Eur Biophys J (2002) 31: 306–316
DOI 10.1007/s00249-002-0217-3

ARTICLE

Jianxin Duan · Lennart Nilsson

The role of residue 50 and hydration water molecules in homeodomain DNA recognition

Received: 18 October 2001 / Revised: 1 March 2002 / Accepted: 1 March 2002 / Published online: 18 April 2002

Abstract We conducted molecular dynamics simulations on several wild-type and mutant homeodomain-DNA complexes to investigate the role of residue 50 in homeodomain-DNA interaction and the behavior of interfacial hydration water. Our results suggest that this residue interacts more favorably with its consensus sequence and thus plays a considerable role in DNA recognition. However, residue 50 was not responsible for DNA recognition alone. Other residues in the vicinity could interact with residue 50 in cooperation upon DNA binding. We also found the lifetime for some water in the protein-DNA interface can be as high as nanoseconds and that a few well-conserved sites for water-mediated hydrogen bonds from protein to DNA are occupied by high-mobility hydrating waters.

Keywords Molecular dynamics · Homeodomain · DNA binding · Hydration water

Introduction

In a biological cell, DNA binding proteins affect several crucial processes such as cell division, differentiation and protein expression. The proteins must be able to discriminate closely related DNA sequences and bind with high affinity to the proper DNA site. Understanding the binding mechanisms and specificity of the protein-DNA interactions is vital in comprehending these crucial processes. Many structures of protein-DNA complexes have been solved and efforts have been made to characterize the protein-DNA interaction in search of a simple "code" (Jones et al. 1999; Kono and Sarai 1999; Mandel-Gutfreund et al. 1995; Nadassy et al. 1999; Pabo and Nekludova 2000; Suzuki and Gerstein 1995). Initially the interactions which contribute to the specificity were considered to be direct sidechain to base contacts, the so-called direct read-out. Since the solution of the X-ray structure of the *Escherichia coli* trp repressor (Otwinowski et al. 1988), in which the interfacial water network seems to mediate specific contacts between protein and DNA, and subsequent biochemical characterization confirmed the observation (Haran et al. 1992), the role of interfacial water molecules in DNA recognition has received increasing attention (Janin 1999; Jones et al. 1999; Nadassy et al. 1999). With high-resolution X-ray crystallography and NMR it is now possible to determine the positions of these water molecules, and NMR experiments are also able to estimate their lifetime in the interface. Yet the questions of their mobility, dynamics and their contribution to specificity and affinity need to be explored (Janin 1999).

Homeodomain proteins are a family of gene regulatory proteins which control cell differentiation during the early stages of embryo development of many eukaryotic organisms (Gehring et al. 1994a, 1994b; Kornberg 1993; Scott et al. 1989; Tullius 1995). The DNA binding structural module, the homeodomain, is a good model to study the factors that determine the affinity and specificity. The biochemical and structural properties of this 60 amino acid residue long module are well documented. Both in vivo and in vitro studies show that the homeodomain is capable of sequence-specific binding. The DNA consensus sequence recognized by homeodomains in vitro is 5′-TAATNN-3′. The structure of this domain is highly conserved and consists of three helices, the third of which is inserted into the major groove of the DNA in a strikingly conserved manner. On the minor groove side, the N-terminal arm of the domain is also contributing to DNA binding. Extensive contacts from the sidechains of the third helix to the bases were documented in all homeodomain structures and the sequence of this helix is also remarkably well conserved. Because the third helix makes contact with

J. Duan · L. Nilsson (✉)
Karolinska Institute, Department of Biosciences at Novum,
141 57 Huddinge, Sweden
E-mail: lennart.nilsson@biosci.ki.se
Tel.: +46-8-6089228
Fax: +46-8-6089290

the 5′-TAATNN-3′ sequence, it is referred to as the recognition helix. The core 5′-TAAT-3′ bases interact with residues 47 and 51 from the major groove side and with residues on the N-terminal arm from the minor groove side (Gehring et al. 1994a, 1994b; Kornberg 1993; Scott et al. 1989; Tullius 1995). The identity of the specificity determining residue for DNA recognition is much argued. It has been hypothesized that residue 50 (glutamine, lysine, serine, histidine, cysteine, isoleucine, arginine or glutamate) is the sequence discriminating residue (Hanes and Brent 1991; Stepchenko et al. 1997; Treisman et al. 1989). Lately this view has been challenged by other biochemical and structural studies (Ades and Sauer 1994; Grant et al. 2000; Tucker-Kellogg et al. 1997; Vershon et al. 1995). In the experimental structures it was found that lysine in this position has a distinct conformation, while serine, glutamine and cysteine were protruding into the protein-DNA interface but there were no specific contacts to DNA. The observation of a network of bound water molecules around these residues led to the hypothesis that the residues are part of a water-mediated hydrogen bond network which facilitates recognition. The importance of residue 50 in DNA recognition is an unsettled issue. The various experimental results either advocate or overthrow the hypothesis, largely depending on which amino acid is in position 50. We choose to address the question by simulating different complexes with different residues (lysine, glutamine, serine and cysteine) in position 50. There are four other naturally occurring amino acid residues, arginine, isoleucine, histidine and glutamate, at that position. Since they are rare in the homeodomain family (isoleucine, 1 case; glutamate, 2 cases; arginine, 2 cases; histidine, 3 cases) and since there are no experimental structures available, we choose to exclude them from this study.

In this work, we attempt to answer questions about the structural and dynamic properties of different sidechains of residue 50 in DNA recognition. We also study the dynamics of the interfacial water molecules and their part in aiding DNA binding. As a scaffold we chose the MAT α2 homeodomain, a yeast mating type regulating protein and mutated in silico the serine at position 50 to lysine, glutamine and cysteine. The mutants were simulated together with different homeodomain-specific DNA sequences using molecular dynamics (MD) techniques. We found that the residue at position 50 is important for DNA recognition and that the interfacial water molecules and their long lifetimes and high mobility are crucial for binding specificity and affinity.

Methods

System set-up

The crystal structures of MAT α2 homeodomain (PDB code: 1apl) (Wolberger et al. 1991), antennapedia homeodomain (PDB code: 9ant) (Qian et al. 1993), engrailed Q50K mutant homeodomain (PDB code: 2hdd) (Tucker-Kellogg et al. 1997), engrailed homeodomain (PDB code: 1hdd) (Kissinger et al. 1990) and POU Oct-1 homeodomain (PDB code: 1oct) (Klemm et al. 1994) were collected from the Protein Data Bank (Bernstein et al. 1977). The backbone of the third helices of 9ant, 2hdd and 1oct were superimposed onto that of 1apl using Modeller (Sali and Blundell 1993). The serine at position 50 in 1apl was mutated to glutamine, lysine and cysteine by transferring the coordinates of the most populated conformer of these residues from 9ant, 2hdd and 1oct, respectively. The protein structures are referred to as S50, Q50, K50 and C50, in accordance with the amino acid residue at position 50. The DNA bases that do not make contacts with the proteins were removed and the backbone of each of the DNA fragments was also superimposed onto that of 1apl. The sequences of the DNA fragments were: 5′-CATGTAATT (1apl), 5′-GCCATTAGA (9ant), 5′-GGGATTACA (2hdd) and 5′-CTTATTTGC (1oct). The DNA structures were then merged with the protein structures, as shown in Table 1. The complexes were named X50WT, if the DNA sequence is the native sequence to the corresponding residue 50 in the experimental structures, or else X50MUT. All together there were seven system set-ups. Hydrogen atoms were added to the molecules using the subroutine HBUILD (Brünger and Karplus 1988) in CHARMM (Brook et al. 1983). The systems were immersed into a pre-equilibrated TIP3P (Jorgensen et al. 1983) water box of size $55\times50\times40$ Å^3. The total surplus of the negative charges in the system was neutralized by adding sodium ions 6 Å away from the phosphorus atom on the bisector of the phosphate oxygen atoms. All water molecules with the oxygen atom closer than 2.8 Å to the solute were deleted.

The minimization and subsequent dynamics simulations were all performed by using the CHARMM program (Brook et al. 1983) and the all-atom version 22 parameters for proteins and water (MacKerell et al. 1998) and version 27 for nucleic acids (Foloppe and MacKerell 2000). The protein and the DNA were initially fixed and the system was minimized with 200 steps of steepest descent (SD) followed by 200 steps of adopted basis Newton-Raphson (ABNR). The restraint was then altered to 10 kcal mol^{-1} Å^{-2} harmonic constraints on the backbones and the system was minimized with 100 steps of ABNR. The constraints were then tuned to 5 kcal mol^{-1} Å^{-2} and another 100 steps of ABNR minimization were conducted. The solutes were then again fixed for a 25 ps constant volume and temperature MD simulation with periodic boundary condition at 300 K, allowing the water and the ions to relax around the macromolecules. Immediately after the simulation, 100 steps SD minimization were performed and the restraints were removed to perform another 200 steps ABNR minimization. The final structures were the input structures to the MD simulations.

MD simulation

In order to mimic a long DNA molecule, harmonic constraints were applied on the Watson-Crick hydrogen bonds at the base pairs at the 3′ and 5′ ends of the DNA. Periodic boundary conditions were applied for all MD simulations. The simulations were carried out at a constant temperature of 300 K and constant pressure of 1 atm by internal virial using the weak coupling algorithm by Berendsen et al. (1984). The compressibility constant was

Table 1. The system set-ups. In the X50MUT simulations the DNA sequences are the same as the S50WT simulation

	Protein	Consensus DNA sequence
S50WT	S50	T T A C A T
K50WT	K50	T A A T C C
Q50WT	Q50	T A A T G G
C50WT	C50	A A A T A A
K50MUT	K50	T T A C A T
Q50MUT	Q50	T T A C A T
C50MUT	C50	T T A C A T

4.63×10^{-5} atm^{-1} and both the temperature and pressure coupling constants were 5 ps. The time-step was 2 fs and all bonds involving hydrogen atoms were constrained by applying the SHAKE algorithm (van Gunsteren and Berendsen 1977). The coordinates were saved every 0.5 ps and the non-bond and image lists were updated when necessary. A constant dielectric constant of 1.0 was applied and the cutoff for long-range interactions was set to 12 Å with an atom-based force shift function. The non-bonded list cut-off was set to 14 Å. Each system was simulated for 1.6 ns. Average structures were calculated from 0.4 ns to the end of the simulations and minimized by running 800 steps SD.

The criterion on a hydrogen bond is that the distance between the acceptor and the hydrogen atom should be less than 2.4 Å. A water bridge is defined to exist between two groups, which simultaneously form hydrogen bonds to the same water molecule. Lifetimes of hydrogen bonds and water-bridged hydrogen bonds were analyzed with 5 ps time resolution. The cutoff distance for van der Waals contact was set to 4 Å.

The calculations were carried out either on DEC alpha workstations or PC clusters of Pentium II 450 MHz processors running Linux.

Results and discussion

In total we describe seven simulations of 1.6 ns each. First, we present the general features of the simulations and then move on to the differences between the simulations in terms of residue 50 and its contacts with DNA and other residues. Lastly, the behavior of the ions and the water molecules will be characterized. In the presentation and the tables the numbering of the nucleotides will follow Table 2.

All simulations were stable, as judged from the time evolution of the overall RMS deviation (Fig. 1) and they seemed to have reached equilibrium at about 400 ps. The backbone RMS deviation of the averaged structures compared to starting structures was around 1.5 Å for all except C50WT, which was 2.8 Å. The residue-based RMS deviations compared to the averaged structures indicate, as expected, that the terminals of the protein and the DNA are more flexible, and that the protein sidechains are more flexible than the backbone, in contrast to DNA whose backbone is less rigid (Fig. 1). The RMS fluctuations were around 1 Å for the backbone and somewhat higher for the sidechains and bases. The typical behavior of our simulations is represented by the S50WT simulation shown in Fig. 1.

Table 2. DNA sequences used in the simulation systems. The conserved ATTA/TAAT sequences are highlighted in bold

C50DNA		1	2	3	4	5	6	7	8	9	
	5-	C	T	T	**A**	**T**	**T**	**T**	G	C	
		G	A	A	**T**	**A**	**A**	**A**	C	G	-5
Q50 DNA	5-	G	C	C	**A**	**T**	**T**	**A**	G	A	
		C	G	G	**T**	**A**	**A**	**T**	C	T	-5
S50 DNA7	5-	C	A	T	**G**	**T**	**A**	**A**	T	T	
		G	T	A	**C**	**A**	**T**	**T**	A	A	-5
K50 DNA	5-	G	G	G	**A**	**T**	**T**	**A**	C	A	
		C	C	C	**T**	**A**	**A**	**T**	G	T	-5
1HDD DNA	5-	A	G	G	T	A	**A**	**T**	**T**	**A**	
		T	C	C	A	T	**T**	**A**	**A**	**T**	-5
		18	17	16	15	14	13	12	11	10	

S50WT simulation

The S50WT simulation started from the X-ray structure of the MAT α2 homeodomain-DNA complex (Wolberger et al. 1991). In the simulation, Ser50 was only briefly in hydrogen bond contact with bases A2 and T3. In accordance with the crystal structure, the serine was basically involved in van der Waals and water-mediated contacts with the DNA. The serine was in van der Waals contact with A2 and T3 during 30% and 87% of the simulation. Water molecules mediating hydrogen bonds from Ser50 to the specific NN bases are listed in Table 3. In addition to the DNA contacts, the hydroxyl group of the serine also formed a hydrogen bond with the Lys46 carbonyl oxygen at an occupancy of 66%.

K50WT and K50MUT simulations

Since, to our knowledge, there is no experimental wild-type homeodomain structure with a lysine at position 50, we used a structure of an engrailed Q50K mutant structure bound to a bicoid-specific DNA sequence (Tucker-Kellogg et al. 1997). In the K50WT simulation the Nζ hydrogen atoms of the lysine bound interchangeably to N7 and O6 atoms of the guanines. The contacts lasted through the entire simulation. Each of the hydrogen atoms participated in hydrogen bonds with the guanines for 33–49% of the simulation (Table 3). Water-mediated hydrogen bonds from the lysine to the guanine bases were much more scarce and of shorter duration. The Lys50 sidechain was not involved in any hydrogen bonds with the rest of the protein.

In the K50MUT simulation, where the DNA sequence is non-specific for lysine, the lysine bound most of the time to the N7 atom of base A2 and only occasionally to the O4 atom of base T3 (Table 3). After 1 ns the lysine turned away from the DNA to form hydrogen bonds with the backbone carbonyl of Lys46. Water-mediated hydrogen bonds from Lys50 to the DNA bases extended less than 10% of the simulation time.

In agreement with the experimental results (Ades and Sauer 1994; Hanes and Brent 1991; Treisman et al. 1989; Tucker-Kellogg et al. 1997), our simulations showed that lysine without doubt was able to distinguish the bicoid-specific 5′-TAATCC by hydrogen bonding to the 5′-GG bases on the complementary strand. Each guanine offers two hydrogen bond acceptors (atoms O6 and N7) whereas lysine has three hydrogen bond donors (the Hζ atoms). This combination maximizes the interaction possibilities and it provides both high affinity and specificity to the interaction. When simulating with a non-specific DNA sequence (the K50MUT simulation), the smaller number of hydrogen bond acceptors on the DNA made the Lys50-DNA contact less stable and the sidechain of Lys50 turned away to form a hydrogen bond with the protein backbone carbonyl of Lys46.

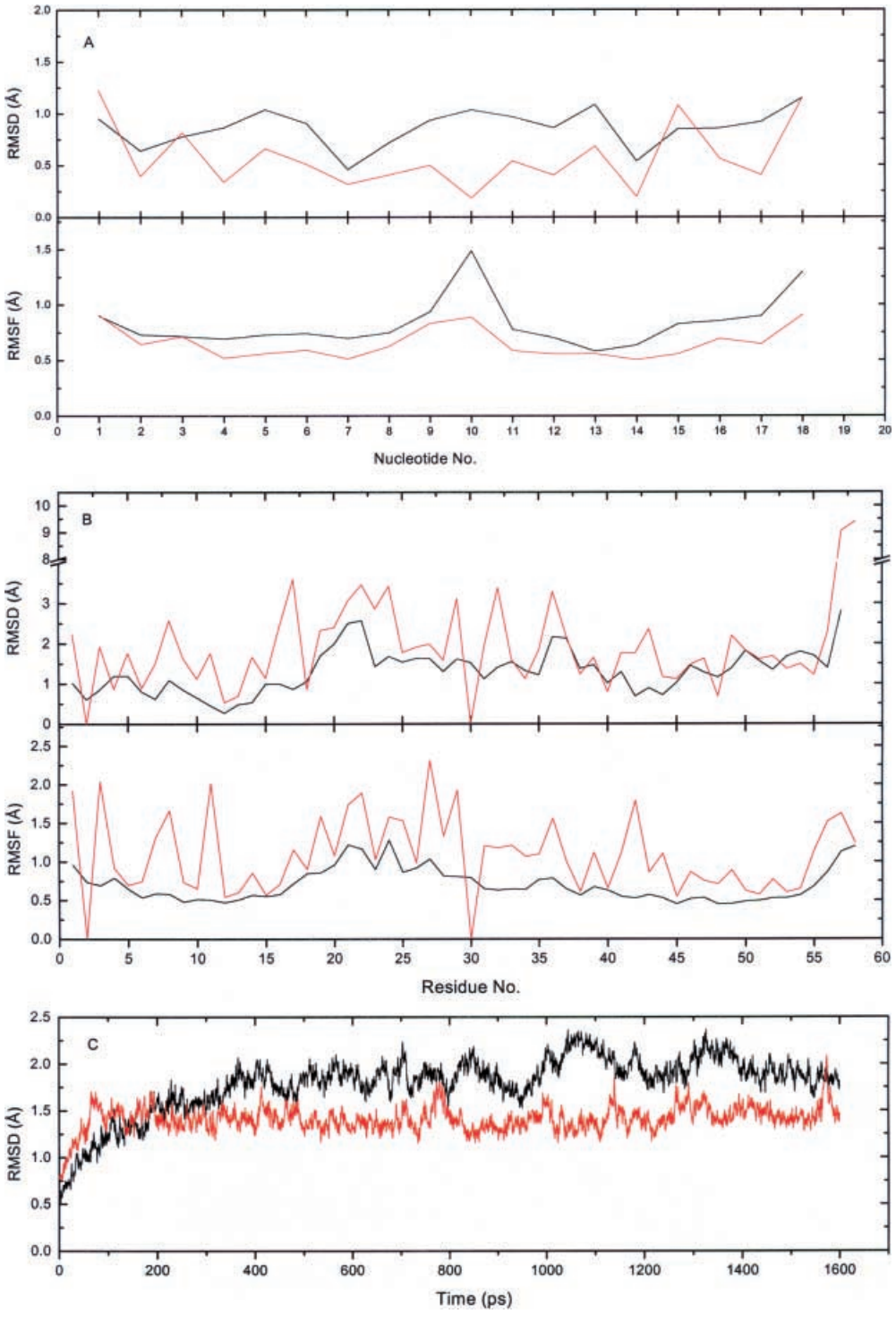


Fig. 1. **A** RMS deviation and fluctuation of DNA bases (*red*) and backbone (*black*) in the S50WT simulation. **B** RMS deviation and fluctuation of protein sidechains (*red*) and backbone (*black*) in the S50WT simulation. **C** Time dependence of RMS deviation of protein backbone (*black*) and DNA backbone (*red*) in the S50WT simulation

Q50WT and Q50MUT simulations

In the experimental homeodomain structures with a glutamine occupying position 50, there are no obvious direct contacts between this residue and the DNA (Billeter et al. 1993; Fraenkel and Pabo 1998; Fraenkel et al. 1998; Gruschus et al. 1999), but possibilities of contacts are discussed in the NMR and crystal structures of the antennapedia homeodomain-DNA complex (Billeter et al. 1993; Fraenkel and Pabo 1998). In the Q50WT simulation we observed direct hydrogen bonds from Oϵ of glutamine to the H42 atom of base C3 (39% occupancy) and from the Hϵ atoms to the phosphate group of G1 (28% occupancy). The Q50MUT simulation exhibited extensive direct hydrogen bonds from Hϵ of Gln50 to N7 at base A2 (Fig. 2) and a much weaker interaction from Hϵ to O4 of base T3. The occupancy of water-mediated hydrogen bonds was slightly higher

Table 3. The dominant direct or water-mediated contacts in each simulation with occupancy over 10% of the whole simulation time. The numbering of the bases corresponds to the numbering in Table 2

	Protein[a]	DNA	Occupancy (%)
S50WT	Ser50 Oγ*	A2 N7	44
	Ser50 Oγ*	A2 H61	14
	Ser50 Oγ*	A2 H62	11
	Ser50 Hγ*	A2 N7	11
	Ser50 Hγ*	T3 O4	12
K50WT	Lys50 Hζ1	G2 N7	12
	Lys50 Hζ1	G3 O6	12
	Lys50 Hζ1	G3 N7	10
	Lys50 Hζ2	G3 O6	12
	Lys50 Hζ2	G3 N7	13
	Lys50 Hζ3	G3 O6	21
	Lys50 Hζ3	G3 N7	19
	Lys50 Hζ1*	G2 O6	10
K50MUT	Lys50 Hζ1	A2 N7	26
	Lys50 Hζ2	A2 N7	14
	Lys50 Hζ3	A2 N7	17
Q50WT	Gln50 Oϵ	C3 H42	39
	Gln50 Hϵ21*	C2 H42	52
	Gln50 Oϵ*	G16 O6	17
Q50MUT	Gln50 Hϵ21	A2 N7	63
	Gln50 Hϵ21	T3 O4	11
	Gln50 Hϵ21*	A3 H61	15
	Gln50 Hϵ21*	A16 H61	42
C50WT	Cys50 Cβ #	T2 C5 M	22
	Cys50 Sγ #	T2 C5 M	55
C50MUT	Cys50 Hγ *	A2 N7	13
	Cys50 Sγ #	A2 N7	25
	Cys50 Sγ #	T3 C5 M	44
	Cys50 Sγ #	T3 O4	42
1HDD	Gln50 Hϵ21*	T14 O4	49
	Gln50 Hϵ22*	T13 O4	51
	Gln50 Hϵ22*	A5 N7	17
	Gln50 Hϵ22*	A5 H62	11

[a]Asterisk (*) indicates water-mediated hydrogen bonds; hash (#) indicates van der Waals contacts

(52% and 17% versus 42% and 15%) in the Q50WT simulation than in the Q50MUT simulation (Table 3). Only in Q50WT was the glutamine interacting with another protein residue sidechain, Lys46. The occupancy for hydrogen bonds involving each of the Hζ atoms on Lys46 was 10–15% (38% in total).

It seems that Gln50 strongly favors the MAT α2 DNA sequence because of the stable hydrogen bond between the Gln50 Hϵ atom and the N7 atom of base A2 in the Q50MUT simulation (Fig. 2). The DNA sequence used when solving the structure of the antennapedia homeodomain (9ant) was, however, not the consensus one. To investigate the behavior of glutamine in the presence of a consensus DNA sequence, a 1.4 ns simulation of engrailed homeodomain (1hdd) with its consensus DNA sequence was executed. There was no direct hydrogen bond between Gln50 and the DNA, but both the Nϵ hydrogen atoms bound to the (TAAT)TA sequence with high occupancy via water-mediated hydrogen bonds (Table 3). The Oϵ atom of the glutamine was locked in hydrogen bond with the Lys46 Nζ hydrogen atoms with occupancies of 56%, 23% and 15%. This hydrogen bond between the two sidechains was not present in the starting structure of the Q50WT and Q50MUT simulation simply because Lys46 in the MAT α2 structure adopts a different conformation compared to the engrailed structure. Because of the absence of this contact in the Q50MUT starting structure, the glutamine sidechain turned to make contact with the A2 base after about 400 ps. Thus the stable hydrogen bond between Gln50 and A2 observed in the Q50MUT simulation is an artifact due to an incorrect starting structure. Owing to the same reason, the hydrogen bond between Gln50 and the base C2 is broken by a rotation of the χ3 angle of Gln50 at 1 ns in the Q50WT simulation. Because the carbonyl oxygen atom of the Gln50 sidechain was hydrogen bonded to Lys46 throughout the 1HDD simulation in which the DNA sequence is the consensus one (TAATTN) (Gehring et al. 1994a; Hanes and Brent 1991), it is tempting to conclude that Lys46 acts as an anchor, locking Gln50 in a correct conformation for favorable water-mediated contacts with DNA bases (Fig. 3). In an extensive list of 387 homeodomain sequences (Gehring et al. 1994a), 326 have a glutamine at position 50. Of these 326 Gln50 homeodomains, only 41 do not have a lysine at position 46, and 25 of these have a glutamine at position 46; the other alternative residues at position 46 are: Ile, Ala, Val, Asn, Glu, Arg and Thr. With a few exceptions (one Ile, two Ala and two Val), the remaining residues all offer hydrogen bonding possibilities. This hydrogen bond seems to be conserved. Mutating Gln to Ala at this position, thus destroying the hydrogen bond, should change the DNA specificity of the Q50 homeodomain.

Grant et al. (2000) solved the engrailed Gln50Ala mutant structure and found that Lys46 is one of the three residues that adopts a different conformation. They also found that the movement of Lys46 and the nearby Arg31 seems to cause a widening of the DNA major groove by 2–2.5 Å. Thus, when removing Gln50 the Lys46 becomes less constrained and this may cause a widening of the major groove. This could be a destabilizing factor for the binding. By studying four experimentally determined homeodomain-DNA complex structures containing Lys46 and Gln50, they concluded that the Lys46-Gln50 contact is not conserved in three of them: HOX-1, ultrabithorax and antennapedia (Grant et al. 2000). However, none of these complexes includes the typical Q50 homeodomain conserved DNA sequence TAATTN.

C50WT and C50MUT simulations

Because of the less polar nature of cysteine compared to serine, there were even fewer water-mediated hydrogen bonds let alone direct hydrogen bonds between the cysteine and the bases in these systems. During 59% of the C50WT simulation the Hγ atom of cysteine was in hydrogen bond contact distance to the carbonyl oxygen of Lys46, whereas the same contact was present only 19% in the C50MUT simulation. The sidechain of Cys50 was in van der Waals contact with the methyl groups from T2 and T3 throughout the C50WT simulation. During

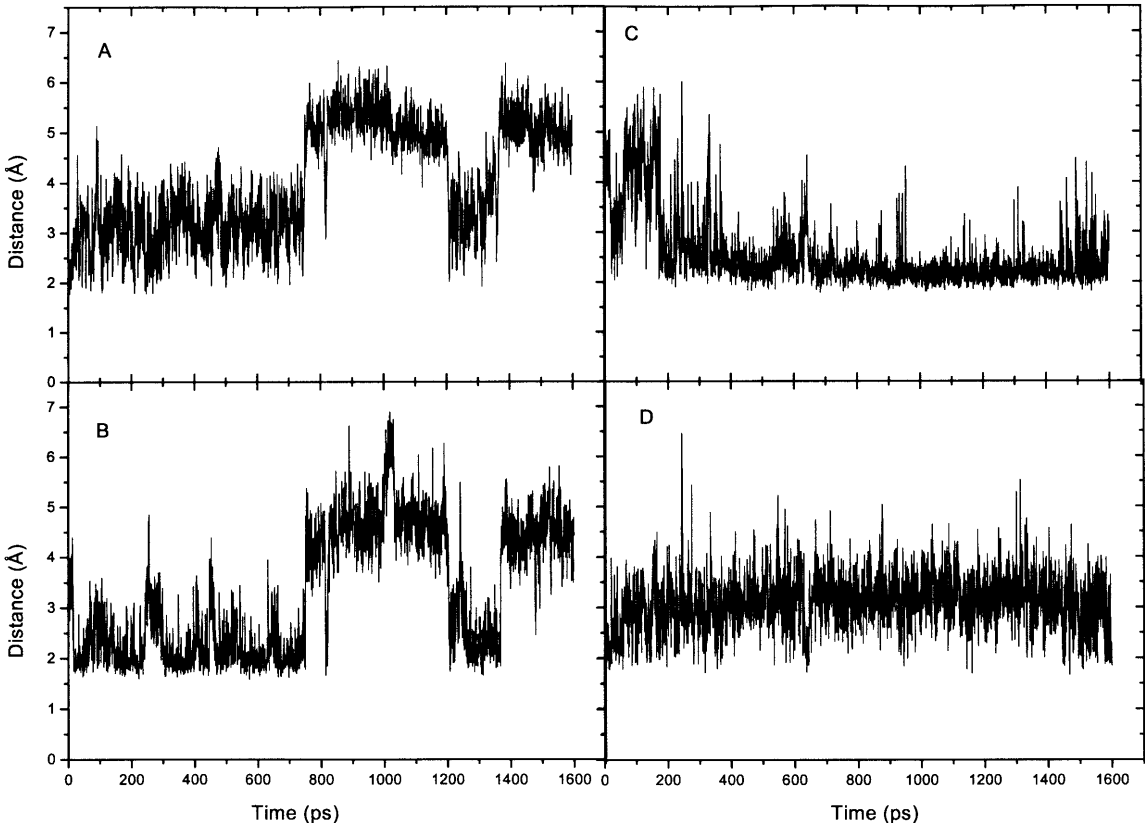
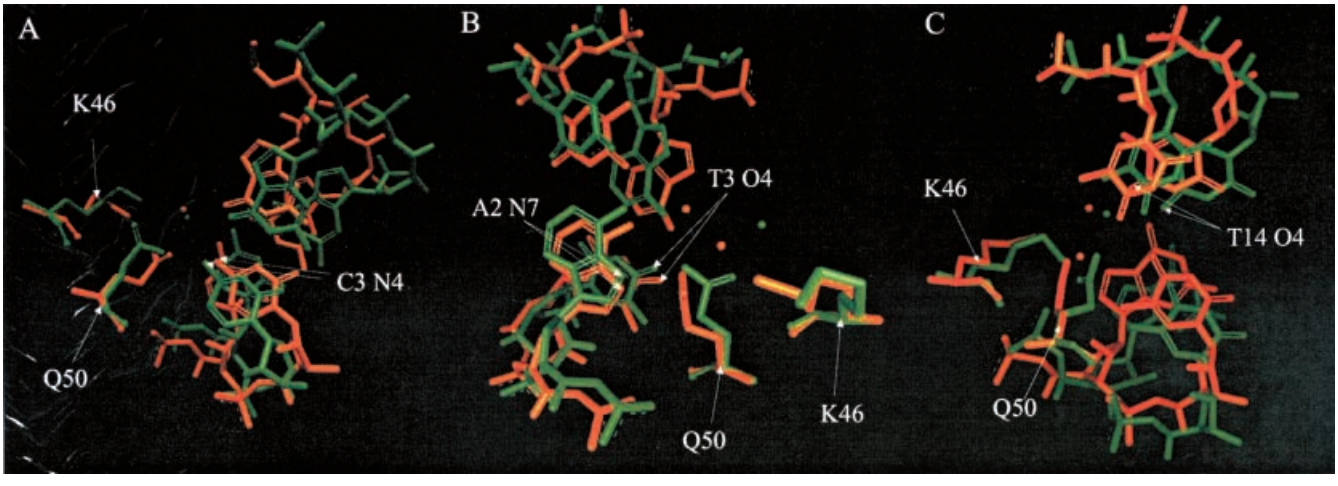


Fig. 2. Time dependence of distances between the Gln50 sidechain and NN bases in the Q50WT and Q50MUT simulations: **A** Q50WT simulation, Oϵ atom to C2 H42; **B** Q50WT simulation, Oϵ atom to C3 H42; **C** Q50MUT simulation, Hδ atom to A2 N7; **D** Q50MUT simulation, Hδ atom to T3 O4

800 ps in the C50MUT simulation, similar contacts to the A2 and T3 could also be noted. Figure 4 shows that the contact distances in C50WT were basically stable during the whole simulation while the contact distances fluctuated more in the C50MUT simulation, which indicates that Cys50 distinguishes the correct sequence with long-term stable van der Waals contacts. Experimental results show that the POU-homeodomain consensus sequence is 5′-TAATNA (Stepchenko et al. 1997; Verrijzer et al. 1992), which is very well reflected in our C50WT simulation where the Cys50 sidechain was interacting with T2 during the whole simulation but not with T3. This implies that the fifth base in the consensus sequence is likely a variable one. Various studies have shown that the C50 homeodomain alone is only capable of weak DNA binding, with low specificity (Ingraham et al. 1990; Verrijzer et al. 1990, 1992). This is explained in our simulations by the fact that the DNA interaction with Cys50 is based purely on van der Waals interactions.

Fig. 3. Two snapshots taken at the beginning (*yellow*) and after 1 ns (*green*) of the Q50WT (**A**), Q50MUT (**B**) and 1HDD (**C**) simulations



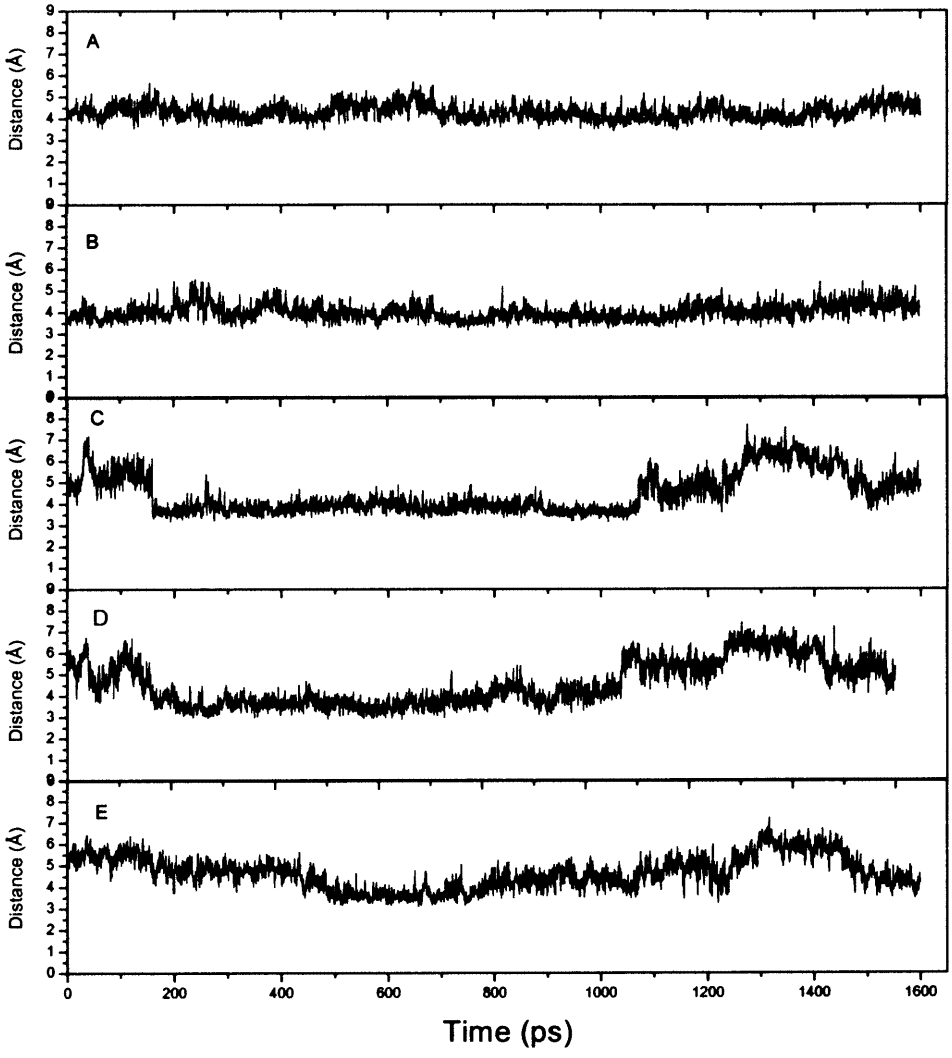


Fig. 4. van der Waals distance between Cys50 and T2 base in C50WT simulation, A2 and T3 bases in the C50MUT simulation: **A** C50WT simulation, Cys50 Cβ atom to T2 methyl carbon; **B** C50WT simulation, Cys50 Sδ atom to T2 methyl carbon; **C** C50MUT simulation, Cys50 Sδ atom to T3 methyl carbon; **D** C50MUT simulation, Cys50 Sδ atom to T3 O4 atom; and **E** C50MUT simulation, Cys50 Sδ atom to A2 N7 atom

As an additional test, we examined contacts from the NN bases in the (TAAT)NN sequence on both strands to any other residue in all seven simulations. We found that these two bases did not form any direct hydrogen bond to other protein sidechains except residue 50 and there were only a few water-mediated hydrogen bonds of noticeable occupancy in some of the simulations (Table 4). Evidently residue 50 is the one that determines the sequence of the NN bases. It is interesting to note that one of few water-mediated hydrogen bonds from the (TAAT)NN sequence went to Arg54, which has been suggested to participate in DNA sequence recognition. This contact was, however, short and not conserved in all simulations. In the S50WT simulation, Asn47 also contacted the (TAAT)NN sequence; this position is usually occupied by a highly conserved Ile and its importance has also been debated. The water-mediated hydrogen bonds in the S50WT simulation were not reproduced in the other simulations with the same DNA sequence (X50MUT), further supporting the notion that only residue 50 plays a non-negligible role in homeodomain DNA differentiation.

Table 4. Water-mediated hydrogen bonds from NN bases to protein sidechains other than residue 50. There were no other direct contacts from the NN bases to any other protein sidechain than residue 50. Occupancy cutoff is 10%. The numbering of the bases is the same as in Table 2

	DNA base	Protein	Occupancy (%)
S50WT	T3 O4	Arg54 HH21	12
	A16 H61	Asn47 Oδ1	34
	A16 H62	Asn47 Oδ1	10
K50WT	G3 N7	Arg54 HH11	34
K50MUT	–	–	–
Q50WT	–	–	–
Q50MUT	–	–	–
C50WT	T2 O4	Lys46 Hζ3	21
C50MUT	T3 O4	Arg54 HH12	13

Hydration of homeodomain-DNA complex

Because in a protein-DNA complex the interfacial solvent molecules, with few exceptions, hydrogen bond to both protein and DNA, we defined a water molecule to be interfacial if it was within 2.4 Å distance from both

the protein and the DNA. This is equivalent to an earlier definition (Nadassy et al. 1999). This definition is prone to slight overestimation because water molecules at the boundary of the interface will be counted. Table 5 shows the average number of water molecules in the protein-DNA interface, sampled every 5 ps. The hydration of the interface varied from 16 up to 40 interfacial water molecules in the snapshots, averaging around 30 for all simulations. This is in concord with earlier studies of high-resolution crystal structures of protein-DNA complexes (Jones et al. 1999; Nadassy et al. 1999), which suggested that a protein-DNA complex with an interface area of the MAT α2 homeodomain complex (3740 Å^2) should have about 25 water molecules on average. The number of water-mediated hydrogen bonds between the protein and DNA that were longer than 50 ps are also shown in Table 5. Most of the single water-mediated hydrogen bonding contacts lasted less than 50 ps, yet a few could be over 200 ps (Table 6). On average, nine water-mediated contacts (range 6–17) had an occupancy of more than 30%. The average lifetimes of these contacts were short (rarely over 60 ps), even for high occupancy (over 30%) water-mediated hydrogen bonds between protein and DNA, suggesting that the contacts constantly break and form in short intervals owing to the movements of the water molecules. The results hint that there are a few conserved hydration sites in the interface which mediate contacts, including hydrogen bonds and steric complementarity, between protein and DNA, but the water molecules in these sites are rarely locked and have high mobility. The RMS fluctuations of the side chains in the recognition helix and DNA bases are low, around 0.7 Å, implying that the flexibility is not caused by sidechain and base mobility.

In order to examine the solvent distribution in the protein-DNA interfaces, we plotted three-dimensional probability density maps of the simulations with specific DNA sequences (S50WT, C50WT, K50WT and 1HDD simulations) (Fig. 5). The density was highest around the DNA backbone phosphates. The density was not uniformly distributed in the major groove but was confined to defined clusters around the sidechains of the recognition helices. In both S50WT and 1HDD simulations, clusters of densities can be found between residue 50 and the (TAAT)NN bases. In the K50WT simulation there were fewer clusters because the Lys50 sidechain protruded into the interface and caused steric hindrance for high occupancy hydration sites. In the C50WT simulation this situation did not occur; rather, it seemed that the major groove was highly hydrated with several sites, confirming the hypothesis of few yet conserved hydration sites in the interfaces.

During the simulation time of 1.6 ns the majority of the water molecules had shorter residence times than 100 ps, but a few water molecules in the interface had long residence times, up to the whole simulation time. Even when prolonging the S50WT simulation to 3.6 ns we observed a 3.3 ns residence time. Of the total of 2908 water molecules in the system, 39% were at some point in contact with both DNA and protein. Of these, 80% had lifetimes less than 100 ps, and less than 0.5% had lifetimes more than 2 ns. This is in agreement with earlier NMR measurements in which a few hydrating

Table 5. The average number of interfacial water molecules in each system and the amount of those who mediate hydrogen bonds longer than 50 ps

	Average number of interfacial waters ($\pm$SD)	Number of water-mediated hydrogen bonds longer than 50 ps
S50WT	32 $\pm$ 4	26
C50WT	29 $\pm$ 4	24
C50MUT	32 $\pm$ 4	36
Q50WT	32 $\pm$ 4	45
Q50MUT	28 $\pm$ 3	48
K50WT	30 $\pm$ 4	83
K50MUT	30 $\pm$ 4	30

Table 6. The top five long-lived water-mediated hydrogen bonds in each simulation. The numbering of the bases is the same as in Table 2

Sidechain			Bases		Water	Lifetime
S50WT						
Arg54	Hϵ	–	G4	O1P	519	520
Arg54	Hϵ	–	G4	O1P	519	435
Arg1	O	–	T13	O3′	360	140
His3	O	–	A14	O2P	2125	125
Phe5	O	–	T13	O2P	853	100
K50WT						
Gln44	O	–	A14	O2P	2244	275
Arg53	HH22	–	G1	O1P	1624	200
Arg53	HH22	–	G1	O1P	1624	180
Trp48	O	–	A13	O2P	1212	175
Trp48	O	–	A13	O2P	868	165
Q50WT						
Arg53	HH22	–	G1	O1P	680	220
Gln50	Hϵ22	–	C2	O1P	1377	160
Gln44	O	–	A14	O2P	2822	155
Arg53	HH22	–	G1	O1P	877	150
Arg54	HH21	–	A4	O2P	1491	150
C50WT						
Gln44	O	–	A14	O2P	907	215
Gln44	O	–	A14	O2P	907	125
Arg4	HH11	–	A13	O4′	1823	110
Arg52	HH12	–	A13	O2P	2124	105
Trp48	O	–	A13	O2P	1845	105
K50MUT						
Arg53	HH12	–	C1	O2P	288	240
Gln44	O	–	A14	O1P	1289	135
Arg54	Hϵ	–	G4	O1P	2536	125
Lys50	Hζ3	–	A2	O2P	495	125
Asn47	Hδ21	–	C15	O1P	1995	120
Q50MUT						
Leu26	HN	–	C1	O2P	21	295
Gln44	O	–	A14	O1P	2303	225
Asn47	Hδ22	–	A14	O1P	2261	185
Gln44	O	–	A14	O1P	2303	180
Gln44	O	–	A14	O1P	2303	175
C50MUT						
Phe5	HN	–	T13	O2P	1851	250
Gln44	O	–	A14	O1P	899	175
Asn51	Hδ22	–	T13	O1P	1317	130
Arg54	Hδ	–	G4	O1P	2472	100
Arg4	HH12	–	T13	O4′	898	95

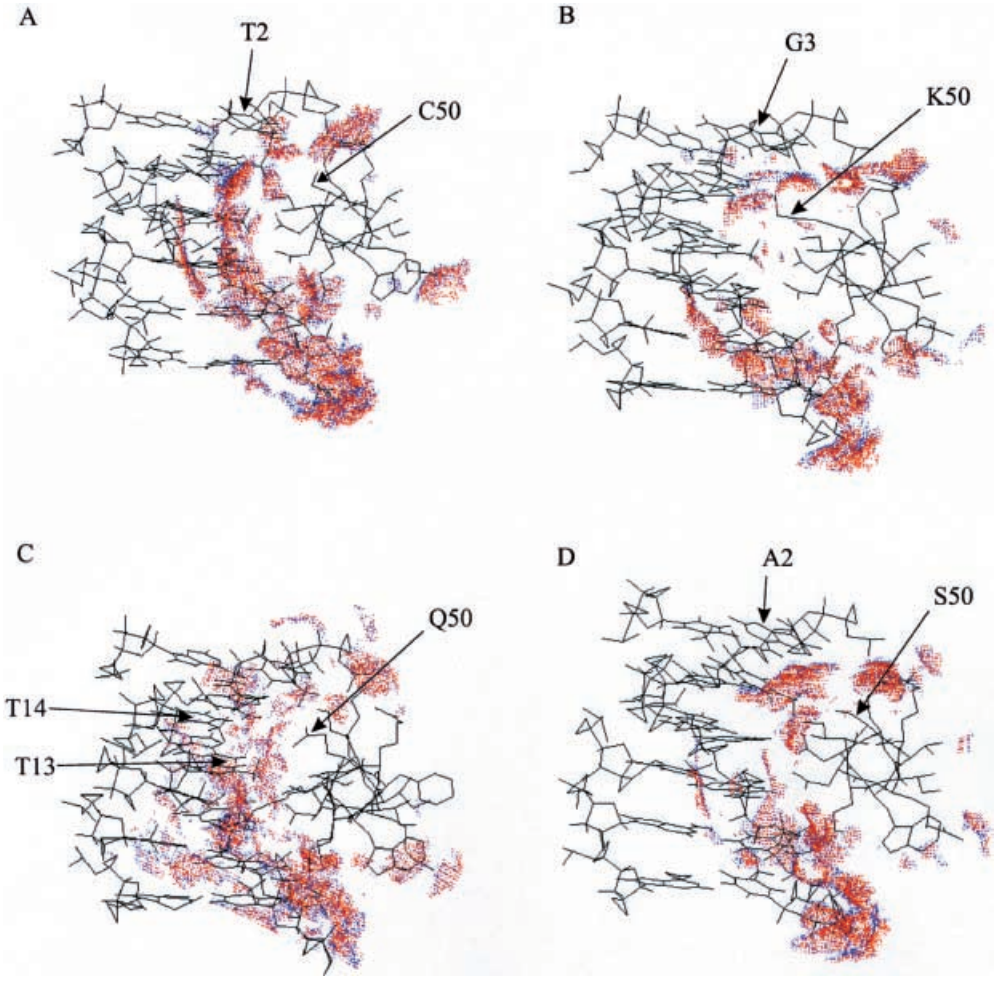


Fig. 5A–D. 3D probability density map of the solvent distributions in the protein-DNA interfaces. Only the recognition helices and the TAATNN DNA sequences are shown. The figures are oriented in such way that we are looking down along the axis of the recognition helix. The *blue color* indicates a σ value of 1.0 and the *red color* marks a σ value of 1.5. **A** C50WT simulation; **B** K50WT simulation; **C** 1HDD simulation; and **D** S50WT simulation

water molecules with 1 ns to millisecond residence times were found (Qian et al. 1993). In Fig. 6 we illustrate the path of a single water molecule that spends around 1700 ps in the protein-DNA interface in the S50WT simulation.

Upon binding, the protein sidechains and DNA bases in the binding interface become more rigid because of the intermolecular interactions between them. Recently, Jayaram et al. (2002) calculated the free energy components of 40 protein-DNA complexes and estimated the entropy (translational, rotational, vibrational and configurational entropies) loss of formation of the MAT α2 homeodomain-DNA complex to be about 47 kcal mol^{-1} ($-T\Delta S$). Dunitz (1994) concluded that a single water molecule bound to a protein cavity contributes no less than about 10 cal mol^{-1} K^{-1} (corresponding to ice and crystalline salts) and no more than 16–17 cal mol^{-1} K^{-1} (corresponding to liquid water) in entropy, i.e. the difference is about 2 kcal mol^{-1} at 300 K. The homeodomain-DNA interface was hydrated by 30 water molecules on average and no more than two of them in each simulation were locked in a fixed position more than 200 ps (Table 6). Thus, these molecules were only weakly bound in position. Using MD (Garcia and Hummer 2000), free energy calculations (Zhang and Hermans 1996) and NMR measurements (Denisov et al. 1997), it was earlier found that water molecules in protein cavities, despite extensive hydrogen bonding, have almost the same entropy as bulk water. Thus the entropy contribution of the 30 mobile interfacial water molecules is about 60 kcal mol^{-1} compared to a situation where they would be kept rigidly in place.

Conclusions

By simulating homeodomain-DNA systems we have found that the protein-DNA recognition is dependent not only on direct and water-mediated indirect read-out between the protein sidechains and the DNA bases, but also on the arrangement of the nearby protein sidechains. We found that the residue at position 50 in the homeodomain is the most important for DNA recognition. The interfacial water molecules occupy several conserved hydration sites and mediate contacts between the protein and the DNA. Some water molecules may have lifetimes in the nanosecond scale in the interface but in general the majority of the water molecules are mobile, which means that the entropic cost of keeping them in the interface region is low. The contacts formed

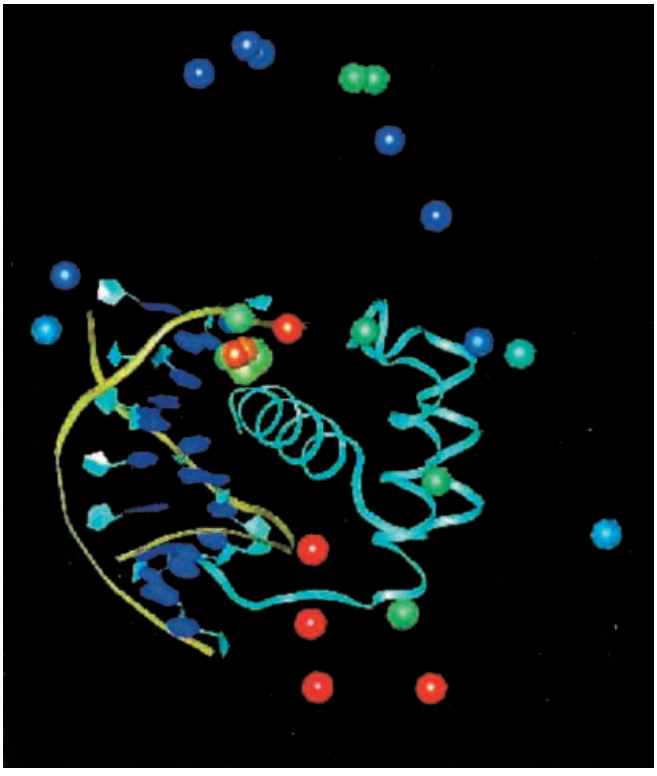


Fig. 6. Snapshots taken every 100 ps from S50WT simulation of 3.6 ns in total to illustrate the trajectory of a single water molecule. The *colored spheres* represent the water molecule. The color scale goes from *red* via *green* to *blue*. After about 400 ps, this water molecule spends 1.7 ns between Ser50, Arg53 and bases A2 and T3 before leaving the interface

between residue 50 and the DNA generally involve the DNA bases at the NN positions in the (TAAT)NN consensus recognition sequence, whereas most of the contacts between the DNA and other residues involve the DNA backbone rather than the base atoms.

Acknowledgements This work was supported by the Swedish Research Council and the Magnus Bergvall Foundation.

References

Ades SE, Sauer RT (1994) Differential DNA-binding specificity of the engrailed homeodomain: the role of residue 50. Biochemistry 33:9187–9194

Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. J Chem Phys 81:3684–3690

Bernstein FC, Koetzle TF, Williams GJ, Meyer EEJ, Brice MD, Rogers JR, Kennard O, Shimanouchi T, Tasumi M (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 112:535–542

Billeter M, Qian YQ, Otting G, Muller M, Gehring W, Wuthrich K (1993) Determination of the nuclear magnetic resonance solution structure of an antennapedia homeodomain-DNA complex. J Mol Biol 234:1084–1093

Brook B, Bruccoleri R, Olafson B, States D, Swaminathan S, Karplus M (1983) Charmm: a program for macromolecular energy, minimization and dynamic calculations. J Comput Chem 4:187–217

Brünger A, Karplus M (1988) Polar hydrogen positions in proteins: empirical energy placement and neutron diffraction comparison. Proteins 4:148–156

Denisov VP, Venu K, Peters J, Dietrich Hörlein H, Halle B (1997) Orientational disorder and entropy of water in protein cavities. J Phys Chem B 101:9380–9389

Dunitz JD (1994) The entropic cost of bound water in crystals and biomolecules. Science 264:670

Foloppe N, MacKerell AD (2000) All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data. J Comput Chem 21:86–104

Fraenkel E, Pabo CO (1998) Comparison of X-ray and NMR structures for the antennapedia homeodomain-DNA complex. Nat Struct Biol 5:692–697

Fraenkel E, Rould MA, Chambers KA, Pabo CO (1998) Engrailed homeodomain-DNA complex at 2.2 Å resolution: a detailed view of the interface and comparison with other engrailed structures. J Mol Biol 284:351–361

Garcia AE, Hummer G (2000) Water penetration and escape in proteins. Proteins 38:261–272

Gehring WJ, Affolter M, Burglin T (1994a) Homeodomain proteins. Annu Rev Biochem 63:487–526

Gehring WJ, Qian YQ, Billeter M, Furukubo-Tokunaga K, Schier AF, Resendez-Perez D, Affolter M, Otting G, Wuthrich K (1994b) Homeodomain-DNA recognition. Cell 78:211–223

Grant RA, Rould MA, Klemm JD, Pabo CO (2000) Exploring the role of glutamine 50 in the homeodomain-DNA interface: crystal structure of engrailed (Gln50→Ala) complex at 2.0 Å. Biochemistry 39:8187–8192

Gruschus JM, Tsao DH, Wang LH, Nirenberg M, Ferretti JA (1999) The three-dimensional structure of the vnd/NK-2 homeodomain-DNA complex by NMR spectroscopy. J Mol Biol 289:529–545

Hanes SD, Brent R (1991) A genetic model for interaction of the homeodomain recognition helix with DNA. Science 251:426–430

Haran TE, Joachimiak A, Sigler PB (1992) The DNA target of the trp repressor. EMBO J 11:3021–3030

Ingraham HA, Flynn SE, Voss JW, Albert VR, Kapiloff MS, Wilson L, Rosenfeld MG (1990) The POU-specific domain of Pit-1 is essential for sequence-specific, high affinity DNA binding and DNA-dependent Pit-1-Pit-1 interactions. Cell 61:1021–1033

Janin J (1999) Wet and dry interfaces: the role of solvent in protein-protein and protein-DNA recognition. Struct Fold Des 7:R277–R279

Jayaram B, McConnell K, Dixit SB, Das A, Beveridge DL (2002) Free-energy component analysis of 40 protein-DNA complexes: a consensus view on the thermodynamics of binding at the molecular level. J Comput Chem 23:1–14

Jones S, van Heyningen P, Berman HM, Thornton JM (1999) Protein-DNA interactions: a structural analysis. J Mol Biol 287:877–896

Jorgensen WL, Chandraseklar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–935

Kissinger CR, Liu BS, Martin-Blanco E, Kornberg TB, Pabo CO (1990) Crystal structure of an engrailed homeodomain-DNA complex at 2.8 Å resolution: a framework for understanding homeodomain-DNA interactions. Cell 63:579–590

Klemm JD, Rould MA, Aurora R, Herr W, Pabo CO (1994) Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA-binding modules. Cell 77:21–32

Kono H, Sarai A (1999) Structure-based prediction of DNA target sites by regulatory proteins. Proteins 35:114–131

Kornberg TB (1993) Understanding the homeodomain. J Biol Chem 268:26813–26816

MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C,

Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 102:3586–3616

Mandel-Gutfreund Y, Schueler O, Margalit H (1995) Comprehensive analysis of hydrogen bonds in regulatory protein-DNA complexes: in search of common principles. J Mol Biol 253:370–382

Nadassy K, Wodak SJ, Janin J (1999) Structural features of protein-nucleic acid recognition sites. Biochemistry 38:1999–2017

Otwinowski Z, Schevitz RW, Zhang RG, Lawson CL, Joachimiak A, Marmorstein RQ, Luisi BF, Sigler PB (1988) Crystal structure of trp repressor/operator complex at atomic resolution. Nature 335:321–329 [erratum: (1988) Nature 335:837]

Pabo CO, Nekludova L (2000) Geometric analysis and comparison of protein-DNA interface: why is there no simple code for recognition? J Mol Biol 301:597–624

Qian YQ, Otting G, Wuthrich K (1993) NMR detection of hydration water in the intermolecular interface of a protein-DNA complex. J Am Chem Soc 115:1189–1190

Sali A, Blundell T (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779–815

Scott MP, Tamkun JW, Hartzell GWd (1989) The structure and function of the homeodomain. Biochim Biophys Acta 989:25–48

Stepchenko AG, Luchina NN, Pankratova EV (1997) Cysteine 50 of the POU H domain determines the range of targets recognized by POU proteins. Nucleic Acids Res 25:2847–2853

Suzuki M, Gerstein M (1995) Binding geometry of alpha-helices that recognize DNA. Proteins 23:525–535

Treisman J, Gonczy P, Vashishtha M, Harris E, Desplan C (1989) A single amino acid can determine the DNA binding specificity of homeodomain proteins. Cell 59:553–562

Tucker-Kellogg L, Rould MA, Chambers KA, Ades SE, Sauer RT, Pabo CO (1997) Engrailed (Gln50→Lys) homeodomain-DNA complex at 1.9 Å resolution: structural basis for enhanced affinity and altered specificity. Structure 5:1047–1054

Tullius T (1995) Homeodomains: together again for the first time. Structure 3:1143–1145

van Gunsteren WF, Berendsen HJC (1977) Algorithms for macromolecular dynamics and constraint dynamics. Mol Phys 34:1311–1327

Verrijzer CP, Kal AJ, Van der Vliet PC (1990) The DNA binding domain (POU domain) of transcription factor Oct-1 suffices for stimulation of DNA replication. EMBO J 9:1883–1888

Verrijzer CP, Alkema MJ, van Weperen WW, Van Leeuwen HC, Strating MJ, van der Vliet PC (1992) The DNA binding specificity of the bipartite POU domain and its subdomains. EMBO J 11:4993–5003

Vershon AK, Jin Y, Johnson AD (1995) A homeodomain protein lacking specific side chains of helix 3 can still bind DNA and direct transcriptional repression. Genes Dev 9:182–192

Wolberger C, Vershon AK, Liu B, Johnson AD, Pabo CO (1991) Crystal structure of a MAT alpha 2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. Cell 67:517–528

Zhang L, Hermans J (1996) Hydrophilicity of cavities in proteins. Proteins 24:433–438