257−62.
42. Zhang, H., Catron, K. M., and Abate Shen, C. (1996) *Proc. Natl. Acad. Sci. U.S.A. 93*, 1764−9.
43. Zhang, H., Hu, G., Wang, H., Sciavolino, P., Iler, N., Shen, M. M., and Abate-Shen, C. (1997) *Mol. Cell. Biol. 17*, 2920−2932.
44. Newberry, E. P., Latifi, T., Battaile, J. T., and Towler, D. A. (1997) *Biochemistry 36*, 10451−62.
45. Bendall, A. J., Rincon Limas, D. E., Botas, J., and Abate Shen, C. (1998) *Differentiation 63*, 151−7.
46. Bendall, A. J., Ding, J., Hu, G., Shen, M. M., and Abate-Shen, C. (1999) *Development 126*, 4965−76.
47. Catron, K. M., Iler, N., and Abate, C. (1993) *Mol. Cell Biol. 13*, 2354−65.
48. DeWees, S., and Geiger, J. H. (1999) *Acta Crystallogr., Sect. D 55*, 2039−40.
49. Otwinowski, Z. (1993) in *Data Collection and Processing* (Sawyer, L., Issacs, N., and Bailey, S., Eds.) pp 56−62, SERC Daresbury Laboratory, Daresbury, U.K.
50. Navaza, J. (1994) *Acta Crystallogr., Sect. A 50*, 157−63.
51. Jones, T. A. (1985) *Methods Enzymol. 115*, 157−71.
52. Jones, T. A., Zou, J. Y., cowan, S. W., and Kjeldgaard, M. (1991) *Acta Crystallogr., Sect. A 47*, 110−9.
53. Brunger, A. T., Kuriyan, J., and Karplus, M. (1987) *Science 235*, 458−460.
54. Brunger, A. T., Krukowski, A., and Erickson, J. (1990) *Acta Crystallogr., Sect. A46*, 585−93.
55. Brunger, A. T. (1992) *X-PLOR, version 3.1, a System for X-ray Crystallography and NMR*, Yale University Press, New Haven, CT.
56. Brunger, A. T. (1992) *Nature 355*, 472−74.
57. Laskowski, R. A., Moss, D. S., and Thornton, J. M. (1993) *J. Mol. Biol. 231*, 1049−67.
58. Clark, K. L., Halay, E. D., Lai, E., and Burley, S. K. (1993) *Nature 364*, 412−20.
59. Konig, P., Giraldo, R., Chapman, L., and Rhodes, D. (1996) *Cell 85*, 125−36.
60. Mondragon, A., and Harrison, S. C. (1991) *J. Mol. Biol. 219*, 321−34.
61. Chandler, S. P., and Fox, K. R. (1996) *Biochemistry 35*, 15038−48.
62. Ts'o, P. O. P., and Eisinger, J. (1974) *Basic principles in nucleic acid chemistry*, Academic Press, New York.

63. Luisi, B. F., Xu, W. X., Otwinowski, Z., Freedman, L. P., Yamamoto, K. R., and Sigler, P. B. (1991) *Nature 352*, 497−505.
64. Schultz, S. C., Shields, G. C., and Steitz, T. A. (1991) *Science 253*, 1001−7.
65. Van Meervelt, L., Vlieghe, D., Dautant, A., Gallois, B., Precigoux, G., and Kennard, O. (1995) *Nature 374*, 742−4.
66. Carson, M. (1991) *J. Appl. Crystallogr. 24*, 958−61.
67. Evans, S. V. (1993) *J. Mol. Graphics 11*, 134−8.
68. Lavery, R., and Sklenar, H. (1988) *J. Biomol. Struct. Dyn. 6*, 63−91.

BI0108148