1440–1448 *Nucleic Acids Research, 2016, Vol. 44, No. 3* *Published online 31 December 2015*
doi: 10.1093/nar/gkv1511

Dynamics and recognition within a protein–DNA complex: a molecular dynamics study of the SKN-1/DNA interaction

Loïc Etheve, Juliette Martin and Richard Lavery*

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

Received September 23, 2015; Revised November 30, 2015; Accepted December 15, 2015

ABSTRACT

Molecular dynamics simulations of the *Caenorhabditis elegans* transcription factor SKN-1 bound to its cognate DNA site show that the protein–DNA interface undergoes significant dynamics on the microsecond timescale. A detailed analysis of the simulation shows that movements of two key arginine side chains between the major groove and the backbone of DNA generate distinct conformational substates that each recognize only part of the consensus binding sequence of SKN-1, while the experimentally observed binding specificity results from a time-averaged view of the dynamic recognition occurring within this complex.

INTRODUCTION

In recent years, a large structural database of protein–DNA complexes has been established, mainly through the contribution of X-ray crystallography. Although this information has undoubtedly been invaluable in understanding many aspects of protein–DNA interactions, it is true that it gives a rather static view of such complexes. The possible role of the dynamics of protein–DNA interfaces has nevertheless been a subject of interest for many years. A significant number of experimental studies have notably aimed at understanding how proteins approach and bind to their DNA targets and how they distinguish non-specific from cognate sites. Both, nuclear magnetic resonance (NMR) and paramagnetic resonance approaches have been used to better characterize non-specific protein binding and to analyze how such largely electrostatic interactions (reliant on arginine or lysine salt bridges with DNA phosphate groups) enable enhanced diffusion along DNA and can be subsequently transformed into specific binding, at least partially through the establishment of direct contacts with the nucleic acid bases (1–5). These mechanisms have also been the subject of a large number of theoretical (6–9) and molecular simulation studies (10–15) at various levels of detail, providing models of recognition mechanisms and suggesting how these mechanisms finally control the kinetics of gene expression at the cellular level (16–18).

The role of dynamics is however not limited to non-specific complexes and search mechanisms. Dynamics can also be important for specific protein–DNA complexes. Flexible, positively charged protein tails are a feature of many transcription factors. These tails, and also flexible linkers between DNA binding domains, can assist binding and can serve to fine tune specificity (19,20). Novel NMR studies using ^{15}N relaxation times and ^{15}N-^{31}P scalar coupling have also shown that lysine-phosphate salt bridges within specific complexes are themselves dynamic and direct interactions are regularly broken and remade (21–23), in line with earlier studies of salt bridges within proteins (24). This finding has recently been supported by all-atom molecular dynamics (MD) studies of homeodomain and Zn-finger complexes with DNA (25). Another aspect of protein–DNA interface dynamics is illustrated in a recent MD study of telomere repeat binding factors (TRF1 and TRF2), where the dynamics of individual amino acids chains suggested that they could contribute to the recognition of more than one base pair, helping to resolve conflicting experimental data (26).

As part of our ongoing attempt to better understand protein–DNA interactions using computer simulation techniques, we decided to couple long MD simulations with a time-dependent analysis of sequence selectivity using a sequence threading technique (ADAPT) that we have developed (27–30). ADAPT enables us to calculate and rank the binding energy of all possible DNA sequences within a protein–DNA complex (energy minimizing the interface structure for every sequence) and thus to obtain a computational position weight matrix (PWM). We already used this approach to study the appearance of base sequence selectivity during the approach of the mammalian transcription factor SRY to its DNA target (12). SRY, which controls the development of the male phenotype, is a member of the SOX (SRY-type HMG Box) family (31). By binding to the DNA minor groove, this protein creates significant DNA

*To whom correspondence should be addressed. Tel: +33 0 4 72 72 26 37; Fax: +33 0 4 72 72 26 04; Email: richard.lavery@ibcp.fr deformation (32). We were able to show that this deformation indeed plays a major role in the resulting binding selectivity and that SRY therefore relies on a so-called indirect recognition mechanism.

Here, we chose to study a very different protein, the transcription factor SKN-1. SKN-1 is a *Caenorhabditis elegans* transcription factor involved in early embryonic development, oxidative stress resistance and aging (33,34). It is homologous to the human Nrf proteins that are also involved in stress response. Although it contains a basic C-terminal helix bound in the major groove of DNA analogous to the bZIP transcription factors (e.g. c-Jun and GCN4), it lacks a leucine zipper and does not dimerize. It also contains a basic N-terminal tail similar to those of the homeodomain proteins (35) that is responsible for high-affinity binding to AT-rich sequences at the 5′-end of the binding site (36). Its consensus binding site involves five base pairs RTCAT (where R ≡ A/G) (37). Genomic studies of genes up- or down-regulated by SKN-1 are consistent with this consensus, but show some modulations in specificity within the consensus site (38,39). The crystal structure of the 84 residue C-terminal DNA binding domain complexed with a cognate DNA oligomer (35) shows that this transcription factor induces only moderate DNA deformation and is consequently expected to recognize its binding site via a direct mechanism involving specific amino-acid base contacts.

In line with the NMR and simulations studies cited above (21,22,25), the 0.5 μs MD simulation of the SKN-1/DNA complex we have carried out shows significant dynamics at the protein–DNA interface. Most interestingly, this involves the breakage of backbone salt bridges and formation of base contacts, recalling the mechanisms described for the passage between non-specific and specific complexes (1,2,11), but here occurring within an existing specific complex.

By coupling our MD simulation with ADAPT sequence threading we have been able to establish that the observed interface dynamics indeed affects sequence selectivity. This suggests that the protein–DNA interfaces of specifically bound transcription factors may be considerably more dynamic than previously expected and, moreover, that an observed binding specificity may, at least in some cases, be the time-averaged result of a number of different sub-states where only parts of the overall cognate sequence are actually recognized.

MATERIALS AND METHODS

MD simulations

The structure of the SKN-1/DNA complex (PDB code 1SKN) was taken from the X-ray study of Rupert *et al.* (35). The single-stranded ends of the DNA oligomer were completed with complementary nucleotides to form a 17-mer (see Figure 1A and B). Hydrogen atoms were added to both the DNA and the protein and the complex was solvated with SPC/E water molecules (40) within a truncated octahedral box, ensuring a solvent shell of at least 10 Å around the solute. The solute was neutralized with 32 potassium ions and then sufficient K^+/Cl^- ion pairs were added to reach a concentration of 150 mM. The ions were initially placed at random, but at least 5 Å from DNA and 3.5 Å from one another. The resulting system contained roughly 10 400 water molecules and 34 000 atoms in total.

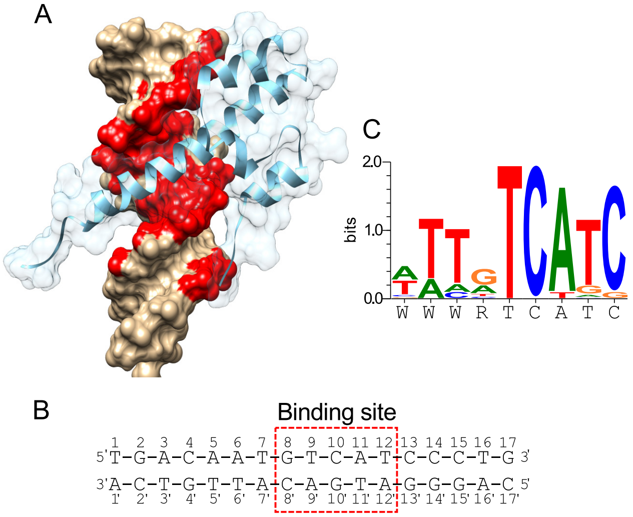


Figure 1. (**A**) Structure of SKN-1 protein–DNA complex (35). SKN-1 is shown in blue indicating its secondary structure and its surface envelope. The DNA oligomer is shown as a brown surface envelope with the protein-binding surface indicated in red. (**B**) DNA sequence used for the MD simulations, with the principal protein-binding site delimited by the red dashed box. Note that the first 'Watson' strand of the oligomer is numbered 1–17 in the 5′-3′ sense. Each complementary nucleotide in the 'Crick' strand has an identical number with a quote. (**C**) Experimental PWM for SKN-1 (W ≡A/T, R ≡A/G) from the JASPAR database (63).

MD simulations were performed with the AMBER 12 suite of programs (41,42) using PARM99 parameters (43) and the bsc0 modifications (44) for the solute and Dang parameters (45) for the surrounding ions. Simulations employed periodic boundary conditions and electrostatic interactions were treated using the particle-mesh Ewald algorithm (46,47) 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.

The system was initially subjected to energy minimization with harmonic restraints of 25 kcal mol^{-1} $Å^{-2}$ on the solute atoms. The system was then heated to 300 K at constant volume during 100 ps. Constraints were then relaxed from 5 to 1 kcal mol^{-1} $Å^{-2}$ during a series of 1000 steps of energy minimization (500 steps of steepest descent and 500 steps of conjugate gradient) followed by 50 ps of equilibration with restraints of 0.5 kcal mol^{-1} $Å^{-2.}$ and 50 ps without solute 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 (48) with a coupling constant of 5 ps and SHAKE constraints were applied to all bonds involving hydrogens (49). Conformational snapshots were saved for further analysis every ps. For comparison purposes, the 17-mer DNA oligomer was also simulated alone using an identical protocol, creating a second 500 ns trajectory.

Conformational and environmental analysis

Average DNA conformation, DNA conformational fluctuations and ion distributions around the SKN-1/DNA complex during the MD simulations were analyzed with the Curves+ program (50) and the Canal and Canion utilities (https://bisi.ibcp.fr/tools/curves_plus/). Using the recently developed ion analysis approach, based on describing ion positions with respect to the DNA helical axis, it was notably possible to calculate average ion molarities within the DNA grooves (51,52). As in our earlier work, the groove limit was set at a radius of 10.25 Å from the DNA helical axis (the average radial position of the backbone phosphorus atoms), while the angular limits were determined by the average position of the sugar C1' atoms. Lastly, hydrogen bond and salt bridges were analyzed using AMBER Tools (53) applying a distance cut-off of $\leq$ 3.5 Å between the relevant heavy atoms and an angle cut-off of $\geq$ 135° at the intervening hydrogen atom.

Clustering the MD trajectory

In order to identify conformational clusters within the MD trajectory, we began by extracting snapshots every 200 ps. Since we were principally interested in the evolution of the protein–DNA binding specificity, we characterized each snapshot by counting the number of contacts between the protein and the DNA bases. Each contact between heavy atoms scored 1 for distances r_{ij} below 4 Å (using shorter distances would result in many transient 'breaks' that add noise to the analysis). In order to further increase the robustness, we used a buffer zone from 4 Å to 5 Å over which the score was modulated with a sigmoidal function *s(i,j)* of the distance r_{ij} between the atoms *i* and *j*:

$$s(i, j) = \frac{1}{1 + e^{10*(r_{ij}-4.5)}}$$

This analysis yielded a 74 (amino acid) by 34 (DNA base) matrix for each snapshot. The overall distance **d(x,y)** between any two such matrices **x** and **y** was then calculated using the Manhattan algorithm (54).

$$\mathbf{d}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{N} \left|\mathbf{x}_i - \mathbf{y}_i\right|$$

Next, the Ward agglomerative hierarchical clustering method (55–57) was used to classify the different snapshots into groups by minimizing the variance within each cluster and increasing the weighted squared distance between cluster centers. The distance matrix and cluster representation were carried out using the R software package (58).

Binding specificity analysis

SKN-1 binding specificity was determined for any chosen snapshot from the MD trajectory (after a brief Cartesian coordinate energy minimization to remove bond length and base plane deformations) using the so-called ADAPT approach (28,29) implemented within the JUMNA program (59). This consists of calculating the complexation energy of the SKN-1/DNA complex for all possible DNA base sequences and deriving a PWM. In order to do this, it is necessary to thread all possible base sequences into the DNA oligomer within the complex, adapting the protein–DNA interface in each case using internal coordinate energy minimization. This was performed with the same AMBER parameterization used for the MD simulations, but replacing the explicit solvent and ion shell with a simple continuum model using a sigmoidal distance-dependent dielectric function and reduced phosphate charges (29). In parallel, an identical base sequence is threaded into the average conformation of the isolated DNA oligomer and energy minimization is again performed. Finally, another energy minimization is performed for the isolated protein (with flexibility limited to the side chains included within the interface cutoff distance, see below). Subtracting the isolated DNA oligomer and protein energies from the SKN-1/DNA complex energy yields the complex formation energy, which can be further analyzed in terms of two components: the DNA deformation energy and the protein–DNA interaction energy. In the present case, the nine central base pairs of the DNA oligomer were scanned, corresponding to the SKN-1 cognate site flanked by two extra base pairs on either end, leading to $4^9 = 262{,}144$ possible sequences. ADAPT calculations were accelerated using a divide-and-conquer technique, breaking each sequence into overlapping 4 bp fragments and thus reducing the total number of calculations to $6 \times 4^4 = 1024$ (for the complex and for the isolated DNA oligomer), without significant loss of accuracy (29). Protein flexibility was also limited to side chains within 20 Å of the protein–DNA interface. The energies resulting from this analysis were converted into PWMs using the WebLogo software (60). Finally, by analyzing the binding specificity derived from the sequence-dependent DNA deformation energy, or from the sequence-dependent protein–DNA interaction energy we could also analyze specificity in terms of its indirect and direct components.

We remark that for the purpose of this study we extended the utility programs associated with ADAPT to be able to derive a single PWM from a number of MD snapshots. In this case, ADAPT calculations were based on sequence-dependent energy differences with respect to the minimum energy for each snapshot, enabling us to overcome sequence-independent energy changes mainly caused by the necessary simplification of the electrostatic calculations (which rely on a rudimentary implicit solvent representation). Using this approach it was possible to describe the sequence selectivity of each of the conformational substates detected by the cluster analysis and to compare this to the consensus selectivity for the entire MD simulation, or to experimental binding data.

RESULTS AND DISCUSSION

We begin by considering the general impact of SKN-1 binding on DNA structure and dynamics. As shown in Figure 1A (see also Supplementary Figure S1) the protein inserts its long C-terminal α-helix in the major groove of the DNA binding site, while its N-terminal arm binds to the adjacent minor groove. In addition to amino acid side chain contacts with the DNA bases, the crystal structure of the

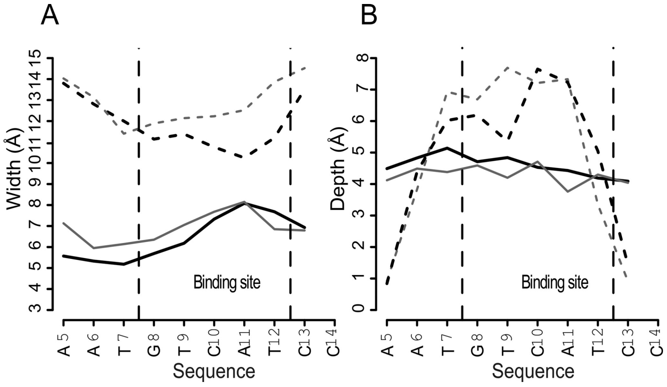


Figure 2. DNA groove dimensions (Å), width (**A**) and depth (**B**), within the isolated DNA oligomer (gray lines) and within the SKN-1/DNA complex (thick black lines). Major groove dimensions are indicated with dashed lines and minor groove dimensions with solid lines. Vertical dashed lines indicate the protein-binding site.

complex is stabilized by seven salt bridges involving seven arginines (R503, R506, R507, R508, R516, R521, R522). Of these residues, four (R503, R506, R507, R508) belong to the central support region (see Supplementary Figure S1) and three (R516, R521, R522) are located in the C-terminal helix of the protein. These interactions link the protein with the phosphate groups at positions G8 and C10 in the Watson strand and positions G10', T11', G14' and G15' in the Crick strand.

Comparing the average structures derived from the MD simulations of the SKN-1/DNA complex and of DNA alone, we see that protein binding has relatively little structural impact. There are no major changes in helical parameters or backbone parameters, although the average twist along the binding site increases by 2° in the presence of the protein. We also observe slight bending of the DNA toward the protein (6.5° versus 2.5° in the isolated DNA oligomer), but this value is less than that seen in the crystal structure (22°). These changes are coupled to a change in groove geometry, as shown in Figure 2. Insertion of the C-terminal α-helix in the major groove leads to a decrease in width of roughly 2 Å at positions C10-C13 and a localized decrease in depth at position T9. The binding of the N-terminal tail has a smaller effect on the minor groove (positions 5–7), where we see a narrowing of roughly 1 Å coupled with a small increase in depth.

Before passing to an analysis of the dynamics of the SKN-1/DNA complex, we lastly consider the effect of protein binding on the ionic environment of DNA. As shown in Figure 3, protein binding, not surprisingly, almost completely removes potassium cations from the major groove between positions T6 and C13, whereas we observe roughly 1–2 M potassium in this region for isolated DNA. In compensation, the K^+ molarity increases in the minor groove of the binding site, notably with a strongly localized ion site at the step G8-T9 that is absent in the isolated DNA oligomer. Secondary increases in potassium molarity are also seen at A11-T12 in the minor groove and at C13-C14 in the major groove.

We now turn to the dynamics of the SKN-1/DNA complex. The first observation is that DNA backbone dynamics decrease in the presence of the protein. This is illustrated in Figure 4 using the root mean square fluctuations of the phosphate atoms. We recall that these values were obtained by analyzing the position of the phosphorus atoms within each MD snapshot using curvilinear helicoidal coordinates with respect to the instantaneous helical axis, and then replotting them in Cartesian space with respect to the helical axis of the average DNA structure (52). This has the effect of removing any fluctuations due to DNA bending, stretching or twisting and gives an accurate view of phosphorus atom mobility. The protein clearly reduces the mobility of the phosphate groups within the binding site and the effect is particularly strong for the phosphates involved in salt bridges with SKN-1 (see Figure 4B).

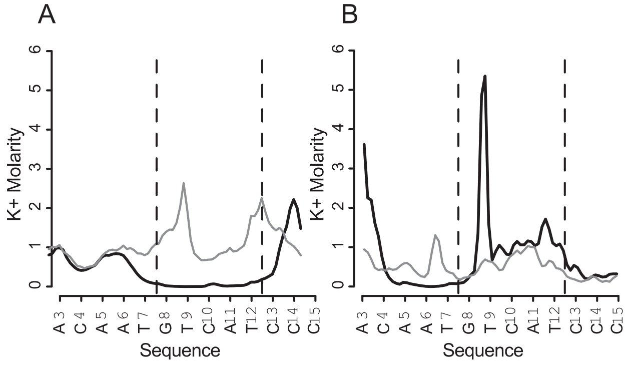


Figure 3. Potassium ion molarity: (**A**) inside the major groove and (**B**) inside the minor groove for the isolated DNA oligomer (gray lines) and for the DNA/SKN-1 complex (thick black lines). The sequences of both strand are shown in the 5′-3′ direction. Vertical dashed lines indicate the protein-binding site.

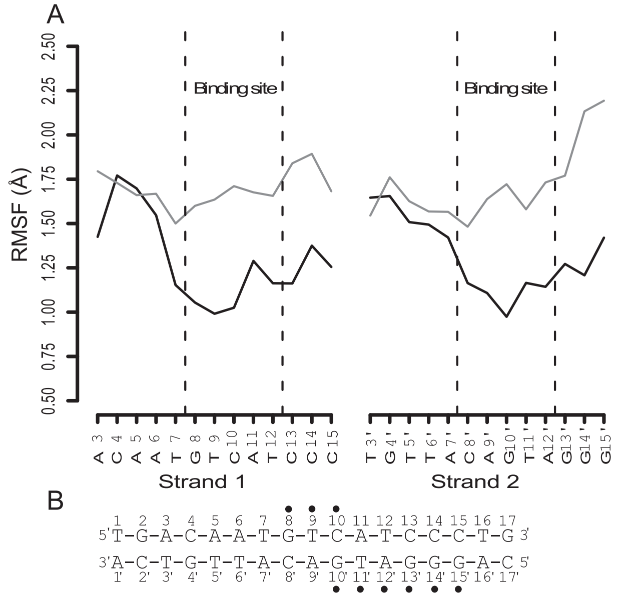


Figure 4. (**A**) Root mean square fluctuation (Å) of phosphorus atoms within the isolated DNA oligomer (gray lines) and within the DNA/SKN-1 complex (thick black lines). The sequences of both strand are shown in the 5′-3′ direction. Vertical dashed lines indicate the protein-binding site. (**B**) Black circles show the position of salt bridges within the DNA/SKN-1 complex.

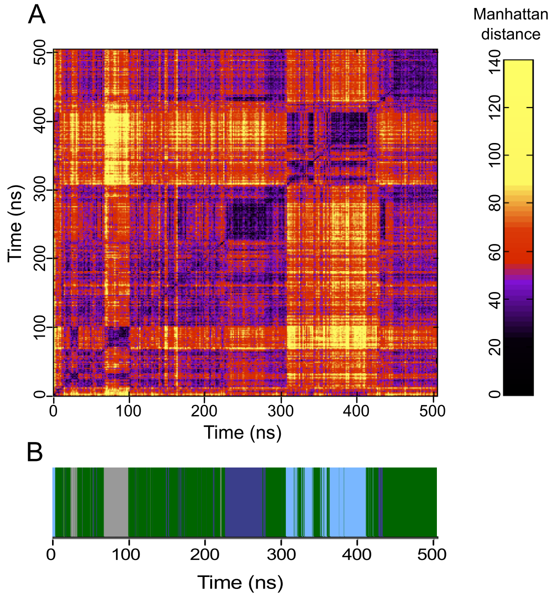
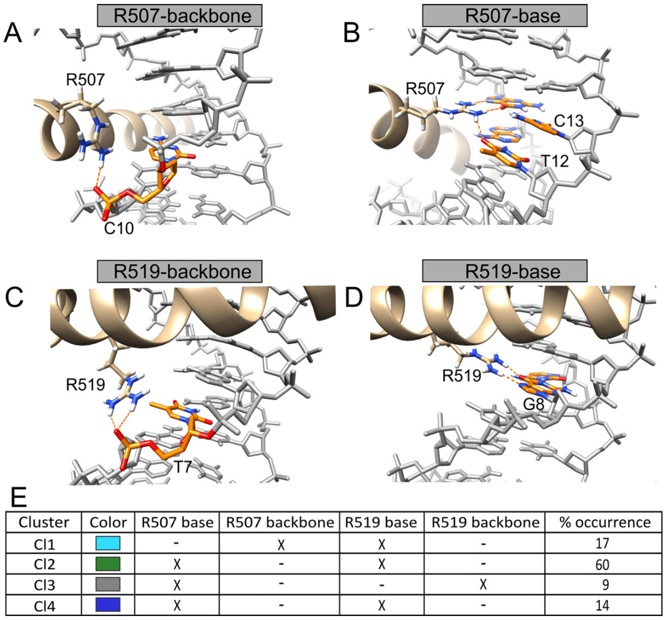


Figure 5. Clustering snapshots from the 500 ns MD trajectory of the DNA/SKN-1 complex: (**A**) Manhattan distance matrix. The vertical black to yellow scale represents increasing distances. (**B**) Clustering using the distance matrix leads to four distinct clusters whose appearance during the trajectory is indicated by the colors cyan (CL1), green (CL2), gray (CL3) and dark blue (CL4).



Cluster	Color	R507 base	R507 backbone	R519 base	R519 backbone	% occurrence
Cl1	🟦	-	X	X	-	17
Cl2	🟩	X	-	X	-	60
Cl3	◼	X	-	-	X	9
Cl4	🟦	X	-	X	-	14

Figure 6. Alternative orientations observed for arginines 507 (**A**, **B**) and 519 (**C**, **D**). Orange dashed lines indicate hydrogen bonds between these arginines and DNA. The table (**E**) shows the link between the clusters observed during the MD trajectory and the R507/R519 orientations in addition to the percentage occurrence of each cluster during the trajectory.

In contrast to this apparent rigidification, we see significant dynamics at the protein–DNA interface. Note that Figure 4B indicates nine salt bridges, in contrast to the seven seen in the crystal structure. This change is indicative of what occurs during the MD simulation where we see many intermittent protein–DNA contacts. Most of these are alternative interactions involving the same side chains that form salt bridges in the crystal structure, although some are completely new, notably involving Arg 457 and Lys 460 within the N-terminal tail. Table 1 shows contacts seen in both the crystal structure and the MD simulation in black, while those appearing only in the simulation are shown in bold/red. From these results, we can see that most interactions are only present for a fraction of the 0.5 μs trajectory, although those observed in the crystal structure are generally the longest lived. It also shows that interactions between given side chains and nucleotides often involve different sets of atoms, in some cases simultaneously, creating bidentate interactions.

On the basis of this finding, we decided to see if the interface dynamics were random or reflected the existence of conformational sub-states. As described in the methodology section we carried out this analysis by building a contact matrix between protein side chains and DNA bases for snapshots every 200 ps along the trajectory, leading to a total of 2500 matrices. Measuring the Manhattan distances between these matrices created a new distance matrix 2500 × 2500 that could then be analyzed to detect conformational clusters. The results shown in Figure 5 indicate that the MD trajectory is in fact composed of four distinct conformational clusters.

The initial cluster, CL1 (cyan) is closest to the X-ray conformation of the complex. It is lost after only 5 ns, but then reappears intermittently during the last third of the trajectory and finally constitutes 17% of the trajectory. The second cluster to appear, CL2 (green) is the most common and reappears throughout the simulation representing in total 60% of the trajectory. A third cluster, CL3 (gray) appears around 70 ns, but only makes up 9% of the trajectory and is not seen after the first 100 ns. The final cluster, CL4 (dark blue) appears in the middle of the simulation and again briefly toward the end, making up 14% of the total.

By extracting snapshots belonging to each of the four clusters we can analyze their structural characteristics. The first observation is that the CL2 (green) and CL4 (dark blue) clusters are very similar to one another, differing only by the position of the N-terminal arm, which interacts with the bases in the DNA minor groove in the more common CL2 (green) cluster (without affecting the groove geometry), but with the DNA backbone in the CL4 (dark blue) cluster. We will consequently temporarily group these two clusters together (and denominate them as CL2/4). The main feature distinguishing the remaining clusters turns out to be to the position of the side chains of two arginines: R507 and R519. In CL1, R507 lies close to the DNA backbone, intermittently forming a salt bridge with the phosphate of C10 or, more rarely, those of A11 and G15'. In contrast, in CL2/4 and CL3 it binds in the DNA major groove forming a bidentate interaction with O6 and N7 of G13' (as seen in other protein–DNA complexes (61,62)) and, intermittently, to O4 of T12. Similarly, in CL1 and CL2/4, R519 also forms a bidentate interaction with O6 and N7 of G8, whereas in CL3 it is close to the backbone, intermittently forming a salt bridge with the phosphate of T7. The alternate conformations of R507 and R519 are illustrated in Figure 6. As summarized in Figure 6E, the combination of these two side

Table 1. SKN-1 salt bridges with the DNA backbone (columns 1–3) and hydrogen bonds with the DNA bases (columns 4–6) are highly dynamic

Protein	DNA Backbone	Percentage	Protein	DNA base	Percentage
ARG 457(N)	**G 8(O2P)**	**28**	**ARG 457(NE)**	**T 5'(O2)**	**15**
LYS 460(N)	**T 9(O2P)**	**37**	**ARG 457(NH1)**	**T 5'(O2)**	**12**
ARG 503(NE)	G 15'(O2P)	52	**ARG 457(NH2)**	**T 5'(O2)**	**19**
ARG 503(NE)	**G 15'(O1P)**	**38**	**ARG 457(NH1)**	**T 6'(O2)**	**16**
ARG 503(NH2)	**G 15'(O2P)**	**20**	**ARG 507(NH2)**	**T 12(O4)**	**24**
ARG 503(NH2)	**G 15'(O1P)**	**49**	**ARG 507(NH1)**	**G 13'(N7)**	**25**
ARG 503(NH1)	**A 16'(O2P)**	**13**	**ARG 507(NH2)**	**G 13'(O6)**	**72**
ARG 503(NH2)	**A 16'(O2P)**	**19**	ASN 511(ND2)	T 11'(O4)	92
ARG 503(NH2)	**A 16'(O1P)**	**15**	ASN 511(OD1)	C 10(N4)	97
ARG 506(NE)	G 14'(O2P)	17	ARG 519(NH1)	G 8(N7)	78
ARG 506(NE)	G 14'(O1P)	96	ARG 519(NH2)	G 8(O6)	79
ARG 506(NH2)	G 14'(O2P)	98			
ARG 507(NH2)	**G 15'(O1P)**	**3**			
ARG 507(NE)	**G 15'(O2P)**	**2**			
ARG 507(NH2)	**A 11(O1P)**	**3**			
ARG 507(NH1)	C 10(O1P)	9			
ARG 508(NE)	**T 9(O2P)**	**20**			
ARG 508(NH2)	**T 9(O2P)**	**37**			
ARG 508(NE)	**T 9(O5')**	**11**			
ARG 508(NH1)	C 10(O1P)	50			
LYS 510(NZ)	**G 13'(O1P)**	**21**			
ARG 516(NE)	**G 8(O1P)**	**42**			
ARG 516(NH2)	G 8(O2P)	30			
ARG 516(NH2)	G 8(O1P)	37			
ARG 519(NH2)	**T 7(O1P)**	**5**			
ARG 519(NH1)	**T 7(O1P)**	**4**			
ARG 521(NE)	T 11'(O1P)	44			
ARG 521(NH2)	T 11'(O1P)	89			
ARG 521(NH1)	**A 12'(O1P)**	**59**			
ARG 521(NH2)	**A 12'(O1P)**	**15**			
ARG 521(NH2)	**A 12'(O5')**	**16**			
ARG 522(NE)	**G 10'(O1P)**	**28**			
ARG 522(NH2)	G 10'(O1P)	42			
ARG 525(NH1)	**T 11'(O2P)**	**12**			

The percentage time each interaction was observed during the 0.5 μs MD trajectory is given in columns 3 and 6. Interactions shown in bold/red are only observed in the MD trajectory, while those in black are seen in the crystal structure (35) and in the MD trajectory.

chain flips gives rise to three conformational sub-states that distinguish the clusters CL1, CL2/4 and CL3.

The dynamical behavior of R507 and R519 are illustrated by the time series of side chain-DNA backbone/base distances in Supplementary Figure S2, which, for reference, also shows the distance fluctuations for the R506 salt bridge with the phosphate of G14'. While the significant perturbations of the R506 interaction occur only occasionally, R507 and R519 show complex fluctuations whether they are interacting with DNA bases or DNA phosphates. Analyzing snapshots every picosecond along the MD trajectory, with distance and angle cutoffs of 3.5 Å and 135°, respectively, leads to lifetimes of less than 30 ps for either base or phosphate interactions. However, ignoring breaks that last no longer than 1 ps typically increases the lifetimes to 100–400 ps. By comparison, the R506 salt bridge has lifetimes of roughly 100 ps or 1800 ps, depending on whether 1 ps breaks are taken into account or ignored.

By applying our sequence-threading approach ADAPT to multiple snapshots belonging to each cluster (7, 12, 10 and 2 for CL1, CL2, CL3 and CL4, respectively), we were able test whether the very localized changes in the two key arginines have any significant impact on how SKN-1 is recognizing the DNA base sequence. The results are shown in Figure 7, where CL2 and CL4 have again been grouped together since they yield identical PWMs. If we concentrate on the bases at positions 8, 12 and 13**,** the results are relatively easy to interpret. When R519 interacts with position 8 in CL1, a 'G' appears strongly at this position in the PWM. Similarly when R507 interacts with positions 12 and 13 in

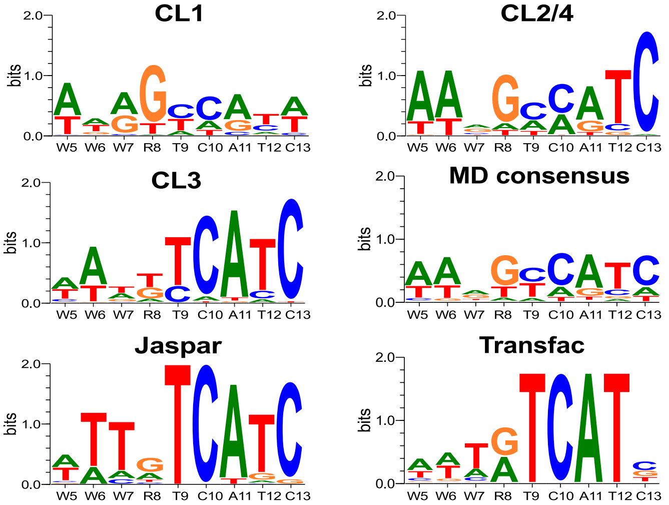


Figure 7. SKN-1 PWMs resulting from the analysis of snapshots belonging to each of the three distinct clusters and also a consensus PWM using the snapshots from the entire MD trajectory. These results can be compared to the experimental results from the JASPAR (63) and TRANSFAC databases (64) (W ≡A/T, R ≡ A/G).

CL2/4 and CL3, a clear 'TC' appears at these positions. Finally, when both arginines bind within the major groove in CL2/4, both a 'G' at position 8 and a 'TC' at positions 12 and 13 dominate. However, we can also see that the R507 groove interaction also impacts positions 10 and 11 at the 3′-end of the binding site and leads to the appearance of the CATC motif in both CL2/4 and CL3. As expected the majority of the recognition in each cluster comes from direct protein–DNA contacts. Although some base pairs show selectivity due to DNA deformation (notably for T at positions 10 and 13, see Supplementary Figure S3), protein–DNA interaction is clearly the dominant factor in the overall PWM.

We remark that the movement of the N-terminal tail does not appear to have any significant impact on the PWM since the A/T-rich preference seen at the 5′-end of the SKN-1 binding site, corresponding to the location of the N-terminal tail, is virtually unchanged whether the tail lies within the minor groove (CL2 and CL3), or closer to the DNA backbones (CL1 and CL4). Supplementary Figure S4 shows one such comparison for the clusters CL2 and CL4. We conclude that its role is largely electrostatic (its cationic residues favoring the more negative minor groove potentials generated by AT base pairs) and does not require binding to a specific base site.

We can make this analysis of selectivity more quantitative by calculating Pearson correlation coefficients (PCCs) between the PWMs of the various clusters and the experimental results. We limit our analysis to the PWM for SKN-1 from the JASPAR database (63), but remark that very similar results are obtained with the equivalent data in TRANSFAC (64). The overall correlation between CL1, CL2/4 and CL3 PWMs with the JASPAR data is 0.50, 0.52 and 0.82, respectively. Thus CL3 is closest to the experimental data (which can be seen visually in Figure 7). However, if we now look at the correlations at each position within the binding site, another picture emerges. At position 8, the correlations for CL1, CL2/4 and CL3 become 0.89, 0.95 and 0.29, respectively. Thus, only CL1 and CL2/4 (where R519 is bound in the DNA groove) reproduce the experimental result. In contrast, at positions 12 and 13, the correlations for CL1, CL2/4 and CL3 change again to (0.84, -0.50), (0.99, 1.0) and (0.97, 1.0) and thus only CL2/4 and CL3 (where R507 is bound in the DNA groove) fit the experiments. This confirms the notion that each conformational sub-state is recognizing only part of the binding site. In addition, we can note that these partial recognition events are not fully compatible with one another since the consensus correlation between the simulation (using all the snapshots extracted from the MD run) and the JASPAR PWM is only 0.57. This loss of selectivity can also be quantified by calculating the total information content of the various PWMs (65), which yields 6.2, 9.0 and 9.5 for CL1, CL2/4 and CL3, respectively, but only 5.3 for the MD consensus. In contrast, if we model recognition events occurring separately in different regions of the binding site, by combining columns 1–4 from the PWM of CL1 with columns 5–9 from the PWM of CL3, the total information content becomes 10.5, close to that of the experimental JASPAR logo (11.6).

CONCLUSIONS

This computational study of the transcription factor SKN-1 bound to its cognate DNA site shows that the protein–DNA interface is dynamic and, notably, that two arginine side chains oscillate between the formation of direct interactions with DNA bases and interactions with the DNA backbone. The cationic N-terminal arm of SKN-1 undergoes similar oscillations. This dynamics is analogous to what has been seen at protein-protein interfaces (66,67) and is compatible with recent NMR studies and simulation studies showing that protein–DNA salt bridges are broken on sub-nanosecond timescales (21,25). In our case, the temporary loss of protein-base interactions significantly alters sequence selectivity and suggests that the observed consensus binding sequence of the transcription factor exists as the time-averaged ensemble of a number of distinct conformational sub-states that each recognize different parts of the binding site. As other authors have already pointed out, the dynamic nature of the protein–DNA interface may aid binding both by making the transition between non-specific and specific sites easier and by reducing the entropic penalty for binding. From a computational point of view the 0.5 μs simulations carried out here led to the detection of four distinct sub-states, but we cannot exclude that this number would grow with longer simulations, or that the the relative sub-state populations could evolve. We conclude that understanding protein–DNA recognition mechanisms using molecular simulations, at least in some cases, may very well require trajectories on the microsecond scale.

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

This work is funded by the ANR project CHROME ANR-12-BSV5-0017-01 and GENCI for a generous allocation of supercomputer resources at the CINES center in Montpellier. L.E. was funded by a doctoral grant from Rhône-Alpes ARC 1 Santé. Funding for open access charge: ANR project CHROME ANR-12-BSV5-0017-01.
Conflict of interest statement. None declared.

REFERENCES

1. Kalodimos,C.G., Biris,N., Bonvin,A.M., Levandoski,M.M., Guennuegues,M., Boelens,R. and Kaptein,R. (2004) Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. *Science*, **305**, 386–389.
2. Kalodimos,C.G., Boelens,R. and Kaptein,R. (2004) Toward an integrated model of protein-DNA recognition as inferred from NMR studies on the Lac repressor system. *Chem. Rev.*, **104**, 3567–3586.
3. Iwahara,J., Zweckstetter,M. and Clore,G.M. (2006) NMR structural and kinetic characterization of a homeodomain diffusing and hopping on nonspecific DNA. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 10562–10567.

4. Iwahara,J. and Clore,G.M. (2006) Direct observation of enhanced translocation of a homeodomain between DNA cognate sites by NMR exchange spectroscopy. *J. Am. Chem. Soc.*, **128**, 404–405.
5. Clore,G.M., Tang,C. and Iwahara,J. (2007) Elucidating transient macromolecular interactions using paramagnetic relaxation enhancement. *Curr. Opin. Struct. Biol.*, **17**, 603–616.
6. Mirny,L., Slutsky,M., Wunderlich,Z., Tafvizi,A., Leith,J. and Kosmrlj,A. (2009) How a protein searches for its site on DNA: the mechanism of facilitated diffusion. *J. Phys. A Math. Theor.*, **42**, 434013.
7. Das,R.K. and Kolomeisky,A.B. (2010) Facilitated search of proteins on DNA: correlations are important. *Phys. Chem. Chem. Phys.*, **12**, 2999–3004.
8. Koslover,E.F., Díaz de la Rosa,M.A. and Spakowitz,A.J. (2011) Theoretical and computational modeling of target-site search kinetics in vitro and in vivo. *Biophys. J.*, **101**, 856–865.
9. Kolomeisky,A.B. (2011) Physics of protein-DNA interactions: mechanisms of facilitated target search. *Phys. Chem. Chem. Phys.*, **13**, 2088–2095.
10. Furini,S., Domene,C. and Cavalcanti,S. (2010) Insights into the sliding movement of the lac repressor nonspecifically bound to DNA. *J. Phys. Chem. B*, **114**, 2238–2245.
11. Temiz,A.N., Benos,P.V. and Camacho,C.J. (2010) Electrostatic hot spot on DNA-binding domains mediates phosphate desolvation and the pre-organization of specificity determinant side chains. *Nucleic Acids Res.*, **38**, 2134–2144.
12. 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.
13. Chen,C. and Pettitt,B.M. (2011) The binding process of a nonspecific enzyme with DNA. *Biophys. J.*, **101**, 1139–1147.
14. Furini,S., Barbini,P. and Domene,C. (2013) DNA-recognition process described by MD simulations of the lactose repressor protein on a specific and a non-specific DNA sequence. *Nucleic Acids Res.*, **41**, 3963–3972.
15. Ando,T. and Skolnick,J. (2014) Sliding of proteins non-specifically bound to DNA: Brownian dynamics studies with coarse-grained protein and DNA models. *PLoS Comput. Biol.*, **10**, e1003990.
16. Sela,I. and Lukatsky,D. (2011) DNA sequence correlations shape nonspecific transcription factor-DNA binding affinity. *Biophys. J.*, **101**, 160–166.
17. Afek,A. and Lukatsky,D.B. (2013) Positive and negative design for nonconsensus protein-DNA binding affinity in the vicinity of functional binding sites. *Biophys. J.*, **105**, 1653–1660.
18. Afek,A. and Lukatsky,D. (2013) Genome-wide organization of eukaryotic preinitiation complex is influenced by nonconsensus protein-DNA binding. *Biophys. J.*, **104**, 1107–1115.
19. 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.
20. Fuxreiter,M., Simon,I. and Bondos,S. (2011) Dynamic protein-DNA recognition: beyond what can be seen. *Trends Biochem. Sci.*, **36**, 415–423.
21. Anderson,K.M., Esadze,A., Manoharan,M., Bru schweiler,R., Gorenstein,D.G. and Iwahara,J. (2013) Direct observation of the ion-pair dynamics at a protein–DNA interface by NMR spectroscopy. *J. Am. Chem. Soc.*, **135**, 3613–3619.
22. Zandarashvili,L., Esadze,A. and Iwahara,J. (2013) NMR studies on the dynamics of hydrogen bonds and ion pairs involving lysine side chains of proteins. *Adv. Protein Chem. Struct. Biol.*, **93**, 37–80.
23. Zandarashvili,L. and Iwahara,J. (2014) Temperature dependence of internal motions of protein side-chain NH3+ groups: insight into energy barriers for transient breakage of hydrogen bonds. *Biochemistry*, **54**, 538–545.
24. Esadze,A., Li,D.W., Wang,T., Brüschweiler,R. and Iwahara,J. (2011) Dynamics of lysine side-chain amino groups in a protein studied by heteronuclear 1H−15N NMR spectroscopy. *J. Am. Chem. Soc.*, **133**, 909–919.
25. Chen,C., Esadze,A., Zandarashvili,L., Nguyen,D., Pettitt,B.M. and Iwahara,J. (2015) Dynamic equilibria of short-range electrostatic interactions at molecular interfaces of protein–DNA complexes. *J. Phys. Chem. Lett.*, **6**, 2733–2737.
26. Garton,M. and Laughton,C. (2013) A comprehensive model for the recognition of human telomeres by TRF1. *J. Mol. Biol.*, **425**, 2910–2921.
27. Lafontaine,I. and Lavery,R. (2001) High-speed molecular mechanics searches for optimal DNA interaction sites. *Comb. Chem. High Throughput Screen.*, **4**, 707–717.
28. Paillard,G. and Lavery,R. (2004) Analyzing protein-DNA recognition mechanisms. *Structure*, **12**, 113–122.
29. Deremble,C., Lavery,R. and Zakrzewska,K. (2008) Protein-DNA recognition: Breaking the combinatorial barrier. *Comput. Phys. Commun.*, **179**, 112–119.
30. Zakrzewska,K., Bouvier,B., Michon,A., Blanchet,C. and Lavery,R. (2009) Protein–DNA binding specificity: a grid-enabled computational approach applied to single and multiple protein assemblies. *Phys. Chem. Chem. Phys.*, **11**, 10712–10721.
31. Berta,P., Hawkins,J.B., Sinclair,A.H., Taylor,A., Griffiths,B.L., Goodfellow,P.N. and Fellous,M. (1990) Genetic evidence equating SRY and the testis-determining factor. *Nature*, **348**, 448–450.
32. 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.
33. Bowerman,B., Eaton,B.A. and Priess,J.R. (1992) skn-1, a maternally expressed gene required to specify the fate of ventral blastomeres in the early C. elegans embryo. *Cell*, **68**, 1061–1075.
34. Bowerman,B., Draper,B.W., Mello,C.C. and Priess,J.R. (1993) The maternal gene skn-1 encodes a protein that is distributed unequally in early C. elegans embryos. *Cell*, **74**, 443–452.
35. 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.
36. 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.
37. 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.
38. Oliveira,R.P., Abate,J.P., Dilks,K., Landis,J., Ashraf,J., Murphy,C.T. and Blackwell,T.K. (2009) Condition-adapted stress and longevity gene regulation by Caenorhabditis elegans SKN-1/Nrf. *Aging Cell*, **8**, 524–541.
39. Staab,T.A., Griffen,T.C., Corcoran,C., Evgrafov,O., Knowles,J.A. and Sieburth,D. (2013) The conserved SKN-1/Nrf2 stress response pathway regulates synaptic function in Caenorhabditis elegans. *PLoS Genet.*, **9**, e1003354.
40. 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.
41. 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.
42. 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.
43. 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.
44. 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.
45. 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.
46. Essmann,U., Perera,L., Berkowitz,M.L., Darden,T., Lee,H. and Pedersen,L.G. (1995) A smooth particle mesh Ewald method. *J. Chem. Phys.*, **103**, 8577–8593.
47. 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.

48. 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.
49. 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.
50. 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.
51. Lavery,R., Maddocks,J.H., Pasi,M. and Zakrzewska,K. (2014) Analyzing ion distributions around DNA. *Nucleic Acids Res.*, **42**, 8138–8149.
52. 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.
53. Case,D.A., Berryman,J., Betz,R.M., Cerutti,D., Cheatham,T. III, Darden,T., Duke,R., Glese,T., Gohlke,H. *et al.* (2015) *AMBER 2015*.
54. Krause,E.F. (1987) *Taxicab geometry: an adventure in non-Euclidean geometry* . Courier Corporation, Dover, London .
55. Ward,J.H. Jr (1963) Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, **58**, 236–244.
56. Kaufman,L. and Rousseeuw,P.J. (2009) *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, New York .
57. Murtagh,F. and Legendre,P. (2014) Wards hierarchical agglomerative clustering method: which algorithms implement wards criterion? *J. Classif.*, **31**, 274–295.
58. R Development Core Team (2009) *R: A language and environment for statistical computing*.
59. Lavery,R., Zakrzewska,K. and Sklenar,H. (1995) JUMNA (Junction Minimization of Nucleic-Acids). *Comput. Phys. Commun.*, **91**, 135–158.
60. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
61. McClarin,J.A., Frederick,C.A., Wang,B.-C., Greene,P., Boyer,H.W., Grable,J. and Rosenberg,J.M. (1986) Structure of the DNA-Eco RI endonuclease recognition complex at 3 A resolution. *Science*, **234**, 1526–1541.
62. Otwinowski,Z., Schevitz,R.W., Zhang,R.G., Lawson,C.L., Joachimiak,A., Marmorstein,R.Q., Luisi,B.F. and Sigler,P.B. (1988) Crystal structure of trp represser/operator complex at atomic resolution. *Nature*, **335**, 321–329.
63. Mathelier,A., Zhao,X., Zhang,A.W., Parcy,F., Worsley-Hunt,R., Arenillas,D.J., Buchman,S., Chen,C.-Y., Chou,A. and Ienasescu,H. (2013) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **42**, D142–D147.
64. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
65. Schneider,T.D., Stormo,G.D., Gold,L. and Ehrenfeucht,A. (1986) Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, **188**, 415–431.
66. Lee,H.J., Hota,P.K., Chugha,P., Guo,H., Miao,H., Zhang,L., Kim,S.-J., Stetzik,L., Wang,B.-C. and Buck,M. (2012) NMR structure of a heterodimeric SAM: SAM complex: characterization and manipulation of EphA2 binding reveal new cellular functions of SHIP2. *Structure*, **20**, 41–55.
67. Zhang,L. and Buck,M. (2013) Molecular simulations of a dynamic protein complex: role of salt-bridges and polar interactions in configurational transitions. *Biophys. J.*, **105**, 2412–2417.