10270–10283 *Nucleic Acids Research, 2017, Vol. 45, No. 17* *Published online 18 July 2017*
doi: 10.1093/nar/gkx627

Decomposing protein–DNA binding and recognition using simplified protein models

Loïc Etheve, Juliette Martin and Richard Lavery*

MMSB UMR 5086 CNRS / Univ. Lyon I, Institut de Biologie et Chimie des Protéines, 7 passage du Vercors, Lyon 69367, France

Received April 12, 2017; Revised July 05, 2017; Editorial Decision July 07, 2017; Accepted July 11, 2017

ABSTRACT

We analyze the role of different physicochemical factors in protein/DNA binding and recognition by comparing the results from all-atom molecular dynamics simulations with simulations using simplified protein models. These models enable us to separate the role of specific amino acid side chains, formal amino acid charges and hydrogen bonding from the effects of the low-dielectric volume occupied by the protein. Comparisons are made on the basis of the conformation of DNA after protein binding, the ionic distribution around the complex and the sequence specificity. The results for four transcription factors, binding in either the minor or major grooves of DNA, show that the protein volume and formal charges, with one exception, play a predominant role in binding. Adding hydrogen bonding and a very small number of key amino acid side chains at the all-atom level yields results in DNA conformations and sequence recognition close to those seen in the reference all-atom simulations.

INTRODUCTION

The very first article published in the Journal of Molecular Biology by Zubay and Doty (1) dealt with an early study of what is now termed chromatin. This article, which appeared less than a decade after Pauling, Corey and Branson's structure of the α-helix (2) and Watson and Crick's model of DNA (3), already addressed the problem of how proteins bind to DNA. Before the structure of any protein–DNA complexes had been resolved, Zubay and Doty pointed out that an α-helix could fit into the major groove of DNA and suggested that, since histones are largely α-helical, this type of interaction could be important for protein–DNA binding. Although this speculation was not correct in the case of the chromatin fiber, it did predict a common binding motif, that we now know is employed by many DNA-binding proteins.

Beyond the dimensions of the DNA grooves, it was also remarked early on that the disposition of hydrogen bonding groups made it easier to distinguish the four possible base pairs (AT, TA, GC, CG) from one another via the, sterically more accessible, major groove (since in the minor groove the acceptor and donor atoms of AT and TA are similarly located, as are those of GC and CG) (4). The importance of hydrogen bonding, and, in particular, bidentate bonding, between amino acids and nucleic acid bases was also stressed in attempts at finding a simple protein–DNA recognition code, although the authors of this proposition also pointed out the potential importance of DNA conformation (5).

That conformation could be important was highlighted by the first single crystal structure of DNA that showed sequence-dependent local changes in structure, or, more subtly, in deformability, susceptible to be 'read' by proteins (6). This finding led to the notion of so-called 'indirect' recognition, as opposed to 'direct' recognition involving hydrogen bonding and steric fit. The existence of such indirect recognition has since been demonstrated in many cases, ranging from the impact of pre-bending on binding the TATA-box protein (7), to the specific binding of the 434 repressor, despite the absence of direct hydrogen bonding from the protein to the DNA bases (8). Many other studies have built on and refined the conformational mechanisms involved (9–16).

Another clearly important factor in protein–DNA interaction is electrostatics, given the charge density of double-stranded DNA. Returning to histone–DNA binding, Mirzabekov and Rich suggested that DNA could be induced to curve towards the histone core as a result of the imbalance in phosphate-phosphate repulsion caused by the neutralization of the phosphates on the face of DNA contacting the positively charged histones (17). The Maher group confirmed such bending experimentally by creating patches of neutralized phosphates (18,19). Another concept resulted from simulations by Elcock and McCammon who showed that putting a low dielectric volume (mimicking a protein) in contact with DNA would shield one face of DNA from water, lower the local dielectric, increase phosphate-phosphate repulsion, again induce a Coulombic imbalance, but, in this

*To whom correspondence should be addressed. Tel: +33 472722637; Fax: +33 472722601; Email: richard.lavery@ibcp.fr

case, causing DNA to bend away from the bound protein (20).

Experimentally it is not easy to deconvolute the different factors that combine to enable a protein to bind to DNA and preferentially select a given base sequence. Molecular simulation, despite necessary approximations, offers a way to do this in a controlled manner. Based on the remarks above, we decided to try and understand the role of different factors in protein binding and recognition using simplified models of proteins, that enable us to separate out factors such as the impact of a low-dielectric volume, point charges, hydrogen bonding and coarse-grain versus all-atom amino acid representations. Using four different transcription factors, two binding in the minor groove of DNA and two in the major groove, we asked to what extent each protein model could reproduce the behavior seen in conventional all-atom molecular dynamics simulations of the corresponding protein–DNA complexes, using a physiologically reasonable water and salt environment. The factors we looked at included DNA conformation, the ion distribution around the complexes and the sequence-recognition. In order to do this we used the conformational and environmental analysis tools provided by Curves+, Canal and Canion (21,22) and the sequence threading analysis ADAPT (11,23,24).

It should be noted that our protein models are rigid, in contrast to the proteins in the all-atom simulations. Consequently, if flexibility is an important factor, as in the case of flexible protein tails (25,26), protein–protein interfaces, or of the protein surface interacting with DNA (27,28) then this will be seen only as the failure of our most detailed models to reproduce the all-atom results.

MATERIALS AND METHODS

Protein structures

The starting configurations of the four protein–DNA complexes were obtained from the Protein Data Bank (29): the crystal structure of human TATA-box binding protein (TBP) (1CDW) (30), resolution 1.9 Å), the NMR structure of human sex-determining Y protein (SRY) (PDB code: 1J46 (31)), the crystal structure of skinhead-1 protein from *C elegans* (SKN) (PDB code: 1SKN (32), resolution 2.5 Å) and the crystal structure of the lambdoid bacteriophage P22 c2 repressor (P22) (PDB code: 2R1J (33), resolution 1.53 Å). The JUMNA program (34) was used when necessary to complete the single-stranded ends of DNA oligomers and to construct complexes with modified base sequences corresponding to those studied experimentally. The length of the oligomers, their sequence, the location and numbering of the binding site are given in the following section. The simplified protein models discussed below were built on the basis of the experimental coordinates, with two exceptions. For SKN, our previous work showed that this protein can adopt several different conformational substates with different sequence selectivities (27). We therefore chose to use the average molecular dynamics conformation of the most common substrate (cluster 2/4 in reference (27)) for the present studies. For P22, our earlier work also pointed to the importance of the Gln 37 residue within each protein monomer in determining sequence selectivity. We therefore modified our models to include two distinct side chain Gln 37 conformations (see Results section).

Simplified protein models (SPMs)

In order to analyze the main elements of protein–DNA binding, we have developed simplified protein models that can include from one to four separate features. The simplest protein representation consists of a rigid, uncharged, low-dielectric volume (V) reproducing the repulsion/dispersion and solvent exclusion properties of the protein. This model is formed of pseudoatoms representing each amino acid. We have based our coarse grain representation on that developed by Zacharias (35), with one pseudoatom placed at Cα, one for the side chains of Ala, Asn, Asp, Cys, Ile, Leu, Pro, Ser, Thr, Val at the geometrical center of the side chain heavy atoms, and two for the side chains of Arg, His, Gln, Glu, Lys, Met, Phe, Trp, Tyr, one at the center of the Cβ-Cγ bond, and the other at the geometrical center of the remaining heavy atoms. This representation has been modified to include supplementary pseudoatoms at the position of the formally charged oxygen or nitrogen atoms of Arg, Asp, Glu and Lys. The Lennard-Jones parameters for the V representation were obtained by analyzing the effective radii of the pseudoatoms using coordinates from 100 crystallographic protein/DNA complexes and then adjusting a common energy well-depth to reproduce the corresponding Lennard-Jones interactions energy for each complex (see Supplementary Table S1 and Figure S1).

The V protein model was held rigid using strong quadratic distance restraints between its pseudoatoms: first, linking the Cα and side chain pseudoatoms along the peptide chain, and second, linking each pseudoatom to three others, avoiding choices that would lead these restraints to be close to being coplanar or to being aligned.

The second protein model feature consists of adding formal charges (C) to the V model. This involves placing ±1.0e or ±0.5e charges on the terminal oxygen or nitrogen atoms of Arg, Asp, Glu and Lys and also on the terminal backbone pseudoatoms of the peptide chains.

The third feature includes the possibility of hydrogen bonding (H) by adding atomic-scale amino groups to Asn, Gln, carbonyl oxygens to Asn and Gln and hydroxyl groups to Ser, Thr and Tyr. In this case, the added atoms carry partial charges taken from the AMBER ff99SB force field (36). The total charge on each side chain was neutralized by adding an equal and opposite net charge to the atom carrying the hydrogen bonding groups.

The fourth feature adds the possibility of representing a certain number of key amino acid side chains at the all-atom level (S). In this case, the side chains are flexible and all their atoms carry partial charges, again taken from the AMBER ff99SB force field.

The simple V model can be combined with any of the other features, but here will mainly consider a progression from V to VC (adding formal charges), to VCH (adding hydrogen bonding) and finally to VCHS (adding selected all-atom side chains).

Molecular dynamics simulations

The simplified model protein/DNA complexes were solvated with SPC/E water molecules (37) within a truncated octahedral box, ensuring a solvent shell of at least 10 Å around the solute. The solute was neutralized with K^+ ions and then sufficient K^+Cl^- ion pairs were added to reach a salt concentration of 150 mM. The ions were initially placed at random, but at least 5 Å away from DNA and 3.5 Å away from one another. Molecular dynamics (MD) simulations were performed with the AMBER 12 suite of programs (38,39) using PARM99 parameters (40) and the bsc0 modifications (41) for the solute and Dang parameters (42) for the surrounding ions. Simulations employed periodic boundary conditions and electrostatic interactions were treated using the particle mesh Ewald algorithm (43) with a real space cutoff of 9 Å. Lennard–Jones interactions were truncated at 9 Å. A pair list was built with a buffer region and a list update was triggered whenever a particle moved by more than 0.5 Å with respect to the previous update.

Each system was initially subjected to energy minimization with harmonic restraints on the position of all solute atoms. The system was then heated to 300 K at constant volume during 100 ps. The position restraints were slowly relaxed during a series of energy minimizations (500 steps of steepest descent and 500 steps of conjugate gradient) followed by 50 ps of equilibration without position restraints. The 500 ns production simulations were carried out at constant temperature (300 K) and pressure (1 bar) with a 2 fs time step. During these simulations, pressure and temperature were maintained using the Berendsen algorithm (44) with a coupling constant of 5 ps and SHAKE constraints (45) were applied to all bonds involving hydrogen.

For comparison purposes, all-atom simulations using an identical protocol were performed for the four protein/DNA complexes and for corresponding isolated DNA molecules. These simulations generated trajectories of 500 ns trajectories (1 μs in the case of P22 c2 complex). Conformational snapshots were saved for further analysis every 5 ps for simulations with the simplified protein models and every 1 ps for the all-atom simulations.

Binding specificity analysis

Binding specificity was determined using the so-called ADAPT sequence threading approach (11,23) implemented within the JUMNA program (34). ADAPT consists of calculating the binding energy of a protein–DNA complex by substituting all possible DNA base sequences into the binding site of a given conformation of the complex. After each sequence change, the DNA binding site and the protein interface (for the all-atom proteins, but only for the all-atom side chains in the case of the simplified protein models) are energy minimized and compared with the energy of the isolated components (the isolated DNA having the same base sequence). ADAPT calculations are accelerated by a divide-and-conquer technique, breaking each sequence down into overlapping 5 bp fragments, dramatically reducing the total number of calculations, without significant loss of accuracy (24). The resulting binding energies are finally used to generate a position weight matrix (PWM) describing the binding specificity. This result can be further broken down into components linked with protein–DNA interaction (so-called direct recognition) and DNA deformation induced during complex formation (indirect recognition).

Here ADAPT was applied to uniformly sampled snapshots derived from the MD trajectories (typically 10–25 per trajectory) described above, after a brief Cartesian coordinate energy minimization to remove bond length and base plane deformations (using the AMBER force field and a simple distance dependent dielectric and reduced phosphate charges to model the environment (11)). Because of these simplifications, the single PWM representing a given trajectory is based on sequence-dependent energy differences with respect to the minimum energy for each snapshot (27,28). Each column j of the PWM matrices is composed of the frequencies (f_{ij}) of each amino acid i at this position along the protein binding site. For visualization, the matrices are represented as sequence logos, where the height of each column j is equal to the information content of the column, $\log_2 4 + \sum_i f_{ij} * \log_2 f_{ij}$. These results were represented graphically using the WebLogo software (46). PWMs from all-atom trajectories are compared with those from the simplified model protein and with experimental results using the Pearson correlation coefficient and the raw PWM frequencies. Experimental PWMs are taken from the Transfac database (47) or, in the case of P22, deduced from relative dissociation constants resulting from the base substitution studies of Watkins *et al.* (33).

Conformational, environmental and sequence analysis

Average DNA conformations of the simulated protein/DNA complexes were analyzed with the Curves+ program and the Canal utility (21), while ion distributions were calculated as local molarities using Canion utility (22,48). For details see the corresponding publications and http://curvesplus.epfl.ch. The protein–DNA interfaces were analyzed with an in-house utility program, using Pauling van der Waals radii, ignoring hydrogen atoms and using a water probe radius of 1.4 Å.

RESULTS AND DISCUSSION

As we have remarked in earlier publications, the ADAPT analysis of protein–DNA sequence selectivity based on all-atom MD trajectories is able to yield results in good agreement with experiment for a variety of proteins (11,14,24). For the four protein complexes we study here this is also the case, as shown by the comparison presented in Supplementary Table S2 and Figure S2. The Pearson correlation coefficients between the MD and experimental PWM results are between 0.60 and 0.86, the lowest value belonging to SKN, which also has smallest protein–DNA interface (see discussion section). We also remark that for P22 we do not see any sequence recognition for the 4 bp spacer between the L and R half-sites, whereas experimentally an A/T preference is observed in the center of the spacer. As noted earlier (28), this sequence preference appears to be linked to cations bound at the center of the protein–DNA interface, observed in the MD simulations, but that we cannot treat at present with the ADAPT method (see also the discussion of P22 binding below).

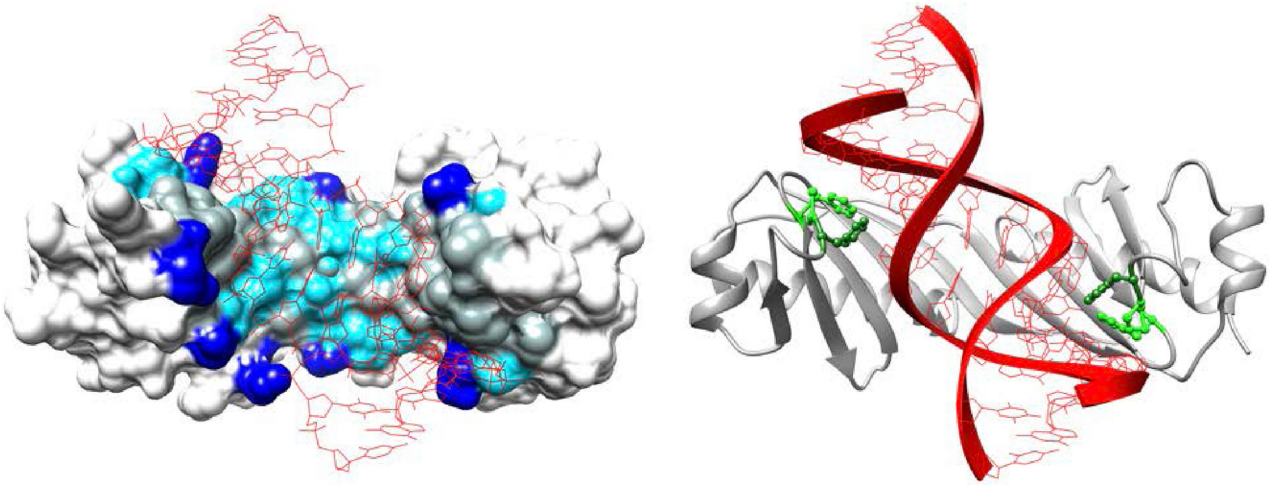

Figure 1. Visualization of the TBP/DNA complex. Left: the protein interface (charged residues: blue, polar residues: cyan, hydrophobic residues: dark gray). Right: ribbon representation of the complex. Key amino acid side chains are shown (intercalating residues Phe 193 and Phe 284: dark green, supporting residues Phe 210, Phe 301, Ala 194 and Pro 285: light green). In both panels, DNA is closest to the viewer and is shown in red.

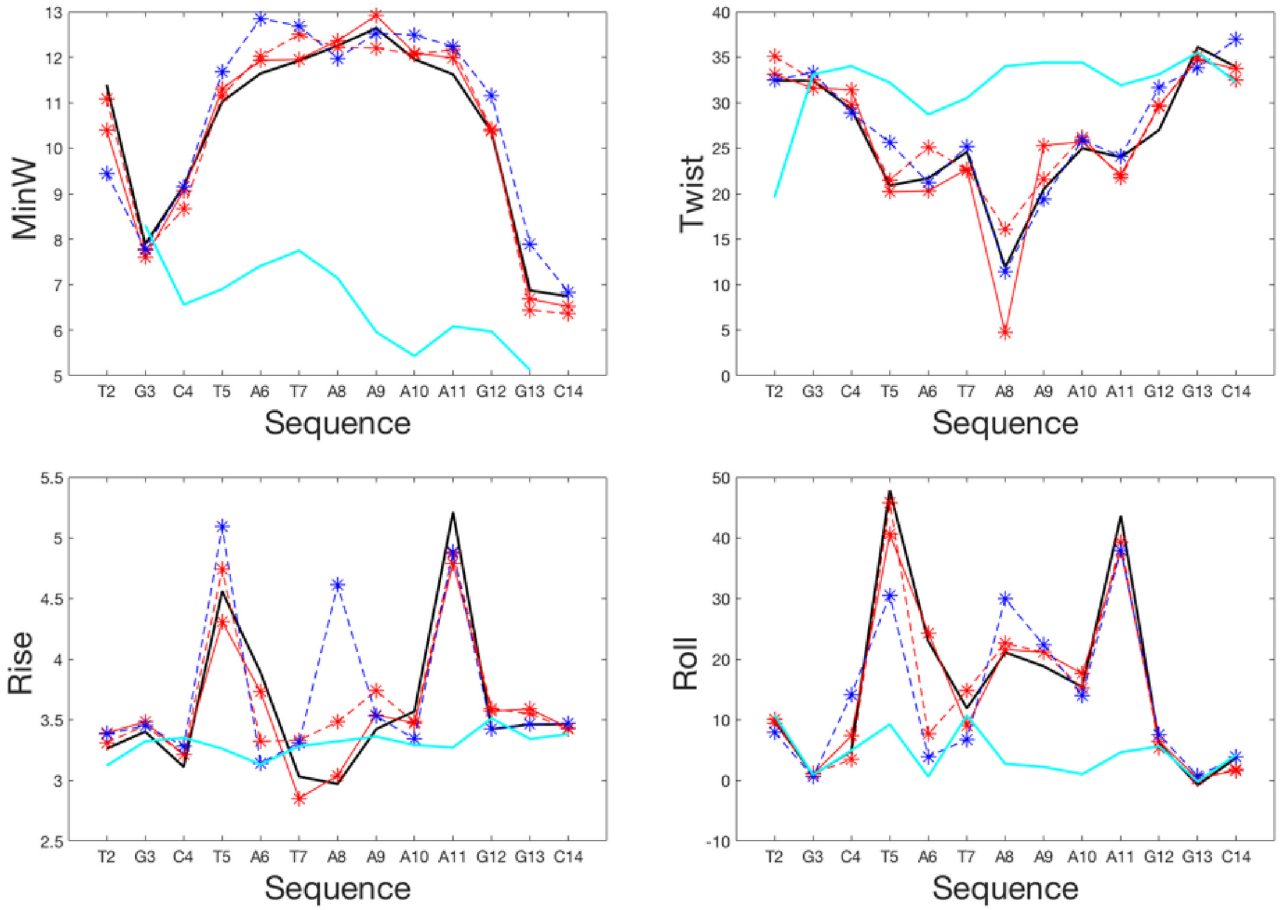

Figure 2. Conformational features of the TBP/DNA complex: minor groove width (Å), twist (°), rise (Å) and roll (°). The average structures from the all-atom simulations of the complex (thick black line) and of the isolated DNA (thick cyan line) are compared with the results from simulations using the simplified model proteins: uncharged (blue) or charged (red); no hydrogen bonding (dotted lines) or with hydrogen bonding (solid lines); no atomistic side chains (circles) or with key atomistic side chains (stars). Note that inter-bp parameters plotted at position i refer to the bp step i-i+1.

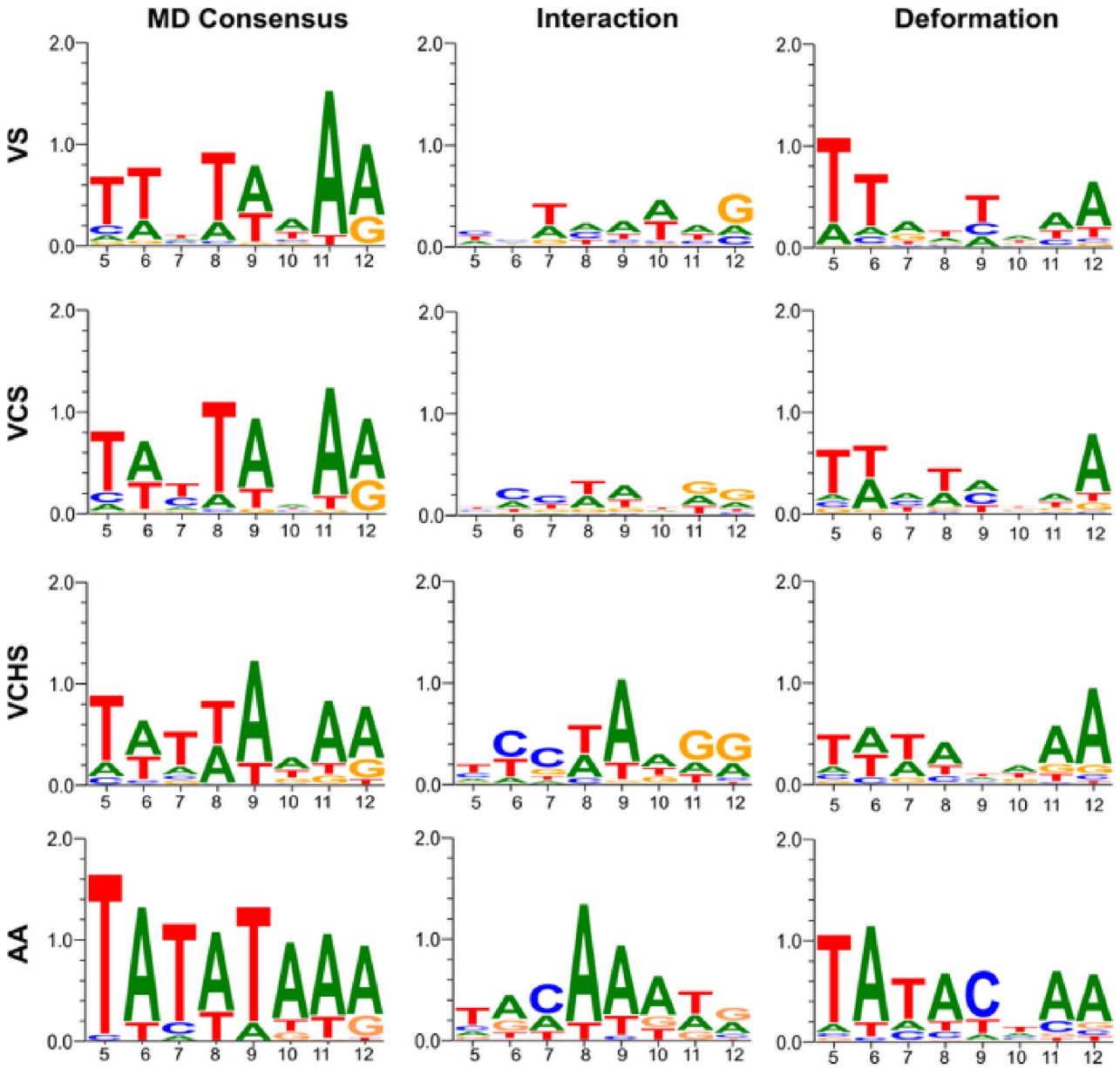

Figure 3. Sequence selectivity of DNA-bound to TBP. ADAPT threading results using structures from simplified model simulations compared with those from all-atom simulations.

In order to focus the present analysis on the components that contribute to sequence recognition, under controlled and well-defined simulation conditions, we will now analyze the four complexes we have chosen to study by comparing the simulations of the simplified protein model with results of the all-atom simulations (AA).

For each of our four complexes, the protein can be represented as an uncharged, low-dielectric, coarse-grain volume (V) that only interacts with DNA via short-range repulsion/dispersion terms, or it can be extended to include the influence of formal charges to the model (C), the possibility of hydrogen bonding (H) and, where necessary, flexible, all-atom amino acid side chains for key interface residues (S). Simulations of these models interacting with DNA are compared with all-atom results in terms of DNA conformation, ionic distributions and sequence selectivity (using the ADAPT threading approach).

Binding in the minor groove

TBP (Figure 1, http://www.uniprot.org/uniprot/P20226) is a component of the TF11D that belongs to the transcriptosome. It binds to the minor groove of DNA via an extended β-sheet and has an 8 base pair (bp) consensus sequence TATAWAWR (W ≡ A/T, R ≡ A/G) (30) Simulations were carried with a 16 bp oligomer having the sequence: CTGC**T$_5$ATAAAAG$_{12}$**GCTG. Bold characters indicate the TBP binding site located in positions 5–12.

The uncharged V protein model of TBP remains stably bound to DNA throughout the simulation, maintaining a wide minor groove. However, a correct bound conformation of DNA is not reproduced, due to the inability of the coarse grain representations of phenylalanine to intercalate at either end of the binding site (steps 5–6 and 11–12) (data not shown). To correct for this, we introduced all-atom side chains for the intercalating residues Phe 193 and Phe 284, as well as for the neighboring Phe 210 and Phe 301 that help to orient the intercalating residues, and for Ala 194 and Pro 285 that contribute to positioning TBP at its recognition site (49). The resulting VS model improves the DNA conformation and already exhibits a correct profile for the minor groove width (see Figure 2). However, it also results in an unusually large rise and roll at the central 8–9 step (not present in the isolated all-atom DNA simulation) and a reduced roll at the 5–6 step where Phe 193 intercalates. The addition of formal charges in the VCS model yields an almost correct DNA conformation, although the rise is still a little high for the 8–9 step that interacts with Asn 165 and Thr 309. The addition of hydrogen bonding in the VCHS model corrects this problem, but also leads to a slightly exaggerated unwinding at this step.

In terms of ion distributions (see Supplementary Figure S3 upper panels), all the charged protein models totally exclude ions from the minor groove binding site. In the major groove, all models also show high K+ molarities associated with the GC-rich regions on either side of the binding site and also, to a lesser extent, in the central 6–11 segment. Again, the addition of formal charges helps to more accurately reproduce the all-atom cation distribution.

If we now consider sequence recognition, we find that the VS model already shows a Pearson correlation coefficient (CC) of 0.65 with the all-atom results (see Supplementary Table S3), in line with a significant contribution of DNA deformation to selective TBP binding (Figure 3). Adding formal charges to the protein model (VCS) improves recognition locally at positions 5 and 6, but completing the model with hydrogen bonding (VCHS) finally increases to correlation to 0.76. Note that in this case, as for the all-atom simulations, the recognition of the 5′-TATA motif is largely due to the sequence-specific deformation of DNA.

SRY (Figure 4, http://www.uniprot.org/uniprot/Q05066) is a transcription factor controlling male sex determination. It binds to a 7 bp minor groove site with the consensus WAACAAW via an α-helix and a cationic C-terminal tail (31). Simulations were carried with a 14 bp oligomer having the sequence: CCT**G_4CACAAA$_{10}$**CACC. Bold characters indicate the SRY binding site, located in positions 4–10.

As for TBP, the uncharged V model of SRY binds stably to DNA. This model also induces a widened minor groove at the binding site, although quantitative details, including the groove width, and the rise, roll and twist of several steps within the binding site do not match the conformation of DNA seen in the all-atom simulations (Figure 5). It is worth noting that the step 8–9, where Ile 13 is partially intercalated has a correct conformation without the need for an all-atom isoleucine side chain. The addition of charges (model VC) improves details in the DNA conformation, while hydrogen bonding (model VCH) has little effect. Note that with either the VC or VCH models there are still some conformational differences in DNA with respect to the all-atom results, probably associated with artificially rigidifying the normally very flexible C-terminal tail.

Given the high charge density at the SRY/DNA interface (see Table 1), it is not surprising that the ion distribution (see Supplementary Figure S3 lower panels) around the complex is poorly reproduced until formal charges are added. In contrast, as for DNA conformation, adding hydrogen-bonding groups has little effect.

In terms of sequence recognition (Figure 6), as for TBP, the simple V model of SRY leads to reasonable sequence selectivity with a CC = 0.60 compared to the all-atom results (Supplementary Table S3). Adding formal charges and/or hydrogen bonding does not significantly affect this result, however the correlation can be improved by adding two all-atom side chains, specifically those of Arg 7 and Asn 10 that interact with position 7. These residues do not affect the DNA conformation (see Figure 5), but ensure recognition of a pyrimidine at position 7 and increase the overall correlation with the all-atom simulations to 0.79.

Binding in the major groove

SKN (Figure 7, http://www.uniprot.org/uniprot/P34707) is a transcription factor involved in development, stress response and neurodegeneration. It binds in the major groove via an α-helix with a 5 bp RTCAT consensus sequence (32,50). Its binding affinity and specificity also involve a basic N-terminal tail binding to an A/T-rich region upstream of the consensus site (51). Simulations were carried with a 17 bp oligomer having the sequence: TGACAAT**G_8TCAT$_{12}$**CCCTG. Bold characters indicate the SKN binding site located in positions 8–12.

In contrast to the minor groove binding proteins, the uncharged V model of SKN rapidly dissociates from DNA. Adding formal charges to make the VC model is however enough to stabilize the interaction and to deform DNA to a conformation similar to that seen in the all-atom study (see Figure 8). Adding hydrogen bonding to the SKN model does not change this situation. Some local variations, notably in twist in the 5′-flanking region (at steps 4–5, 7–8 and 8–9), rise (steps 7–8 and 8–9) and roll (4–5) can probably be ascribed to the artificially rigid N-terminal tail (c.f. the C-terminal tail of SRY mentioned above).

Both the VC and VCH models reproduce the main peak positions in the ion distribution surrounding the complex (see Supplementary Figure S4 upper panels), although they fail to generate strong molarity peaks in the minor groove of the 5′-flanking region, again possibly because of the artificially rigid N-terminal tail.

In terms of sequence specificity, both the VC and VCH models show a high correlation with the all-atom results (CC $\geq$ 0.80, see Figure 9 and Supplementary Table S3). Recognition is mainly due to direct interaction terms and the simplified protein models reproduce this result.

P22 (Figure 10, http://www.uniprot.org/uniprot/P69202) is a bacteriophage repressor protein that maintains the lysogenic state. It is a homodimer that binds DNA with α-helices in the major grooves of two half sites (P22L - left and P22R - right) separated by one helical turn (33). The central four base pairs of the binding site are not con-

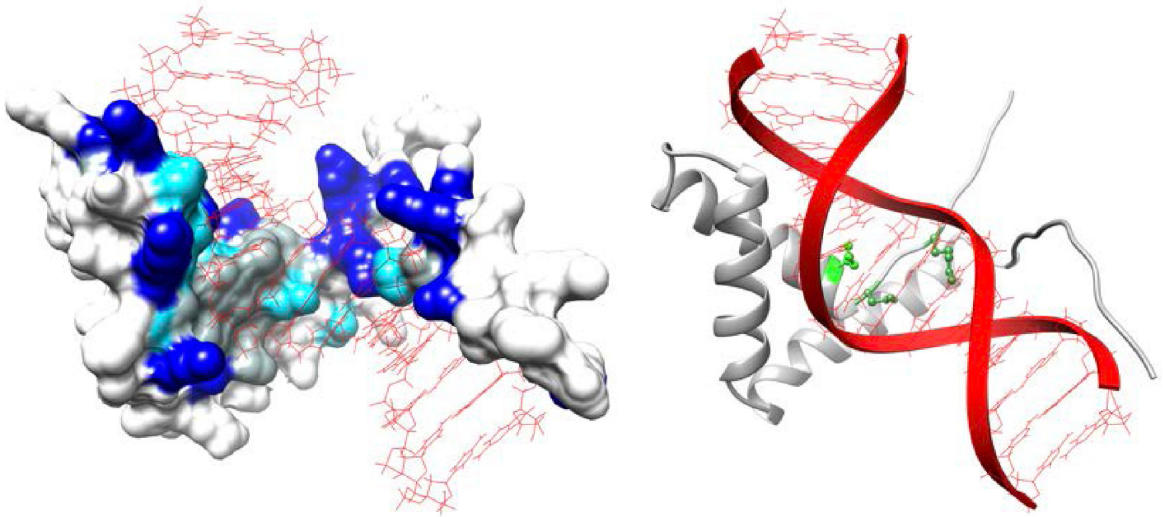

Figure 4. Visualization of the SRY/DNA complex. Left: the protein interface (charged residues: blue, polar residues: cyan, hydrophobic residues: dark grey). Right: ribbon representation of the complex. Key amino acid side chains are shown (Arg 7 and Asn 10: dark green, partially intercalating residue Ile 13: light green). In both panels, DNA is closest to the viewer and is shown in red.

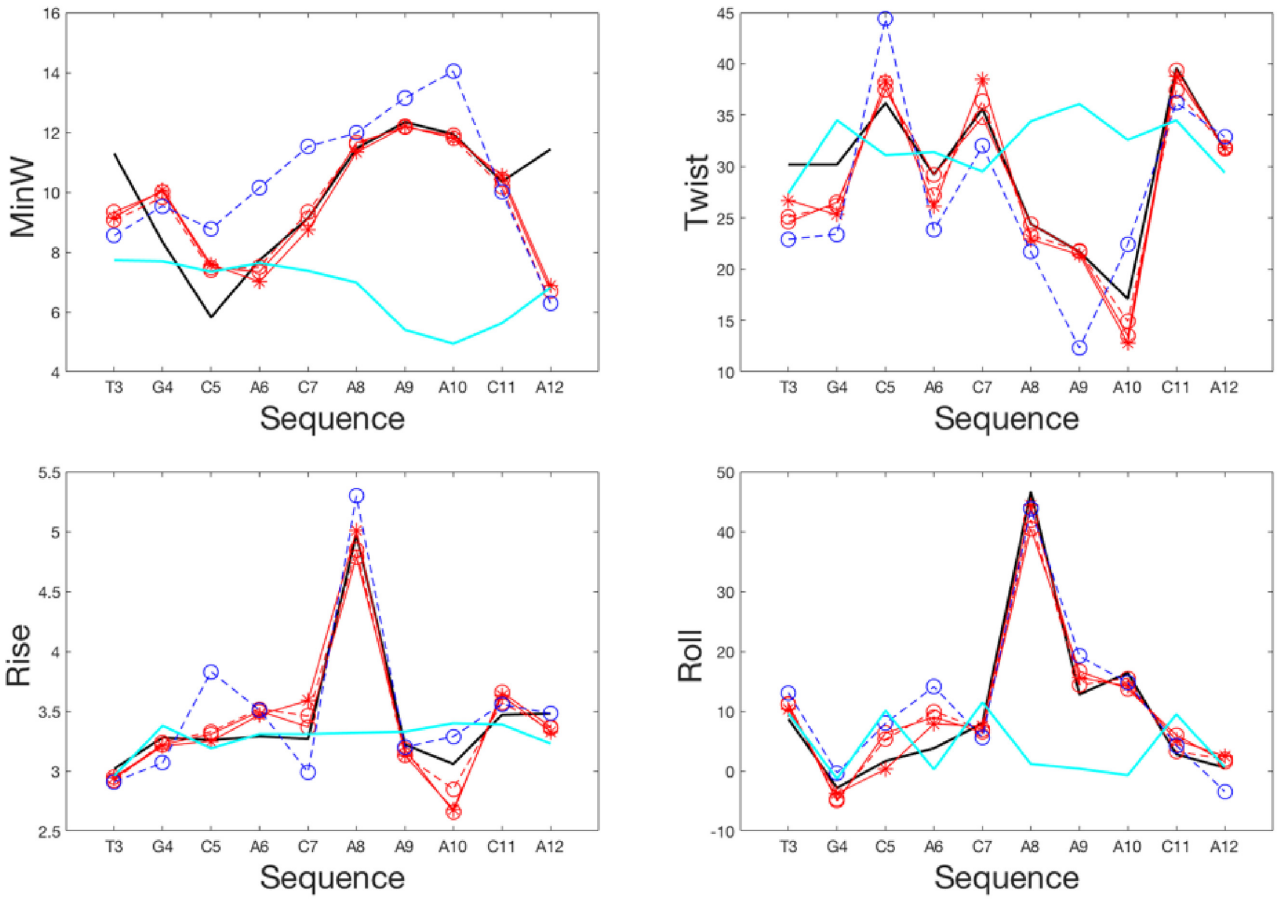

Figure 5. Conformational features of the SRY/DNA complex: minor groove width (Å), twist (°), rise (Å) and roll (°). The average structure from the all-atom simulations of the complex (thick black line) and of the isolated DNA (thick cyan line) are compared with the results from simulations using model proteins: uncharged (blue) or charged (red); no hydrogen bonding (dotted lines) or with hydrogen bonding (solid lines); no atomistic side chains (circles) or with key atomistic side chains (stars). Note that inter-bp parameters plotted at position i refer to the bp step i-i+1.

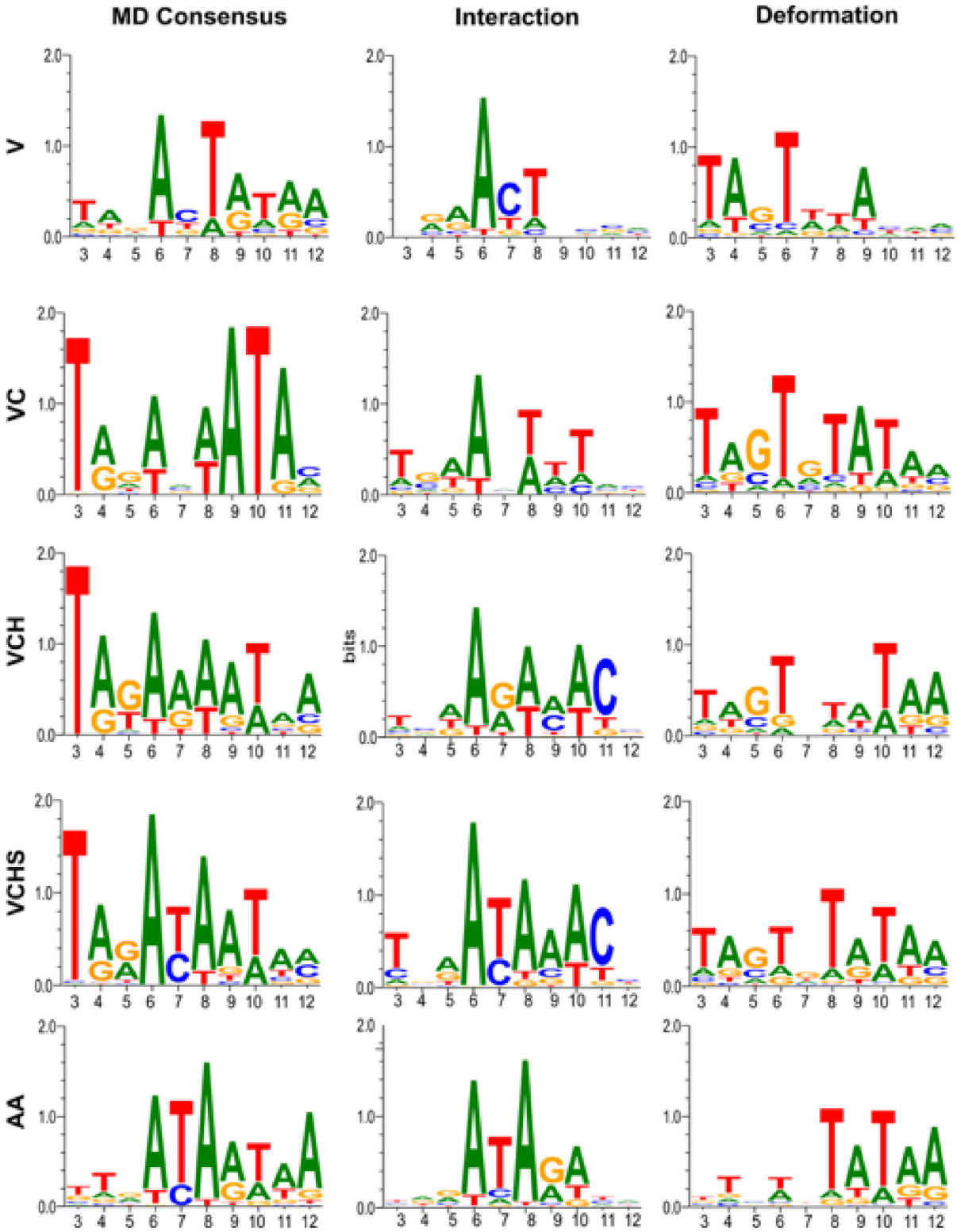

Figure 6. Sequence selectivity of DNA-bound to SRY. ADAPT threading results using structures from simplified model simulations compared with those from all-atom simulations.

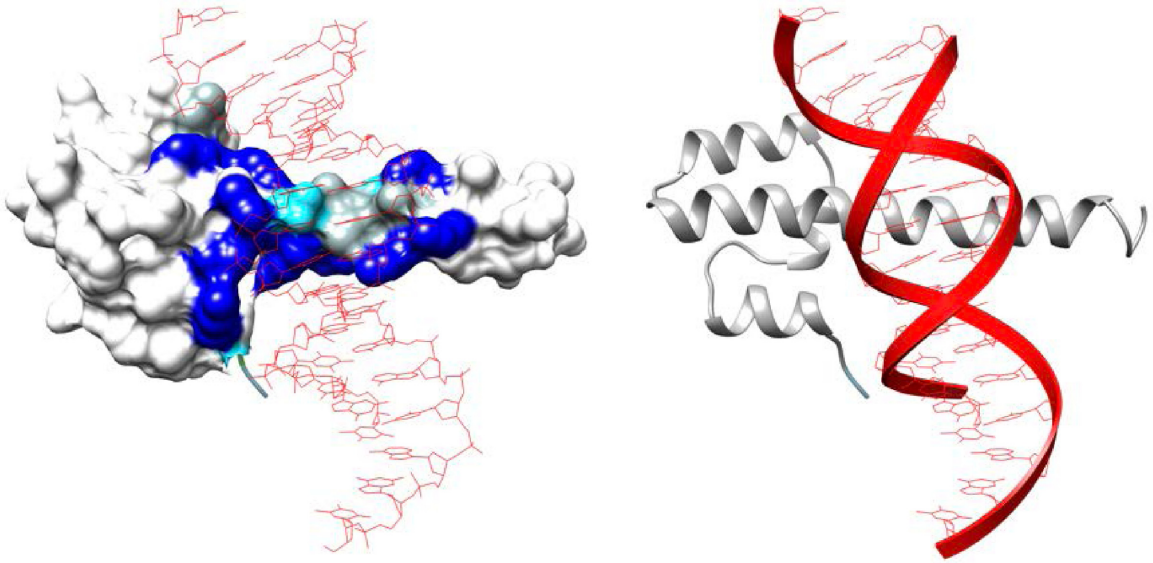

Figure 7. Visualization of the SKN/DNA complex. Left: the protein interface (charged residues: blue, polar residues: cyan, hydrophobic residues: dark gray). Right: ribbon representation of the complex. In both panels, DNA is closest to the viewer and is shown in red.

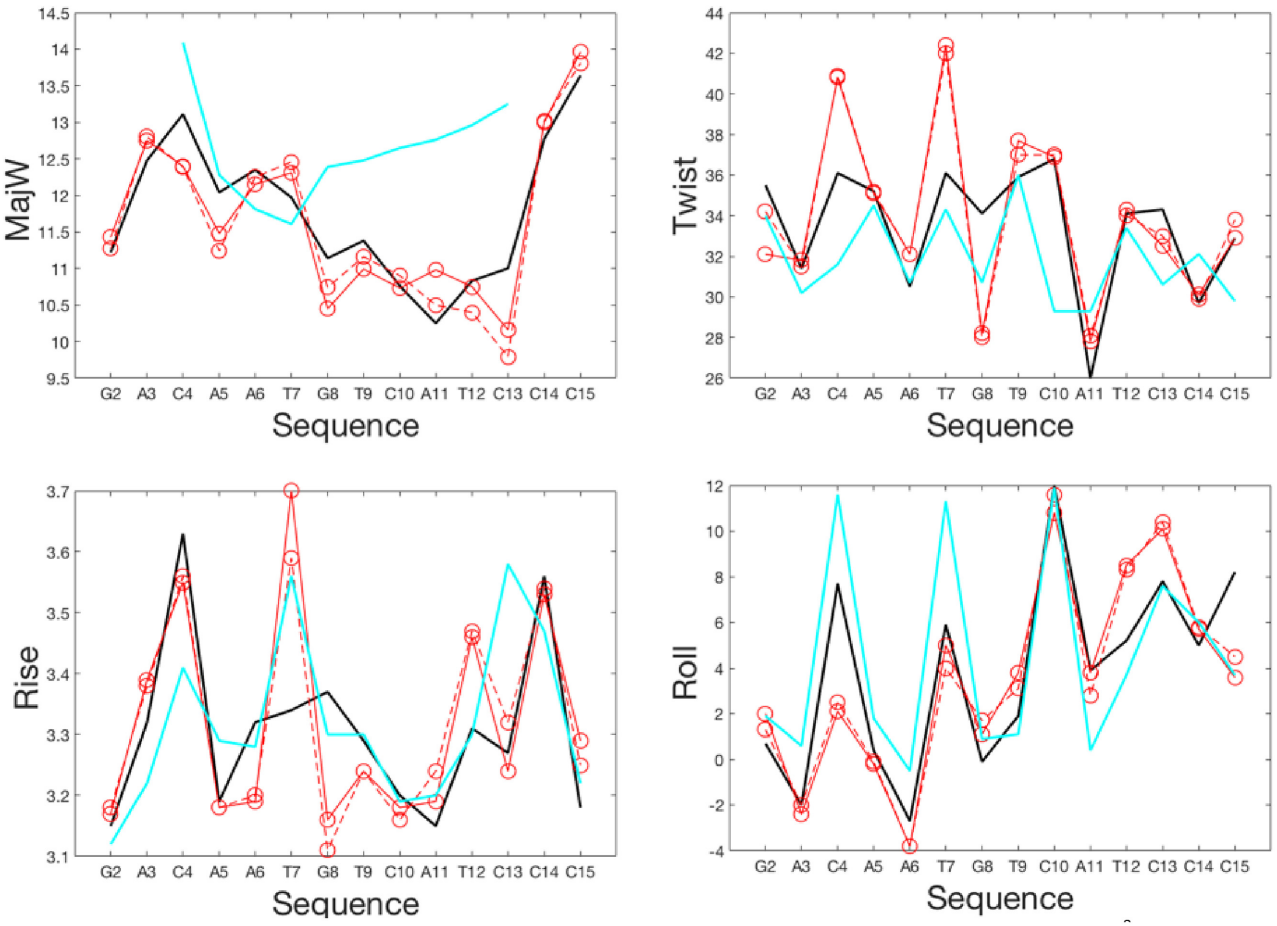

Figure 8. Conformational features of the SKN/DNA complex; minor groove width (Å), twist (°), rise (Å) and roll (°). The average structures from the all-atom simulations of the complex (solid black line) and of the isolated DNA (thick cyan line) are compared with the results from simulations using model proteins: uncharged (blue) or charged (red); no hydrogen bonding (dotted lines) or with hydrogen bonding (solid lines); no atomistic side chains (circles) or with key atomistic side chains (stars). Note that inter-bp parameters plotted at position i refer to the bp step i-i+1.

Table 1. Physical characteristics of the binding sites for the complexes studied

Complex	Binding site	BSA (Å^2)	Char.	Pol.	Hϕ	HB
TBP	Minor groove	1437	8+	14	22	7
SRY	Minor groove	1316	18+/1-	8	14	4
SKN	Major groove	787	13+/1-	3	7	2
P22-L	Major groove	1285	5+/3-	8	8	6
P22-R		L635/6R50	5+/3-	8	8	6

n.b. The binding site is shown in bold and its first and last positions are numbered. BSA refers to the buried surface area. Each complex is characterized by the number of charged (Char.), polar (Pol.) and hydrophobic (Hϕ) residues at the interface as well as the number of hydrogen bonds (HB).

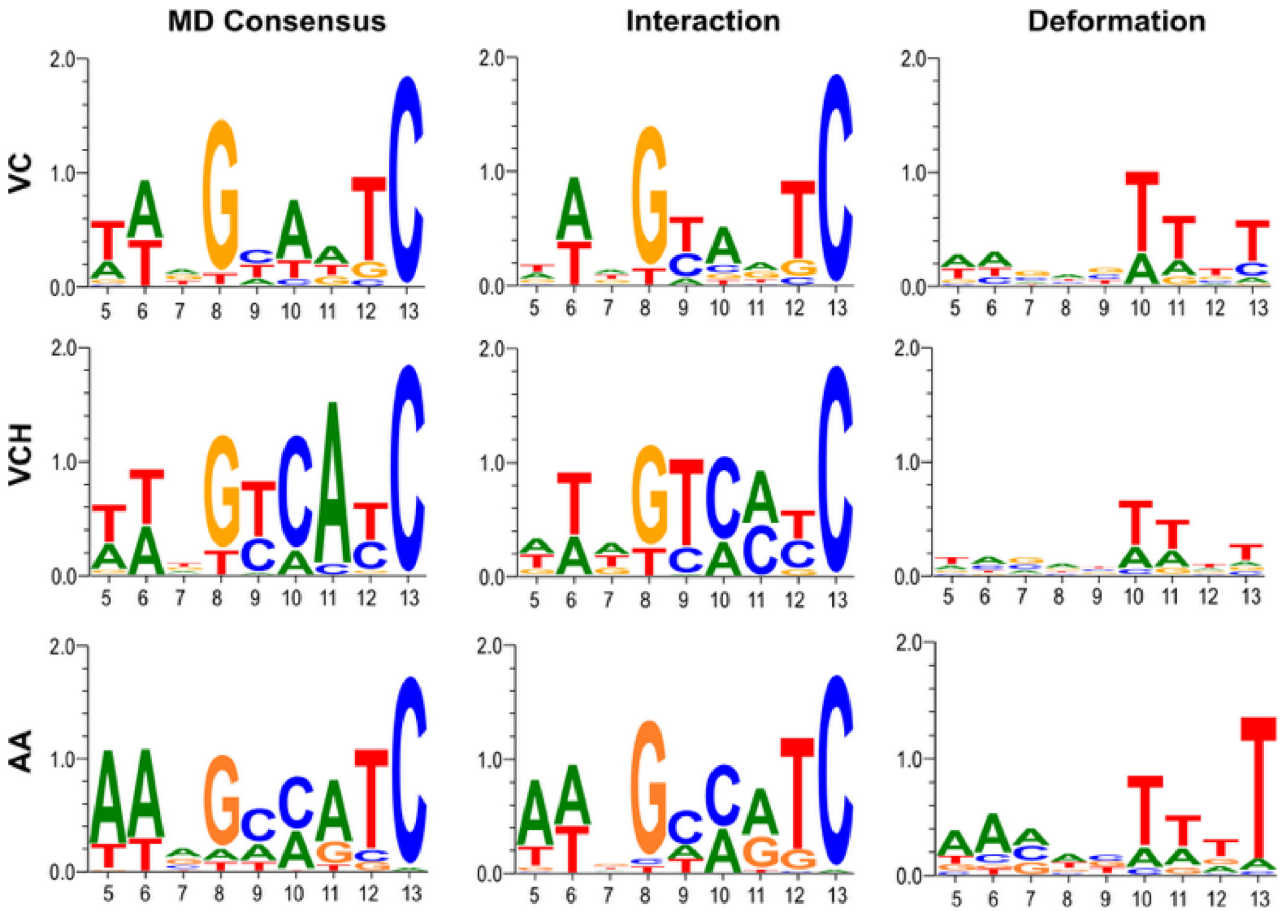

Figure 9. Sequence selectivity of DNA-bound SKN. ADAPT threading results using structures from simplified model simulations compared with those from all-atom simulations.

tacted by the protein, but show a preference for A/T base pairs. The consensus sequence deduced from the base substitution studies of Watkins et al. (33) is WTWAAG-WWCTTWAW (where the dashed indicate the absence of experimental data for the corresponding positions). Simulations were carried with a 20 bp oligomer having the sequence: TA**T$_3$TTAAG**ATAT**CTTAAA**$_{18}$TG. Bold characters indicate the P22 half sites located between positions 3–18.

An important feature of P22 binding is the hydrophobic interaction between a valine residue and four thymine methyl groups within each half site (Val 33 and segment 4–7 in P22L) (33). Because of this, we chose to include the all-atom Val 33 side chains in both monomers. Despite this choice, both the VS and VCS protein models rapidly dissociate from DNA. The addition of hydrogen bonding (model VCHS) stabilizes the complex and reproduces the main features of DNA deformation, although there are visible differences in both minor and major groove widths and some local variations in helical parameters (see Figure 11). This may again be due to a lack of flexibility, in this case, at the interface between the two monomers of P22.

Despite the minor conformational differences, the VCHS model accurately reproduces the ionic distribution around the complex (see Supplementary Figure S4 lower panels), and notably the striking presence of two K+ binding sites in the central minor groove at positions 9–10 and 11–12 that undoubtedly help to offset the repulsion caused by glutamic acid residues (Glu 44 and Glu 48) of P22 close to this region.

We recall that as mentioned in the methods section, we built the P22 model proteins using two different coarse-

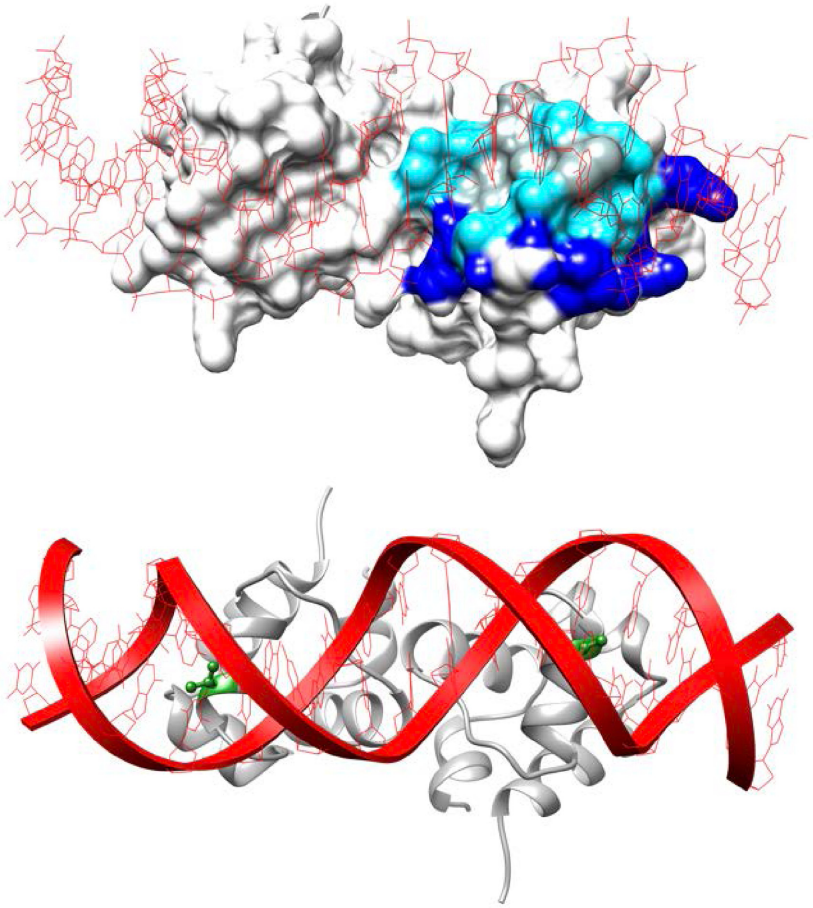

Figure 10. Visualization of the P22/DNA complex. Top: the P22R protein interface (charged residues: blue, polar residues: cyan, hydrophobic residues: dark gray). Bottom: ribbon representation of the complex. Key amino acid Val 33 side chains are shown in dark green. In both panels, DNA is closest to the viewer and is shown in red.

grain Gln 37 side chain conformations. Based on our earlier work showing the importance of these residues in recognition (28), we built the models with only Gln 37 in the P22R monomer directly interacting with a DNA base (C13). As seen above, this small difference has no visible impact on DNA deformation or on ion distribution, but it does modify sequence recognition. Although the VCHS model qualitatively reproduces the sequence specificity of P22, it shows significantly less correlation with the all-atom results for P22L (CC = 0.53) compared to P22R (CC = 0.91) (Figure 12 and Supplementary Table S3).

While ADAPT shows no sequence specificity for the four central base pairs with any of the simplified protein models, the presence of potassium ions in the central minor groove ions noted above would favor A/T base pairs in the center of the binding site, given the more negative potentials they generate in the minor groove (52,53). However, this effect is not seen by ADAPT since it ignores the effect of specifically bound ions. In passing, the small, central A/T preference seen with the all-atom simulations must be a result of the DNA deformation, specifically the narrowing of the minor groove, characteristic of the so-called B' conformation, also believed to play a role in P22 specificity (33,52).

DISCUSSION AND CONCLUSIONS

Before we discuss the behavior of the simplified protein models, it is worth pointing out that, for all the proteins we have studied, the isolated DNA binding site differs significantly from DNA bound to the protein. Although this is most striking for proteins binding in the minor groove, the major groove binders also change DNA by locally reducing the major groove width and selectively changing helical parameters (e.g. C_{10} twist and C_{13} rise for SKN, T_6 and T_{16} twist and rise and T_{11} roll for P22). In terms of ion distributions, all proteins not only displace ions from their binding sites, but also influence ions in the opposing groove (the most striking example being the A_9–T_{10} and A_{11}–T_{12} positions in the central minor groove of the P22 complex). Therefore, the model proteins must exert an effect on DNA if they are to reproduce, even partially, the structural, environmental or sequence-selective properties of the all-atom simulations.

To help to understand the results found with the simplified protein models, it is useful to consider the structural features of the complexes (see Table 1). The most striking distinction is the size of the buried interfaces for the minor groove compared to the major groove binding proteins. If

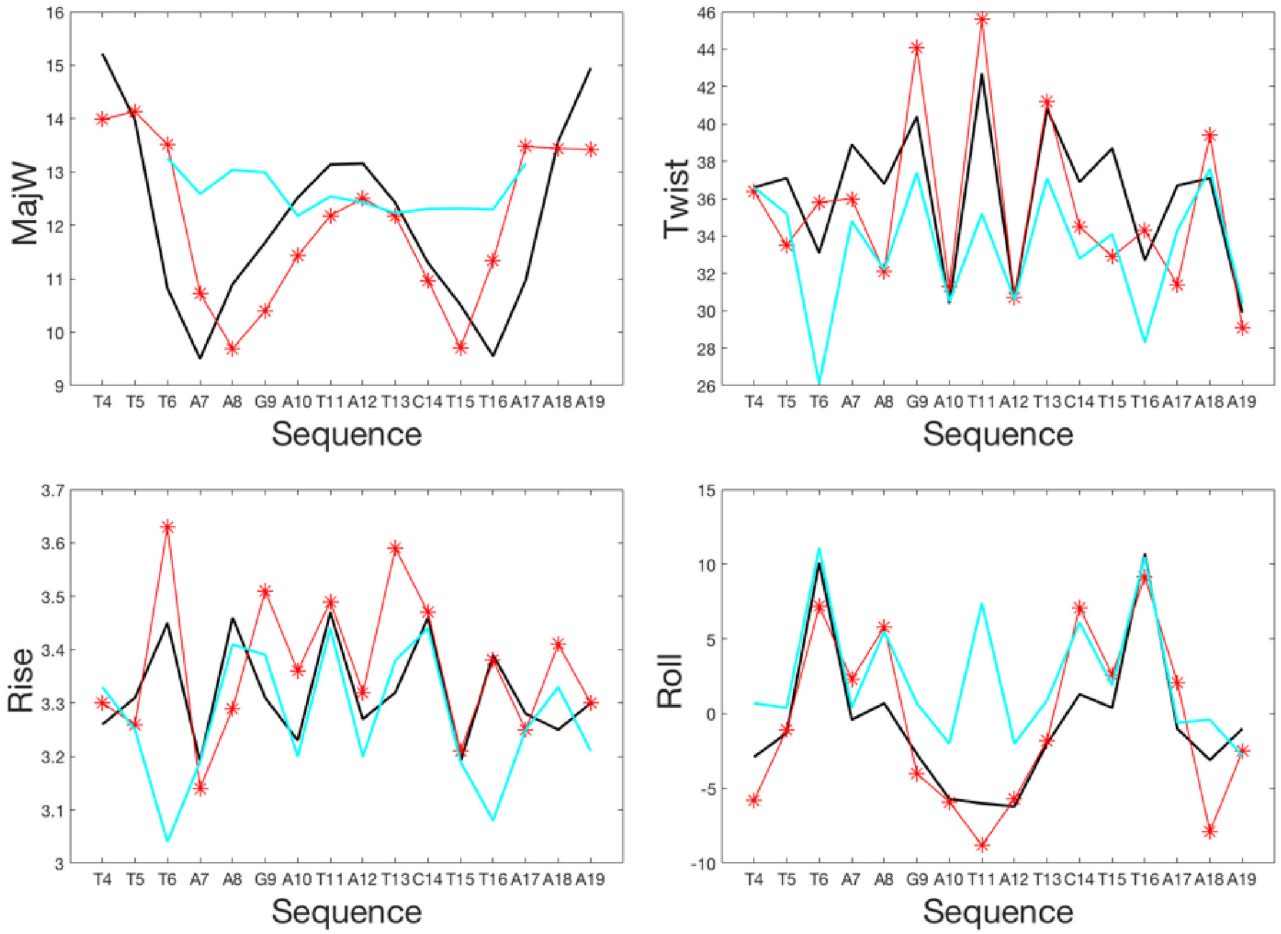

Figure 11. Conformational features of the P22/DNA complex: minor groove width (Å), twist (°), rise (Å) and roll (°). The average structures from the all-atom simulations of the complex (solid black line) and of the isolated DNA (thick cyan line) are compared with the results from simulations using model proteins: uncharged (blue) or charged (red); no hydrogen bonding (dotted lines) or with hydrogen bonding (solid lines); no atomistic side chains (circles) or with key atomistic side chains (stars). Note that inter-bp parameters plotted at position i refer to the bp step i-i+1.

we treat each monomer of P22 separately, the major groove interfaces for P22L, P22R and SKN are roughly half the size of those of TBP and SRY in the minor groove. This size difference is also reflected in the total number of amino acids forming the protein interface: 44 and 41 for TBP and SRY, versus 24 for SKN and P22L/R.

The large buried surface area of complexes involving minor groove binding can explain why the uncharged, low-dielectric (and effectively hydrophobic) V protein model remains stably bound to both TBP and SRY. Although taking the two monomers together yields a similar surface area for P22, the contact of each monomer is mainly with a single strand (see Figure 10) possibly explaining why neither the VS nor the VCS models remain bound. However, other factors in this case are the weak net positive charge on each monomer interface and the numerous hydrogen bonds, which effectively turn out to be essential in this case. The TBP interface also has a significant number of hydrogen bonds and these are again important for establishing local conformational features of DNA in the presence of the protein.

In contrast, both SRY and SKN have relatively few interface hydrogen bonds, but highly charged protein surfaces. In these cases the charged VC protein model reproduces the main features of the DNA deformation and most of the sequence selectivity.

For all the proteins we have studied, only very few amino acid side chains need to be represented at the all-atom level in order to correctly deform individual DNA steps or to locally improve sequence selectivity. In the case of aromatic residues, such as the intercalating phenylalanine groups of TBP, this is in part due to their poor representation using spherical pseudoatoms (in contrast to the virtually spherical isoleucine residue intercalated at the SRY/DNA interface).

Overall, while we have only studied a small group of DNA-binding proteins, it is interesting to see that, in three cases out of four, representing a bound protein by a rigid coarse-grain low-dielectric volume, supplemented with Lennard-jones interactions and formal amino acid charges does a surprisingly good job of explaining the main features of DNA deformation, environmental perturbation and sequence specificity seen with detailed, all-atom simulations. The absence of flexibility, understandably limits

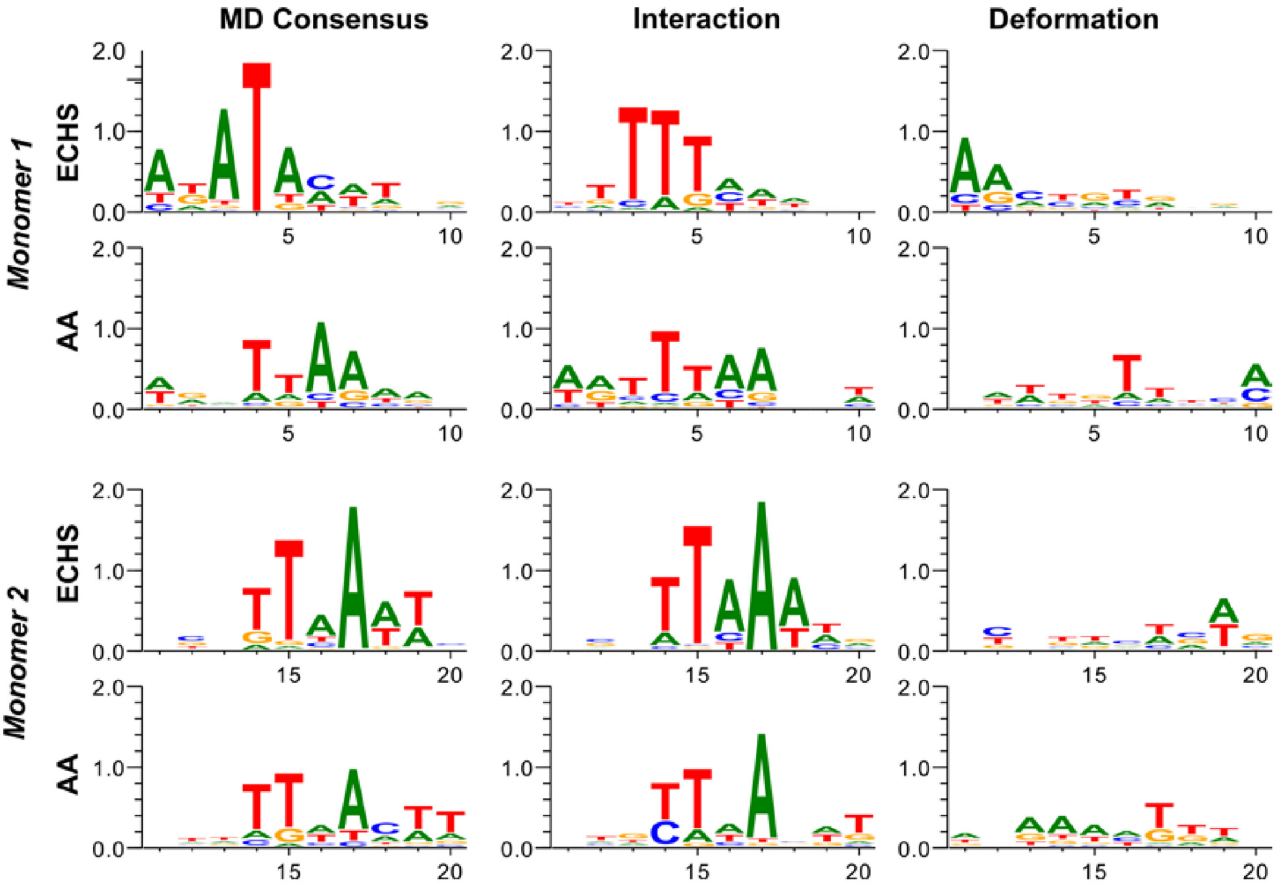

Figure 12. Sequence selectivity for the half-sites of DNA bound to P22. ADAPT threading results using structures from simplified model simulations compared with those from all-atom simulations.

the accuracy of the models in the case of flexible N- or C-terminal tails or of dimeric protein interfaces.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors wish to acknowledge GENCI for a generous allocation of supercomputer resources at the CINES center in Montpellier.

FUNDING

ANR project CHROME [ANR-12-BSV5-0017-01]; Rhône-Alpes ARC 1 Sante Doctoral Grant (to L.E.). Funding for open access charge: ANR project CHROME [ANR-12-BSV5-0017-01].
Conflict of interest statement. None declared.

REFERENCES

1. Zubay,G. and Doty,P. (1959) The isolation and properties of deoxyribonucleoprotein particles containing single nucleic acid molecules. *J. Mol. Biol.*, **1**, 1–20.
2. Pauling,L., Corey,R.B. and Branson,H.R. (1951) The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U.S.A.*, **37**, 205–211.
3. Watson,J.D. and Crick,F.H. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, **171**, 737–738.
4. Dickerson,R.E. and Geis,I. (1969) *The Structure and Action of Proteins*. Harper & Row, NY.
5. Seeman,N.C., Rosenberg,J.M. and Rich,A. (1976) Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **73**, 804–808.
6. Wing,R., Drew,H., Takano,T., Broka,C., Tanaka,S., Itakura,K. and Dickerson,R.E. (1980) Crystal structure analysis of a complete turn of B-DNA. *Nature*, **287**, 755–758.
7. Parvin,J.D., McCormick,R.J., Sharp,P.A. and Fisher,D.E. (1995) Pre-bending of a promoter sequence enhances affinity for the TATA-binding factor. *Nature*, **373**, 724–727.
8. Koudelka,G.B. and Carlson,P. (1992) DNA twisting and the effects of non-contacted bases on affinity of 434 operator for 434 repressor. *Nature*, **355**, 89–91.
9. Thorogood,H., Grasby,J.A. and Connolly,B.A. (1996) Influence of the phosphate backbone on the recognition and hydrolysis of DNA by the EcoRV restriction endonuclease. A study using oligodeoxynucleotide phosphorothioates. *J. Biol. Chem.*, **271**, 8855–8862.
10. Flatters,D. and Lavery,R. (1998) Sequence-dependent dynamics of TATA-Box binding sites. *Biophys. J.*, **75**, 372–381.
11. Paillard,G. and Lavery,R. (2004) Analyzing protein–DNA recognition mechanisms. *Structure*, **12**, 113–122.
12. Rohs,R., West,S.M., Sosinsky,A., Liu,P., Mann,R.S. and Honig,B. (2009) The role of DNA shape in protein–DNA recognition. *Nature*, **461**, 1248–1253.

13. Rohs,R., Jin,X., West,S.M., Joshi,R., Honig,B. and Mann,R.S. (2010) Origins of specificity in protein–DNA recognition. *Annu. Rev. Biochem.*, **79**, 233–269.
14. Bouvier,B., Zakrzewska,K. and Lavery,R. (2011) Protein–DNA recognition triggered by a DNA conformational switch. *Angew. Chem. Int. Ed. Engl.*, **50**, 6516–6518.
15. Abe,N., Dror,I., Yang,L., Slattery,M., Zhou,T., Bussemaker,H.J., Rohs,R. and Mann,R.S. (2015) Deconvolving the recognition of DNA shape from sequence. *Cell*, **161**, 307–318.
16. Yang,L., Orenstein,Y., Jolma,A., Yin,Y., Taipale,J., Shamir,R. and Rohs,R. (2017) Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Mol. Syst. Biol.*, **13**, 910.
17. Mirzabekov,A.D. and Rich,A. (1979) Asymmetric lateral distribution of unshielded phosphate groups in nucleosomal DNA and its role in DNA bending. *Proc. Natl. Acad. Sci. U.S.A.*, **76**, 1118–1121.
18. Strauss,J.K. and Maher,L.J. 3rd (1994) DNA bending by asymmetric phosphate neutralization. *Science*, **266**, 1829–1834.
19. Strauss,J.K., Roberts,C., Nelson,M.G., Switzer,C. and Maher,L.J. 3rd (1996) DNA bending by hexamethylene-tethered ammonium ions. *Proc. Natl. Acad. Sci. U.S.A.*, **93**, 9515–9520.
20. Elcock,A.H. and McCammon,J.A. (1996) The low dielectric interior of proteins is sufficient to cause major structural changes in DNA on association. *J. Am. Chem. Soc.*, **118**, 3787–3788.
21. Lavery,R., Moakher,M., Maddocks,J.H., Petkeviciute,D. and Zakrzewska,K. (2009) Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.*, **37**, 5917–5929.
22. Lavery,R., Maddocks,J.H., Pasi,M. and Zakrzewska,K. (2014) Analyzing ion distributions around DNA. *Nucleic Acids Res.*, **42**, 8138–8149.
23. Lafontaine,I. and Lavery,R. (2000) ADAPT: a molecular mechanics approach for studying the structural properties of long DNA sequences. *Biopolymers*, **56**, 292–310.
24. Deremble,C., Lavery,R. and Zakrzewska,K. (2008) protein–DNA recognition: breaking the combinatorial barrier. *Comput. Phys. Commun.*, **179**, 112–119.
25. Tóth-Petróczy,A., Simon,I., Fuxreiter,M. and Levy,Y. (2009) Disordered tails of homeodomains facilitate DNA recognition by providing a trade-off between folding and specific binding. *J. Am. Chem. Soc.*, **131**, 15084–15085.
26. Vuzman,D. and Levy,Y. (2011) Intrinsically disordered regions as affinity tuners in protein–DNA interactions. *Mol. Biosyst.*, **8**, 47–57.
27. Etheve,L., Martin,J. and Lavery,R. (2015) Dynamics and recognition within a protein–DNA complex. *Nucleic Acids Res.*, **44**, 1440–1448.
28. Etheve,L., Martin,J. and Lavery,R. (2016) Protein–DNA interfaces: a molecular dynamics analysis of time-dependent recognition processes for three transcription factors. *Nucleic Acids Res.*, **44**, 9990–10002.
29. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
30. Nikolov,D.B., Chen,H., Halay,E.D., Hoffman,A., Roeder,R.G. and Burley,S.K. (1996) Crystal structure of a human TATA box-binding protein/TATA element complex. *Proc. Natl. Acad. Sci. U.S.A.*, **93**, 4862–4867.
31. Murphy,E.C., Zhurkin,V.B., Louis,J.M., Cornilescu,G. and Clore,G.M. (2001) Structural basis for SRY-dependent 46-X, Y sex reversal: modulation of DNA bending by a naturally occurring point mutation. *J. Mol. Biol.*, **312**, 481–499.
32. Rupert,P.B., Daughdrill,G.W., Bowerman,B. and Matthews,B.W. (1998) A new DNA-binding motif in the Skn-1 binding domain–DNA complex. *Nat. Struct. Biol.*, **5**, 484–491.
33. Watkins,D., Hsiao,C., Woods,K.K., Koudelka,G.B. and Williams,L.D. (2008) P22 c2 repressor-operator complex: mechanisms of direct and indirect readout. *Biochemistry*, **47**, 2325–2338.
34. Lavery,R., Zakrzewska,K. and Sklenar,H. (1995) JUMNA (Junction Minimization of Nucleic-Acids). *Comput. Phys. Commun.*, **91**, 135–158.
35. Zacharias,M. (2003) protein–protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Sci.*, **12**, 1271–1282.
36. Hornak,V., Abel,R., Okur,A., Strockbine,B., Roitberg,A. and Simmerling,C. (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins*, **65**, 712–725.
37. Berendsen,H.J.C., Grigera,J.R. and Straatsma,T.P. (1987) The missing term in effective pair potentials. *J. Phys. Chem.*, **91**, 6269–6271.
38. Pearlman,D.A., Case,D.A., Caldwell,J.W., Ross,W.S., Cheatham,T.E., DeBolt,S., Ferguson,D., Seibel,G. and Kollman,P. (1995) AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.*, **91**, 1–41.
39. Case,D.A., Cheatham,T.E., Darden,T., Gohlke,H., Luo,R., Merz,K.M., Onufriev,A., Simmerling,C., Wang,B. and Woods,R.J. (2005) The Amber biomolecular simulation programs. *J. Comput. Chem.*, **26**, 1668–1688.
40. Cheatham,T.E. 3rd, Cieplak,P. and Kollman,P.A. (1999) A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.*, **16**, 845–862.
41. Pérez,A., Marchán,I., Svozil,D., Sponer,J., Cheatham,T.E., Laughton,C.A. and Orozco,M. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.*, **92**, 3817–3829.
42. Dang,L.X. (1995) Mechanism and thermodynamics of ion selectivity in aqueous-solutions of 18-crown-6 ether—a molecular dynamics study. *J. Am. Chem. Soc.*, **117**, 6954–6960.
43. Darden,T., Perera,L., Li,L. and Pedersen,L. (1999) New tricks for modelers from the crystallography toolkit: the particle mesh Ewald algorithm and its use in nucleic acid simulations. *Structure*, **7**, R55–R60.
44. Berendsen,H.J.C., Postma,J.P.M., van Gunsteren,W.F., DiNola,A. and Haak,J.R. (1984) Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, **81**, 3684–3690.
45. Ryckaert,J.P., Ciccotti,G. and Berendsen,H.J.C. (1977) Numerical-integration of Cartesian equations of motion of a system with constraints—molecular-dynamics of N-alkanes. *J. Comput. Phys.*, **23**, 327–341.
46. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
47. Krull,M., Voss,N., Choi,C., Pistor,S., Potapov,A. and Wingender,E. (2003) TRANSPATH: an integrated database on signal transduction and a tool for array analysis. *Nucleic Acids Res.*, **31**, 97–100.
48. Pasi,M., Maddocks,J.H. and Lavery,R. (2015) Analyzing ion distributions around DNA: sequence-dependence of potassium ion distributions from microsecond molecular dynamics. *Nucleic Acids Res.*, **43**, 2413–2423.
49. Spencer,J.V. and Arndt,K.M. (2002) A TATA binding protein mutant with increased affinity for DNA directs transcription from a reversed TATA sequence in vivo. *Mol. Cell. Biol.*, **22**, 8744–8755.
50. Blackwell,T.K., Bowerman,B. and Weintraub,H. (1994) Formation of a monomeric DNA binding domain by Skn-1 bZIP and homeodomain elements. *Science*, **266**, 621–628.
51. Kophengnavong,T., Carroll,A.S. and Blackwell,T.K. (1999) The SKN-1 amino-terminal arm is a DNA specificity segment. *Mol. Cell. Biol.*, **19**, 3039–3050.
52. Watkins,D., Mohan,S., Koudelka,G.B. and Williams,L.D. (2010) Sequence recognition of DNA by protein-induced conformational transitions. *J. Mol. Biol.*, **396**, 1145–1164.
53. Harris,L.A., Watkins,D., Williams,L.D. and Koudelka,G.B. (2013) Indirect readout of DNA sequence by p22 repressor: roles of DNA and protein functional groups in modulating DNA conformation. *J. Mol. Biol.*, **425**, 133–143.