doi:10.1016/j.jmb.2010.06.026 J. Mol. Biol. (2010) 401, 415–432

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Mechanism of DNA Recognition by the Restriction Enzyme EcoRV

Mai Zahran[1], Isabella Daidone[2], Jeremy C. Smith[3] and Petra Imhof[1]*

[1]*Computational Molecular Biophysics, IWR, University of Heidelberg, 69120 Heidelberg, Germany*

[2]*Chemistry Department, University of L'Aquila, Via Vetoio (Coppito 1), 67010 L'Aquila, Italy*

[3]*Center of Molecular Biophysics, Oak Ridge National Laboratory, PO Box 2008, MS6164, Oak Ridge, TN 37831-6164, USA*

Received 22 December 2009; received in revised form 11 June 2010; accepted 13 June 2010
Available online
18 June 2010

Edited by D. Case

EcoRV, a restriction enzyme in *Escherichia coli*, destroys invading foreign DNA by cleaving it at the center step of a GATATC sequence. In the EcoRV–cognate DNA crystallographic complex, a sharp kink of 50° has been found at the center base-pair step (TA). Here, we examine the interplay between the intrinsic propensity of the cognate sequence to kink and the induction by the enzyme by performing all-atom molecular dynamics simulations of EcoRV unbound and interacting with three DNA sequences: the cognate sequence, GATATC (TA); the non-cognate sequence, GAATTC (AT); and with the cognate sequence methylated on the first adenine $GA_{CH_3}TATC$ (TA-CH_3). In the unbound EcoRV, the cleft between the two C-terminal subdomains is found to be open. Binding to AT narrows the cleft and forms a partially bound state. However, the intrinsic bending propensity of AT is insufficient to allow tight binding. In contrast, the cognate TA sequence is easier to bend, allowing specific, high-occupancy hydrogen bonds to form in the complex. The absence of cleavage for this methylated sequence is found to arise from the loss of specific hydrogen bonds between the first adenine of the recognition sequence and Asn185. On the basis of the results, we suggest a three-step recognition mechanism. In the first step, EcoRV, in an open conformation, binds to the DNA at a random sequence and slides along it. In the second step, when the two outer base pairs, GAxxTC, are recognized, the R loops of the protein become more ordered, forming strong hydrogen-bonding interactions, resulting in a partially bound EcoRV–DNA complex. In the third step, the flexibility of the center base pair is probed, and in the case of the full cognate sequence the DNA bends, the complex strengthens and the protein and DNA interact more closely, allowing cleavage.

Keywords: protein–DNA interaction; EcoRV; direct/indirect readout; sequence recognition

Introduction

Significant structural alteration of DNA upon binding by specific proteins is commonplace in molecular biology. DNA bending is important in DNA replication, repair, recombination and methylation.[1–7] This bending can be relatively smooth, as in the curvature of nucleosomal DNA to facilitate packing, or can be sharp and localized, such as, for example, the DNA complexed to catabolite activator protein.[8–10] Many DNA-bending proteins are also site-specific enzymes for which the bending event is coupled to the reaction rate at the cognate site.

X-ray crystallographic studies have shown that most specific protein–DNA complexes possess extensive hydrogen bonding between the protein and specific functional groups on the DNA bases, constituting a "direct readout" mechanism.[11] However, in some protein–DNA structures, the direct hydrogen-bonding interactions are insufficient to explain the experimentally observed specificity.[12–14] Therefore, an "indirect readout" mechanism has been proposed in which the sequence-dependent conformation of DNA structure is recognized through protein contacts with the sugar–phosphate backbone and/or with non-specific portions of the bases.[14] Thus, specificity in protein–DNA interac-

**Corresponding author.* E-mail address: petra.imhof@iwr.uni-heidelberg.de.

Abbreviations used: MD, molecular dynamics; RMSF, root-mean-square fluctuation.

tions can be mediated by both the recognition of sequence-dependent DNA conformations and the way in which the base sequence influences the induced-fit transitions often required to form productive complexes.[15]

Restriction endonucleases provide interesting model systems for the investigation of sequence-specific protein–DNA interactions. More than 3000 type II restriction endonucleases have been identified, representing over 200 different sequence specificities.[16] The ability of bacterial cells to resist invading foreign DNA is dependent on the extraordinarily high fidelity of this recognition process, in which target sites are selected from an enormous molar excess of structurally similar non-specific DNA.[17,18] Restriction enzymes enhance the rate of the DNA strand scission at specific sites by an estimated 10^{15}-fold.[19] A single incorrect base pair in a 4- to 6-bp target site reduces $k_{\mathrm{cat}}/K_{\mathrm{M}}$ by 10^6-fold or more.[20–22] During the cognate sequence search process, these enzymes bind non-cognate sequences without inducing a bend.[23,24] Recognition sequences in bacterial DNA are protected from cleavage by methylation of the recognition site.[25] Together with their usefulness as "DNA scissors" in molecular biology, the high specificity of restriction enzymes makes them important systems for understanding protein–DNA interactions.

EcoRV, a restriction enzyme from *Escherichia coli*, is one of the best-characterized endonucleases. It consists of a dimer of 245 amino acid residues per monomer. The enzymatic role of EcoRV is to cleave the foreign sequence 5′-GATATC-3′ at the center TA step in a blunt-ended fashion, generating 5′-phosphate groups in a Mg^{2+}-dependent reaction.[26] *E. coli*'s own DNA is methylated at the first adenine of the recognition sequence, $\mathrm{GA}_{\mathrm{CH}_3}\mathrm{TATC}$, being thus protected against cleavage.[27] The existence of an open conformation was not ruled out, although there is no experimental evidence for an open state of free EcoRV in solution.[28] Crystal structures of EcoRV and of the cognate EcoRV–DNA complex (Fig. 1a and b) reveal extensive conformational changes in both protein and DNA upon complex formation.[29] In the complex, the central TA is directly recognized only *via* hydrophobic contacts with the thymine methyl groups and the DNA is sharply bent by a 50° roll into the major groove at this position, unstacking the bases.[30] Consistent with these findings, theoretical calculations and exhaustive analysis of known DNA and protein–DNA structures indicate that the TA step is highly flexible and easier to unstack than other dinucleotides.[31–33]

Thus, sequence-dependent differences in free energies for unstacking the center step represent an intrinsic property that EcoRV may exploit in order to generate cleavage specificity.

When complexed to EcoRV, the DNA is positioned in a cleft between the two protein monomer chains (Fig. 1a) and makes contacts primarily with two peptide loops from each monomer as well as with other segments of the protein contacting the DNA phosphate groups. The R (recognition) loop, comprising residues 182–188, lies in the major groove of the DNA and makes several hydrogen bonds with bases of the recognition sequence. The Q loop, so-called due to the presence of two glutamines between residues 67 and 72, interacts extensively with the sugar–phosphate backbone in the minor groove, placing the scissile phosphodiester bond in the active site of the enzyme. In the crystal structures of the free protein without the DNA, both the R and the Q loops are largely disordered.[26,34]

To further understand the origins of DNA sequence discrimination by EcoRV, a detailed kinetic and crystallographic study of the interaction of EcoRV with the cognate TA-sequence GATATC and with a non-cognate AT-sequence, GAATTC has been performed.[35] The latter is recognized by another restriction enzyme, EcoRI. Analysis of the DNA binding and bending by equilibrium and stopped-flow fluorescence quenching and fluorescence resonance energy transfer methods demonstrated that the capacity of EcoRV to bend the AT sequence site is severely limited compared to that of the cognate complex (EcoRV–TA).[35]

To better understand the role of the two central bases (TA), several substitutions of the above two nucleotides have been made.[29,35] The structural consequences of the substitutions were well characterized by the crystallographers.[29,35] The results suggest that indirect readout by EcoRV depends significantly on the different free-energy cost of unstacking the central TA base-pair step relative to other base pairs at this position.[29] Structural adaptation at the protein–DNA interface was seen directly in the crystal structure of EcoRV with an analog site reduced in activity by nearly 8 orders of magnitude.[29] Direct interactions appear insufficient to explain the observed specificities, and structure-based mutational analysis has confirmed the importance of sequence-dependent DNA conformational preferences.[36,37]

The current knowledge of sequence-dependent DNA conformation is insufficient to determine whether a particular DNA deformation observed in a protein–DNA complex is induced by protein binding or is an inherent property of a particular nucleotide sequence. The sequence-dependent deformability of the DNA duplexes has been studied with a range of experimental methods[38,39] as well as computational modeling and simulation. The latter have been based on molecular mechanics calculations or on molecular dynamics (MD) simulation in explicit solvent.[32,40–43]

For complex systems, such as protein–DNA complexes, MD simulation is a powerful way of obtaining information on structure and dynamics at the atomic level and has been used to analyze interactions between proteins and DNA.[44–49] For example, Falconi *et al.* investigated the structural dynamics of the DNA binding domains of the human papillomavirus strain 16 and the bovine papillomavirus strain 1, complexed with their DNA targets, using both MD and nuclear magnetic resonance. [49] They observed a good agreement of

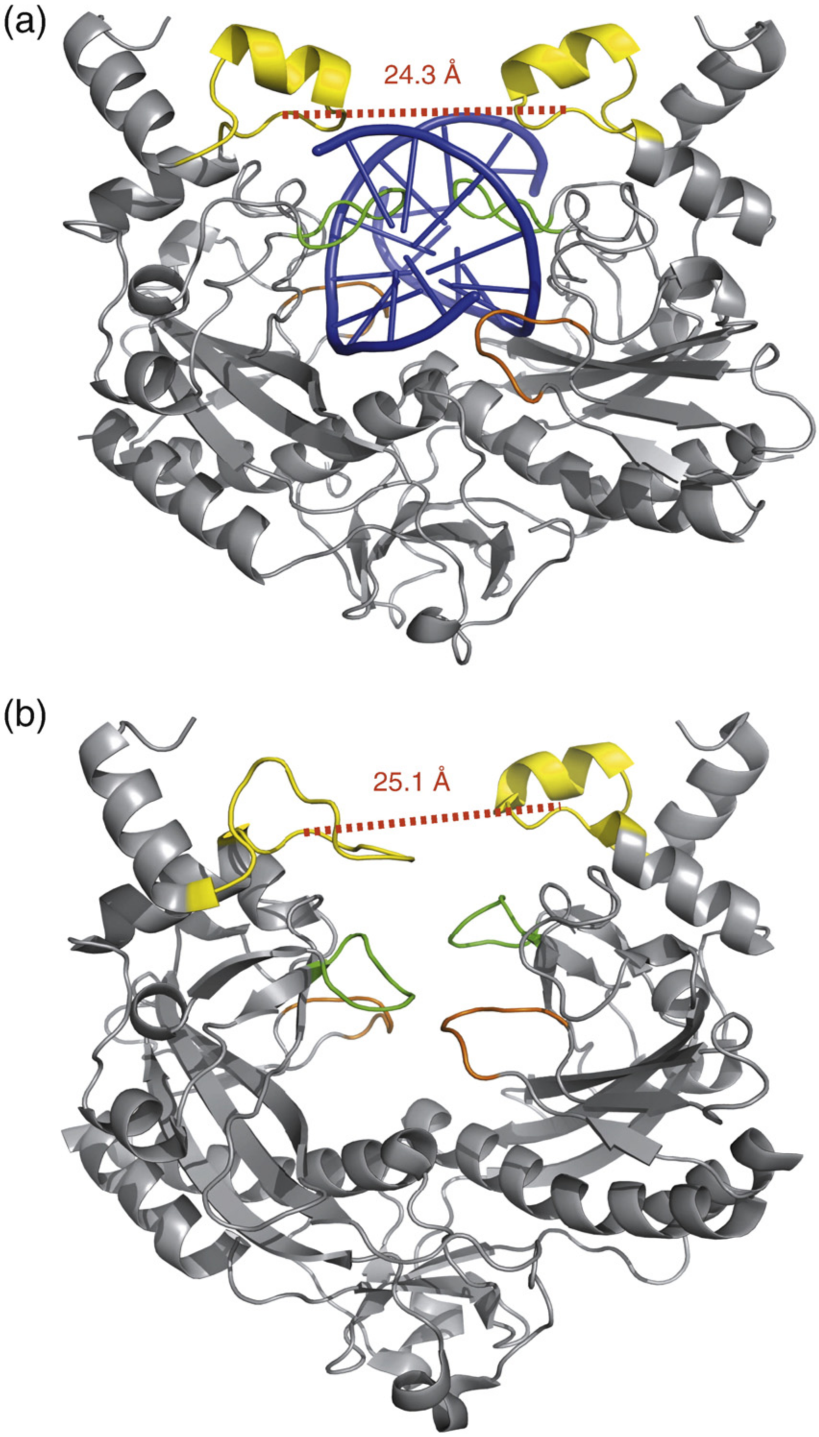

Fig. 1. Ribbon representation of the crystal structures of EcoRV in bound (a) and unbound (b) states, with 1SX8 and 1RVE as PDB identifier, respectively. The C-terminal domains, R loops and Q loops are shown in yellow, green and orange, respectively.

the root-mean-square fluctuation (RMSF) MD values with respect to the crystallographic *B*-factors and a good agreement of the hydrogen bonds found in MD with those found in the X-ray and NMR structures.[49] Studies on proteins have shown that structural properties such as the RMSF, radii of gyration, solvent-accessible surface, secondary structure, or hydrogen-bond propensities are in agreement with experimental data and relatively insensitive to the force field used.[50,51]

Agreement between the simulated and experimental data in the above studies indicates that the force fields currently in use provide an adequate description of proteins and protein–DNA interactions. Furthermore, simulation techniques can reproduce local structural variation in duplexes, as shown, for example, in studies revealing that the sequence-dependent deformability of dimeric steps of DNA obtained by MD simulations is consistent with that found in crystal structures.[52] Fujii *et al.* also observed that the conformational entropy of the dimeric steps from the MD simulation and the crystal data were remarkably well correlated, with a correlation coefficient of 0.90.[52] Recent investigations of protein–DNA complexes have also demonstrated the usefulness of the simulation approach, as reviewed by Mackerell and Nilsson,[53] and how MD simulation of protein–DNA complexes can reveal the interplay between the structural properties of nucleic acids and the impact of the interactions with the protein partners.

The sequence-dependent deformability of DNA has been studied with a range of experimental[39] and computational methods.[38,52] For example, to estimate the sequence-dependent deformability of DNA base-pair steps, Matsumoto *et al.* performed a normal-mode analysis of model DNA fragments, calibrated against elastic constants of a generic DNA.[54] Lankas *et al.* performed MD simulations of DNA fragments and evaluated the harmonic potential energy functions for all 10 unique steps with respect to the six helical parameters.[55] In both studies, it was found that, in the case of the roll angle, the pyrimidine–purine steps are the most flexible, consistent with the present results. To our knowledge, the methylated EcoRV cognate sequence has not been previously investigated.

Here, we use MD simulation to characterize the structural and dynamic properties of the protein and the DNA in the EcoRV–DNA complex and the relative roles played by direct and indirect readout in sequence-specific recognition by the enzyme. We focus on understanding the interactions between EcoRV and DNA sequences after the recognition of the two outer base pairs GAxxTC. Thus, all-atom MD simulations are performed of three DNA sequences: the cognate sequence AAGATATCTT (TA), the non-cognate sequence AAGAATTCTT (AT) studied in Ref. 35 and the cognate methylated sequence AAGA$_{CH_3}$TATCTT (TA-CH$_3$), which is the same as the cognate sequence, but methylated on the first adenine of the recognition site. Simulations are performed in aqueous solutions of EcoRV unbound and bound to each of the three DNA sequences. We analyze specifically localized chemical or structural molecular perturbations at atomic detail and determine the interactions stabilizing the complexes in solution. The results suggest a three-stage model for the recognition of the cognate sequence, strongly influenced by the intrinsic bending propensities of the DNA free in solution.

The article is organized as follows. First, we describe the free and bound DNA in solution for three different sequences TA, AT and TA-CH$_3$. Then we report on MD simulations of the dynamic behavior of the uncomplexed protein in solution and in crystal environment. Finally, we analyze the interactions between the three different protein–DNA complexes in solution.

Results

DNA free in solution

Experiments have shown that the central base pair of the cognate DNA sequence exhibits a roll angle ρ of 50° when bound to EcoRV, compared to 1.5° for an unbound canonical B-form DNA duplex.[56] To determine how much of this bend, if any, is intrinsic to the structure of the uncomplexed DNA sequences, we performed MD simulations of three sequences (TA, AT and TA-CH$_3$) "free" (i.e., not complexed to the protein) in aqueous solution, starting from a B-DNA form.

The free-energy profile along ρ at the center step for the three unbound DNA sequences, computed from the MD probability distributions (see Materials and Methods), is shown in Fig. 2a. The free-energy minima of the roll angles at the center step differ by

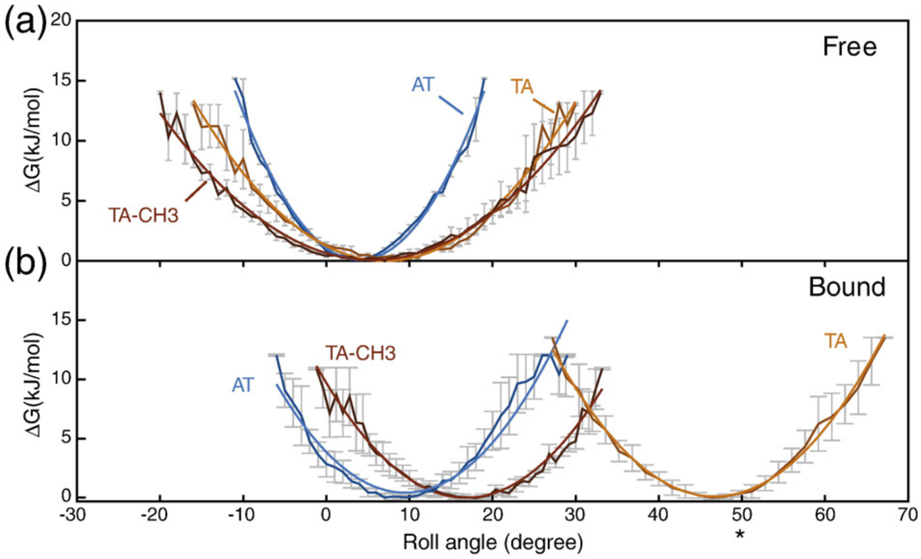

Fig. 2. Free energy profile along the roll angle ρ at the center step of the DNA sequences (a) free in aqueous solution and (b) bound to EcoRV. The TA, AT and TA methylated structures are shown in orange, blue and red, respectively. Corresponding harmonic fits are shown in orange, blue and red, respectively. The star indicates the angle measured from the X-ray structures of EcoRV–TA.

Table 1. Results from a harmonic fit to the free-energy profile along the roll angle ρ at the central step

$0.5a_0(x-a_1)^2+a_2$	TA free	AT free	TA-CH$_3$ free	TA bound	AT bound	TA-CH$_3$ bound
a_0 [kJ (mol °)$^{-1}$]	0.13	0.05	0.04	0.07	0.08	0.07
a_1 (°)	4.01	7.05	5.53	16.82	9.48	46.73
a_2 (kJ/mol)	0.07	−0.01	0.07	0.22	0.92	0.29

The corresponding curves are plotted in Fig. 2.

~4° between TA and AT, whereas the values are almost identical for TA and TA-CH$_3$ (7.9°±0.7 for TA, 4.2°±0.1 for AT and 6.7°±0.1 for TA-CH$_3$). In contrast, the range of angles with a free-energy value within kT of the free-energy minimum (the thermal energy at T=300 K is 2.5 kJ/mol) differs significantly, being wider for the cognate TA sequence, with a range between −3° and 17°, compared to −3° to 10° for AT. The above results show that an exchange of the two central nucleotides of the recognition sequence induces a difference in the flexibility of the uncomplexed DNA molecules, TA and AT. This is in agreement with Refs. 52, 54, and 55.

In contrast to AT, unbound TA-CH$_3$, with the sequence methylated on the adenine 4, behaves in a similar way as TA, that is, with a roll angle ρ at the free-energy minimum of ~7°, and exploring values between −7° and 16° within kT of the minimum. This finding suggests that addition of a methyl group at the first adenine of the recognition site does not change the conformation and flexibility of the uncomplexed DNA molecule significantly compared to the cognate TA sequence. The above statement is valid for all helical parameters. None of the helical parameters shows a significant difference between the TA and the methylated sequence. Similar behavior has been previously observed in case of the EcoRI recognition site; crystallographic work has shown that the methylation of the EcoRI DNA recognition site does not alter its conformation compared to that of the non-methylated sequence.[57]

For the cognate unbound TA sequence, two simulations were performed with different starting conformations: one starting from the B-DNA conformation and the other from the bent conformation found in the EcoRV–DNA crystallographic complex. During simulation of the unbound TA starting in the bent state, the roll angle ρ decreased from its initial value of 50° to 7.5°±0.2°, thus reaching the same minimum as that of the TA starting from the B-form. Consequently, according to the present calculations, the DNA conformation in the complex is an intrinsically unstable and energetically strained form that is stabilized by the EcoRV endonuclease.

To estimate the free energy required to bend the unbound DNA to a roll angle ρ of 50°, as observed in the crystal structure of the EcoRV–DNA complex, we fitted a harmonic potential to the free-energy profiles (Table 1). According to this fit, which, due to anharmonicity likely very approximately represents an upper bound, bending the central step to ρ=50° requires ~45 kJ/mol for unbound TA, whereas for unbound AT the requirement is >130 kJ/mol. The energy required to bend to 50° for TA-CH$_3$ is again similar to that of unbound TA. The above results indicate that the cognate DNA sequences TA and

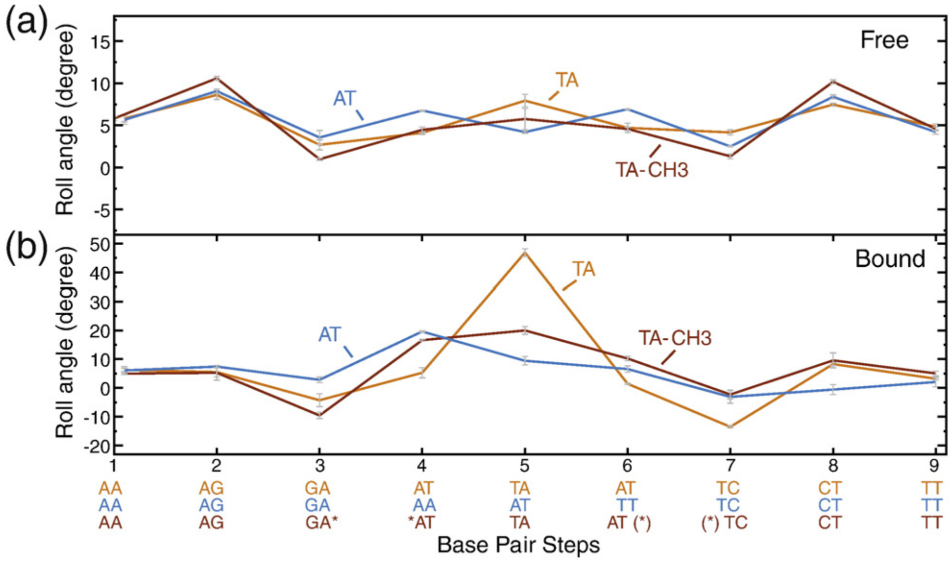

Fig. 3. Average local roll angle ρ values at each base-pair step of (a) the free DNA sequences and (b) the bound DNA. The TA, AT and TA methylated structures are shown in orange, blue and red respectively. The two plots are not on the same scale. The line connecting the points is there to guide the eyes. * indicates where the CH$_3$ group has been added, and (*) indicates that the CH$_3$ group is placed at this base pair on the other strand.

TA-CH_3 have significantly larger intrinsic propensities to induce a kink at the central step than does the non-cognate AT sequence.

DNA in the complex

We examine the effect of binding to the protein on the roll angle, ρ, at the center step. Figure 2b shows the free-energy profiles of the roll angles at the center step of the three DNA sequences complexed to EcoRV. The TA sequence has a free-energy minimum at a ρ of 47°±1.2°, whereas for the AT sequence this minimum is observed at 10°±1.3°. Both values are similar to the angles observed in the crystal structures.[35] The complexed TA-CH_3 sequence has a free-energy minimum at 20°±1.2°. As was shown in the comparison of the roll angles of the free TA and TA-CH_3, this reduction of ρ by ~30° is not an intrinsic feature of the methylation but rather arises from changes in interaction with the protein, which is further analyzed below.

The change of sequence affects not only the center step, but also neighboring steps, as can be seen in Fig. 3 in which the roll angles for nine base pairs are reported. In the unbound DNA there is no significant effect of DNA sequence on ρ at the noncentral base-pair steps (Fig. 3a). However, the ρ profile of the TA complex differs significantly from those of the other two complexes at the central step (Fig. 3b), being 47° for TA and 10° and 20° in the AT and TA-CH_3 sequences, respectively. Furthermore, comparison of ρ of the bound TA and TA-CH_3 molecules at each base-pair step shows that the changes are more pronounced at the center step 5 (TA) and on the steps containing the methylated adenine: steps 3 (GA), 4 (AT), 6 (AT) and 7 (TC). This suggests that addition of the methyl group perturbs the interactions with the protein, reducing the angle at the center step and perturbing also the rest of the base-pair steps of the recognition sequence.

Conformational changes in the protein upon complex formation

To investigate structural changes in the protein on complexation, we performed MD simulations of the complexed and the unbound protein for 50 ns, both starting from the protein conformation in the bound form. The main findings of these simulations are a reduction of the protein flexibility upon complex formation and an opening of the "arms," that is, the two C-terminal domains, which were observed only in the simulation of the unbound protein. These findings are described in more detail in the following.

To examine the structural flexibility of the protein, we calculated the RMSFs (Fig. 4). The region comprising residues 162–244 exhibits markedly higher fluctuations in the unbound protein. This region can be decomposed into two parts: residues 171–186, defined as the recognition (R) loop (in green in Fig. 1), and residues 187–244, which constitute the C-terminal domain (in yellow in Fig. 1) of the protein. A change in flexibility upon binding to DNA is also seen in other parts of the protein; for example, the Q loop and subdomains close to the DNA are more flexible in the absence of DNA and become more ordered upon DNA binding. Hence, the presence of the DNA reduces the global flexibility of the protein in general and, in particular, the R loop and the C-terminal domain.

To probe the effect of binding to the cognate or to the non-cognate sequences on the interdomain flexibility, we calculated the distance between the C-terminal subdomains of each subunit. Figure 5 shows the free-energy profile along the distance

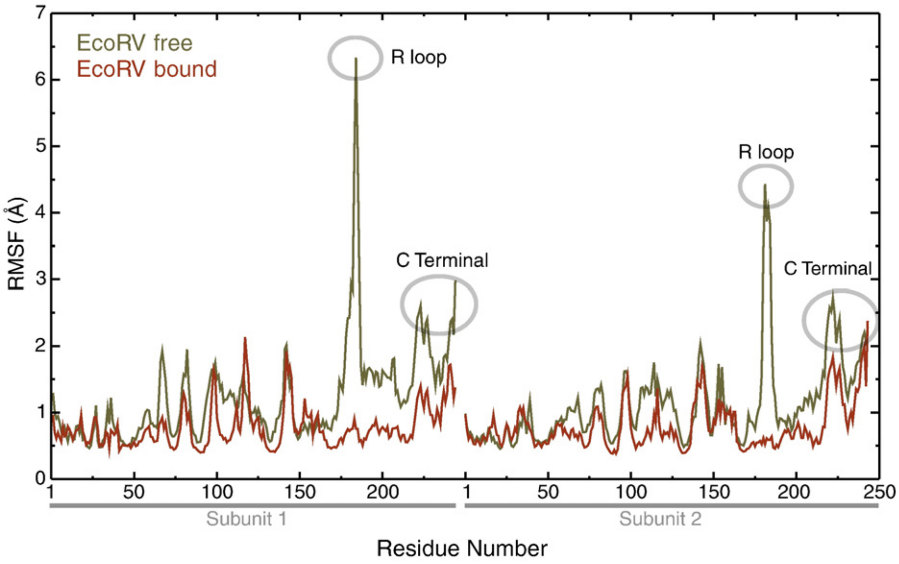

Fig. 4. RMSF for both EcoRV subunits as a function of residue number, calculated for the C^{α} atoms. The black line shows the RMSF of EcoRV from the simulation of the unbound state, and the red line the RMSF from the simulation in the cognate complex structure.

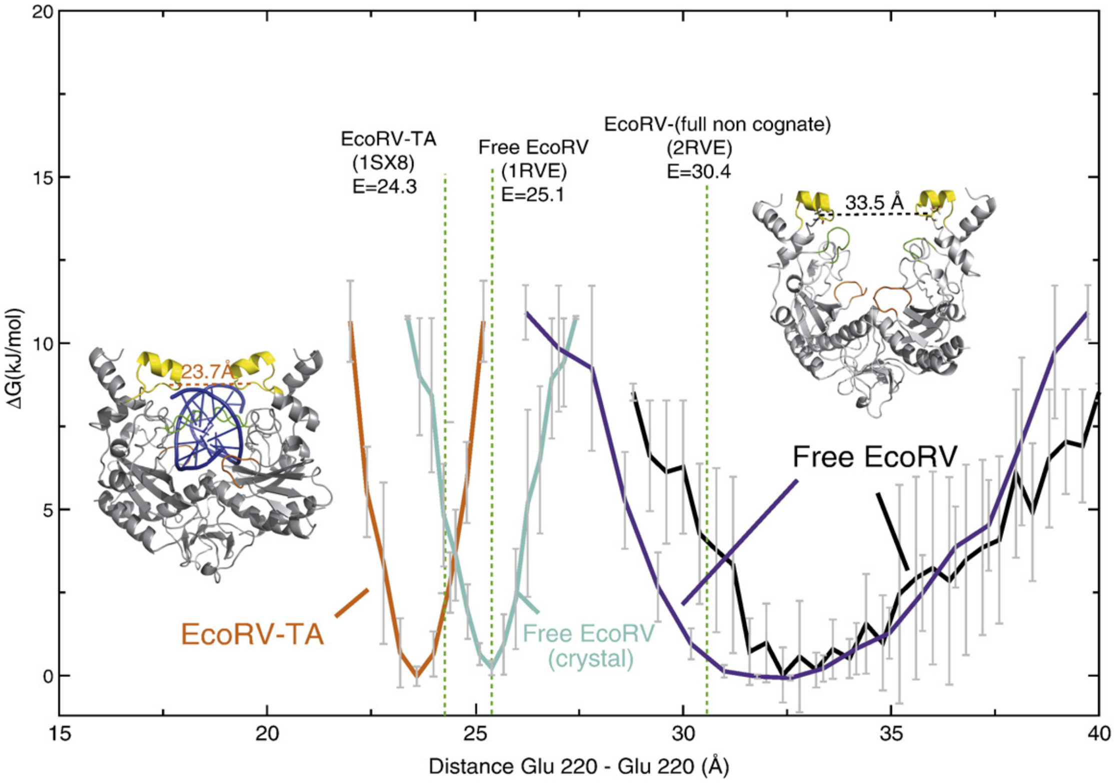

Fig. 5. Free-energy profile along the distance between residue Glu220 of the C-terminal subdomain in the two subunits of EcoRV. Vertical green lines represent the distance *E* between the two C-terminal subdomains from the crystal structures. The black and violet lines correspond to the distance of the unbound EcoRV simulated in solution, the black line shows the C-terminal distance during the simulation without restraints and the violet line shows the distance for the simulation with initial restraint s after removing them (see Materials and Methods for more details). The turquoise line is the distance of the unbound EcoRV simulated in crystal, and the orange line shows the distance of the EcoRV complexed to the cognate TA sequence.

between the C$^{\alpha}$ atom of Glu220 of the two monomers. The unbound protein in solution has a minimum free energy at a distance of 33.4 Å±1.1 between the two C-terminal domains, longer than observed in the crystal of the free protein (25.1 Å in 1RVE). To make sure that this finding is not an artifact, we performed two control simulations. The first control evaluated whether the simulation can reproduce the crystal structure, when the protein is simulated in the crystalline state. Indeed, the unbound protein in the crystal simulation remains in the closed state, with the two C-terminal subdomains close to each other at about the same distance as observed in the crystal structure. The second control simulation was performed in aqueous solution. During this simulation, restraints were imposed on the C-terminal domains for 8 ns to keep them close to the same distance as in the crystal, and then the restraints were removed for the rest of the simulation time. The aim of the constraints was to let the protein relax in the crystal-state conformation (with closed C-terminal domains) and accommodate to the solution environment. The behavior of the unbound protein in solution is different in the crystalline state. After the 8 ns during which the distance was constrained to ~25 Å, the C-terminal subdomains move further away from each other, and at equilibrium a longer interdomain distance is reached (~31.7 Å), comparable to that observed in the simulation without the restraints (see Fig. 5).

Table 2. Opening width of EcoRV measured as the radii of gyration and the distance between the two C-terminal domains (GLU220)

		Radii of gyration (Å)	Distance Cterm–Cterm (Å)
Free protein (simulation)	In solution	25.4±0.2	31.7±0.3
	In crystal	25.0±0.1	25.4±0.1
Protein–DNA (solution simulation)	EcoRV–TAseq	24.6±0.1	23.6±0.6
	EcoRV–ATseq	24.9±0.1	25.4±0.1
	EcoRV-mTAseq	24.8±0.1	26.1±0.8
Crystal structures	EcoRV free (1RVE)	24.4	25.1
	EcoRV–TAseq (1SX8)	24.2	24.3
	EcoRV-fully non-cognate (2RVE)	25.1	30.4

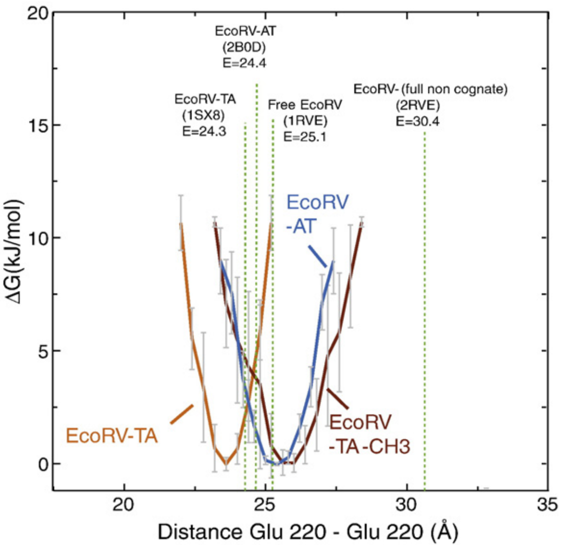

Fig. 6. Free-energy profile along the distance between residue Glu220 of the C-terminal subdomain in the two subunits of EcoRV. Vertical green lines represent the distance E between the two C-terminal subdomains from the crystal structures. The orange, blue and red lines show the distance of the EcoRV complexed to the cognate TA sequence, non-cognate AT sequence and the TA methylated sequence, respectively.

These data show two important results. First, the protein behaves differently in the crystal than in solution: the maintenance of the closed conformation observed in the crystal is most probably due to the crystal packing. Second, these data confirm the existence of an "open state" of unbound EcoRV in solution. This state has not been observed experimentally, but was, according to Schulze *et al.*,[28] "never ruled out."

In Ref. 28, it was noted that the radii of gyration of EcoRV free and in complex with specific DNA, as determined by neutron scattering, are identical within the limits of error. We have now computed the radius of gyration for the simulations of the protein free in the crystal, in solution and the protein complexed to DNA. The results are reported in Table 2. All three radii of gyration are very similar, differing by at most 1 Å, in agreement with the experimental results. In Table 2, we have added the values of the distance between the two C-terminal domains. We can see, for example, in the cases of the unbound protein in solution and in the crystal, that large differences in the distance between the two C-terminal domains occur, even though the values of the radii of gyration are very similar.

The simulations indicate that the opening of the C-terminal domains is primarily triggered by the high level of flexibility of the R loops. When the protein is complexed to DNA, the R loops make very strong interactions with the DNA, stabilizing the whole system and maintaining the protein in the closed state. Upon removal of the DNA, the R loops initially interact with the Q loops of the other monomer, but these hydrogen-bond interactions break rapidly and the R loops interact with the C-terminal domains of the same monomer. At that point, no interactions remain that keep the C-terminal domains close together, and this leads to the opening.

Figure 6 reveals the distance between the two subdomains for the bound protein. Differences are seen between the cognate complex (EcoRV–TA, orange profile) and the two non-cognate complexes (EcoRV–AT and EcoRV–TA-CH_3, blue and red profile, respectively): the distance is longer when EcoRV is bound to a non-cognate sequence. The average distance measured during the simulation of the cognate complex is 23.6±0.6 Å, showing excellent agreement with the distance of 24.3 Å observed in the crystal structure (1SX8). EcoRV complexed to a full non-cognate sequence was not simulated, but in the crystal structure (2RVE) the distance is 30.4 Å, which

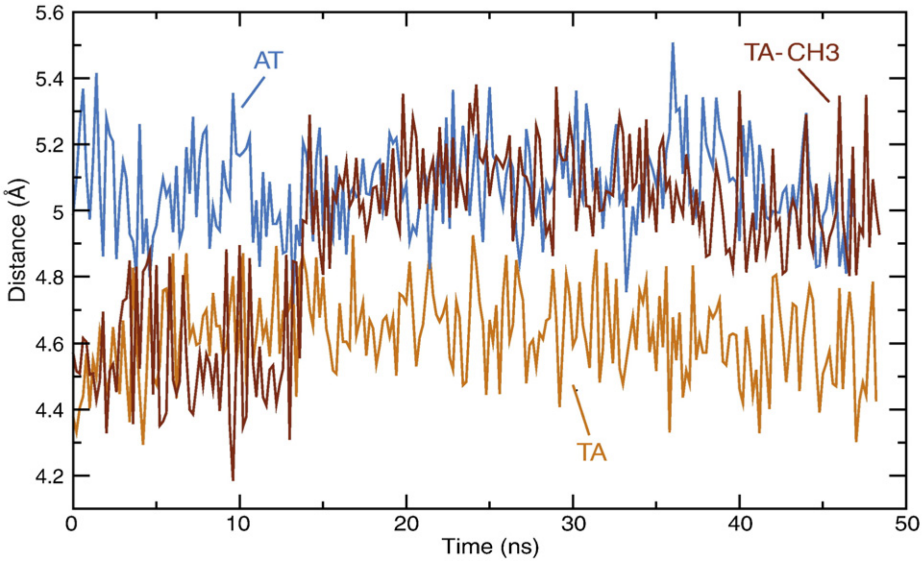

Fig. 7. Evolution of the average distance *D* between the phosphorous atom of T7 and the C^{α} atom of Lys92 for both subunits. The TA, AT and TA methylated structures are shown in orange, blue and red, respectively.

is larger than the distance observed during the simulation of the two non-cognate complexes EcoRV–AT and EcoRV–TA-CH_3.

EcoRV–DNA interaction in the complexes

The complex EcoRV–TA is unique in having a highly distorted DNA with a sharp central kink of ~50°, which renders the inner two base pairs inaccessible to the protein. Therefore, recognition of these two base pairs cannot be direct. However, a number of amino acid residues contact the phosphate groups of the DNA both in and outside of the recognition sequence, which may compensate for the lack of direct base contacts.

Figure 7 shows the time series of the distance between the phosphate atom of the T_7 base of each DNA sequence and the C^{α} atom of Lys92, a residue that has been found to be essential for DNA cleavage.[58,59] After equilibration, this distance is longer for the two non-cognate sequences AT and TA-CH_3. Initially, the TA-CH_3 sequence is close to the protein. The starting point for the EcoRV–TA-CH_3 complex, having been taken as the same conformation as the EcoRV–TA complex, then moves further away during the MD simulation. The protein–DNA distance continues to increase up to approximately 15 ns simulation time.

In the following we analyze the differences in how the three DNA sequences interact with the protein. In the light of the results obtained on the behavior of the DNA when free and bound, an understanding will be reached of how the intrinsic flexibility of the DNA at the center step influences the interactions between EcoRV and the DNA. Figure 8 shows the protein–DNA interface taken from a snapshot of the EcoRV–AT, together with all the residues of the protein that form very stable hydrogen bonds with the cognate DNA (with occupancies >80%). Most of these residues are located within, or close to, the Q

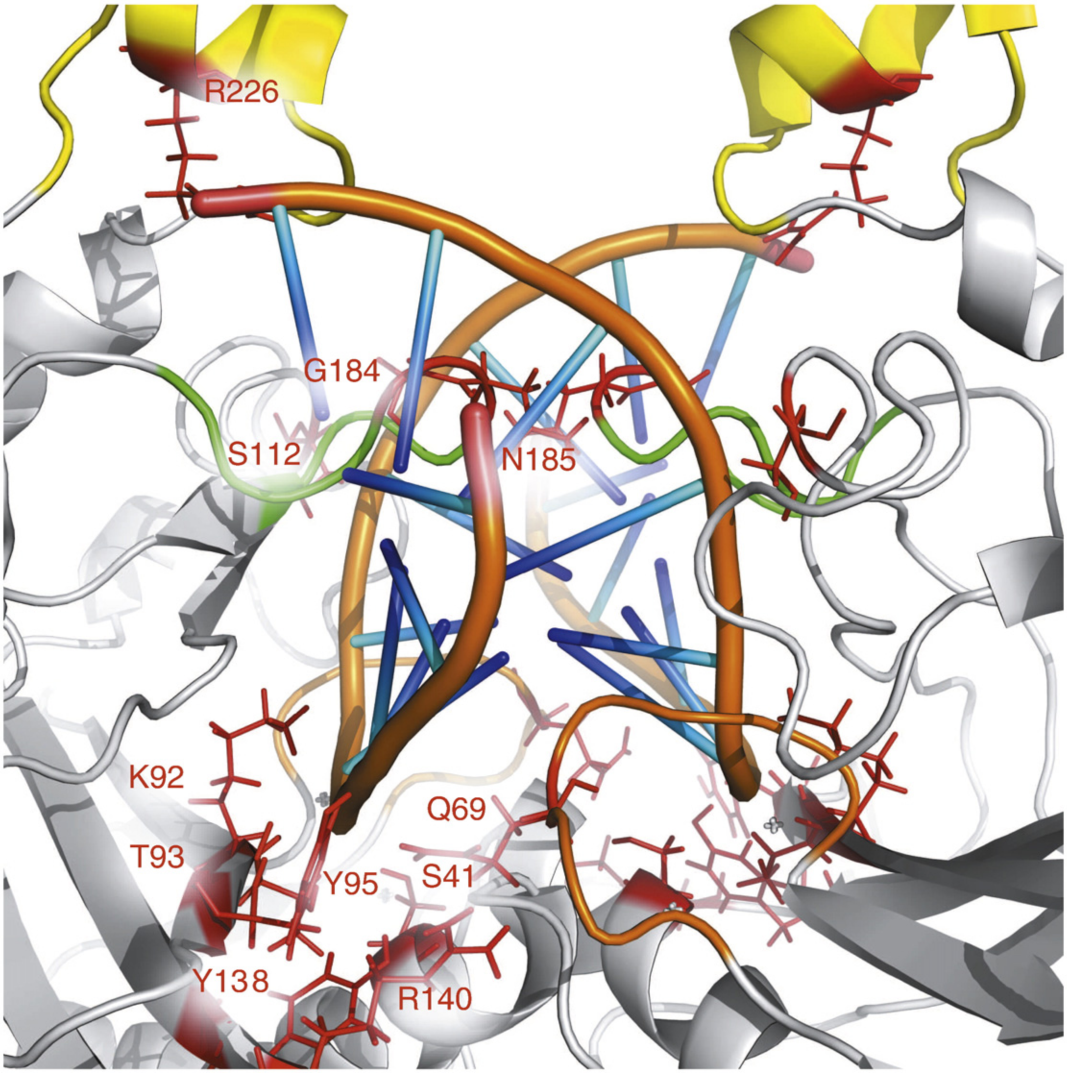

Fig. 8. Expanded view of the regions of EcoRV containing the DNA from the MD simulation of the EcoRV–TA complex. The C-terminal subdomain, R loop and Q loop are shown in yellow, green and orange, respectively. Residues of the protein forming stable hydrogen-bond interactions with the DNA are shown in red. Stable hydrogen bonds are formed with residues of the R and Q loops and of the C-terminal domains.

loop. Three residues are within or close to the R loop, and one residue is in the C-terminal subdomain. To simplify the comparison of the hydrogen-bond patterns, a schematic view of the most stable hydrogen-bond interactions formed between the two subunits of the protein and the two strands of the DNA is shown in Fig. 9. Table A of the Supplementary Material lists these most stable hydrogen-bond interactions. A distinction is made between hydrogen bonds with specific atoms of the base and "non-specific" hydrogen bonds with atoms of the backbone.

Comparison of hydrogen bonding in EcoRV–TA and EcoRV–AT

Three stable, specific hydrogen bonds are formed between the G_3 and A_4 bases on each strand of the two DNA sequences TA and AT and residues Gly184 and Asn185 of the R loop of the protein. Many other residues exhibit similar occupancies between the two complexes, that is, Gln69, Thr93, Asn185 and Arg140 of subunit A and Arg226, Gly184, Asn185 and Ser112 of subunit B. However, differences also exist. In most cases, the hydrogen bonds in the complex EcoRV–TA have a higher occupancy than in EcoRV–AT; that is, Ser112, Lys92, Gly184 and Tyr95 of subunit A and Arg140, Tyr138, Thr93 and Gln69 of subunit B. Ser112 forms hydrogen bonds with the backbone atoms of both DNA sequences but not with the same atom, as is also the case for Ser41 (Fig. 9). These results show that the presence of a lower roll angle at the central step for AT is accompanied by a weakening of the hydrogen-bonding interactions between the protein and the DNA backbone.

Further significant differences are seen in the interaction of EcoRV with the cognate and noncognate DNA sequences. In case of the cognate DNA, the two subunits interact in a symmetrical way; that is, for most of the residues of subunit A that form stable hydrogen bonds the corresponding residues in subunit B also exhibit strong hydrogen bonds. In contrast, in case of the non-cognate DNA AT, the two subunits are asymmetric. Whereas Tyr138, for example, forms hydrogen bonds with an occupancy of 99% with subunit A, Tyr138 of subunit B does not form a hydrogen bond with the DNA.

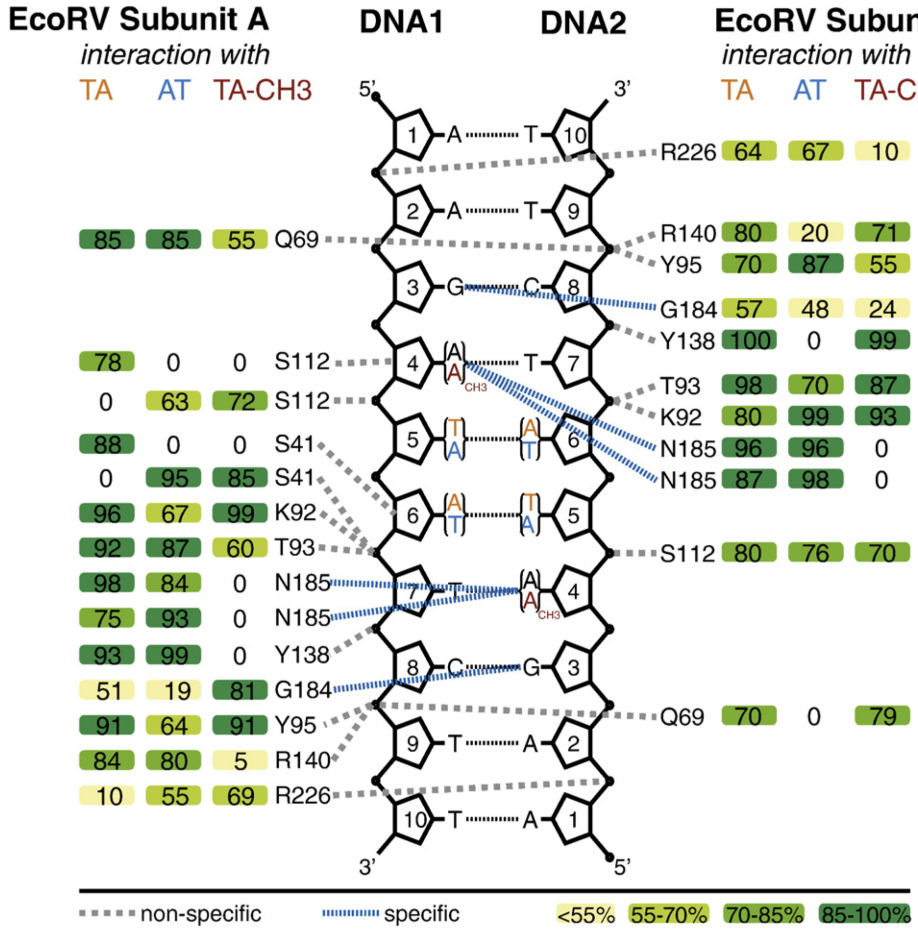

Fig. 9. Schematic view of the hydrogen-bond interactions of the DNA with each subunit of the protein for the TA, AT and TA-CH_3 complexes. Specific interactions are represented as blue dotted lines and interactions with the backbone are shown with dashed gray lines. The occupancies of the hydrogen bonds are highlighted by a color gradient from pale yellow for low occupancy to dark green for high occupancy.

Experimental work has shown that mutations of Thr93 and Ser112 selectively affect the overall enzymatic catalytic rate of the reaction.[60] In the present simulation, hydrogen bonds of Thr93 in both subunits with both DNA sequences are very stable, consistent with the catalytic importance of this residue. In contrast, hydrogen bonds of Ser112 of subunit A with the two DNA sequences are different; the TA sequence forms a hydrogen bond with the sugar of the A_4 base, whereas for AT, the phosphate of the T_5 base is hydrogen-bonded to Ser112. The hydrogen bonds with the other subunit behave similarly (see Table A of the Supplementary Material).

The above hydrogen-bond analysis shows that the AT sequence forms roughly the same hydrogen bonds with EcoRV as does the TA sequence, while the majority of these hydrogen bonds exhibit lower occupancies, that is, are weaker. The interactions between the two subunits of the protein and the DNA are asymmetric with the non-cognate sequence. This asymmetry in non-cognate protein–DNA complex has already been discussed in Ref. 61, in which the structural effects of symmetric/asymmetric and cognate/non-cognate binding on specificity are examined.

Comparison of hydrogen bonding in EcoRV–TA and EcoRV–TA-CH_3

In the EcoRV–TA-CH_3 complex, the amino group of the first adenine of the recognition sequence A_4 is methylated (AAGATATCTT). Figure 9 shows that the stable hydrogen bonds formed by A_4 in the TA and AT sequences are inexistent for the methylated species TA-CH_3. In particular, interactions with Asn185 are lost in the presence of the methyl group. Indeed, from the three specific interactions observed in the two other complexes (Fig. 9), only one, between Gly184 and the G_3 base, remains in the methylated complex. Except for the above difference, the comparison of TA and TA-CH_3 complexes shows that the two DNA sequences form very stable interactions with many of the residues of the protein, with similar hydrogen-bond occupancies for residues Lys92 and Tyr95 of subunit A, and Arg140, Tyr138, Thr93, Lys92, Ser112 and Gln69 of subunit B. Interactions with residues Ser112 and Ser41 of subunit A are, as for the EcoRV–AT complex, with different atoms of the backbone than those formed in the EcoRV–TA complex.

The occupancies of the protein–DNA hydrogen-bond interactions in EcoRV–TA-CH_3 are lower than in the cognate complex and an asymmetry is observed, similar to the one seen in EcoRV–AT. Again, this involves Tyr138, which in subunit A does not hydrogen-bond with the DNA, whereas Tyr138 of subunit B forms a hydrogen bond with an occupancy of 99%. Asymmetry is also observed for Arg140 and Arg226, which form weak hydrogen bonds in subunit A and stronger ones in subunit B (Fig. 9).

The comparison of EcoRV–TA with EcoRV–TA-CH_3 reveals two important differences: the TA-CH_3 sequence does not reach the full bend of 50° in the presence of EcoRV (as observed in the cognate complex), and two stable hydrogen-bond interactions between Asn185 and A_4 are lost when A_4 is methylated. In contrast to the AT sequence, the roll angle ρ of 20° observed at the central step is not due to the lack of an intrinsic ability of the DNA sequence to roll (see the section DNA free in solution), but rather arises from the loss of these two stable hydrogen bonds, ultimately leading to the loss of the binding energy required to stabilize the larger roll angle.

To accommodate the methylated adenine, we performed energy minimization and equilibration

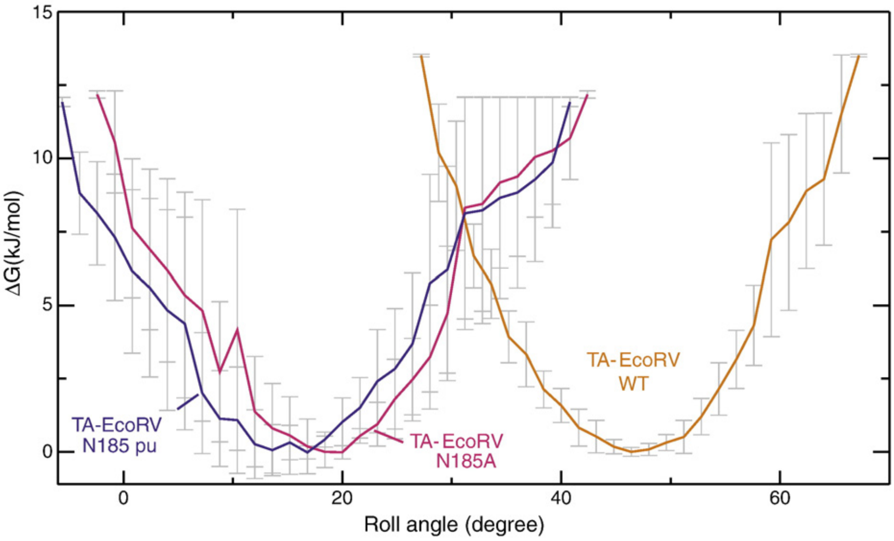

Fig. 10. Free-energy profile along the roll angle ρ at the center step of the DNA bound to wild-type EcoRV and complexed to the N185A and N185pu mutants. The free energy is calculated from the 50-ns MD simulations. TA, N185A and N185pu structures are shown in orange, pink and indigo respectively.

for 2 ns in which the system was restrained, except for a region of 5 Å around the methyl group. Any local steric conflicts in the starting structure were eliminated by the energy minimization and constrained equilibration, during which harmonic restraints were gradually lifted from the solute atoms. After that, we performed a production run without restraints and observed that the roll angle at the central step then decreases to an equilibrium value between 10° and 20°, reached after ~4 ns (see Fig. B of the Supplementary Material). The converged value (~20°) of the roll angle does not depend on the starting structure of the simulation.

To examine more closely the role of the Asn185 in the stability of the roll angle at the center step, we performed two mutations. In the first, the partial charges of the atoms of the NH_2 group of Asn185 were set to zero (N185pu) to prevent the formation of hydrogen bonds. In the second, Asn185 was mutated to alanine. Figure 10 shows the free-energy profiles along the roll angle ρ at the central step of the cognate complex, the N185A complex and N185pu. For both N185A and N185pu the minimum of the free energy is shifted from ~47° for the wild type to ~17° to 20° for the two mutated complexes.

The N185A complex has two fewer interactions between N185A and the TA DNA sequence than the wild type (Fig. 11), while only one hydrogen bond is lost in N185pu. A DNA cleavage rate reduced to 1/5000 of the activity of the wild-type enzyme has been measured for N185A.[62,63] This reduction in rate can be explained by the loss of the specific hydrogen-bond interaction with the NH_2 group of the Asn185 residue, which perturbs the protein–DNA interactions leading to the absence of a full bend. The above simulations of the mutants indicate that direct interactions between the first adenine of the recognition sequence and residue 185 of EcoRV are crucial for maintaining the roll angle at the central step at ρ = 50°.

Discussion

EcoRV has the ability to recognize a specific DNA sequence, GATATC (TA), within a large molar excess of non-specific DNA. The present MD simulations of DNA free and bound to the enzyme allow the understanding of the properties of the cognate TA sequence compared to those of a non-cognate sequence, GAATTC (AT). The simulations

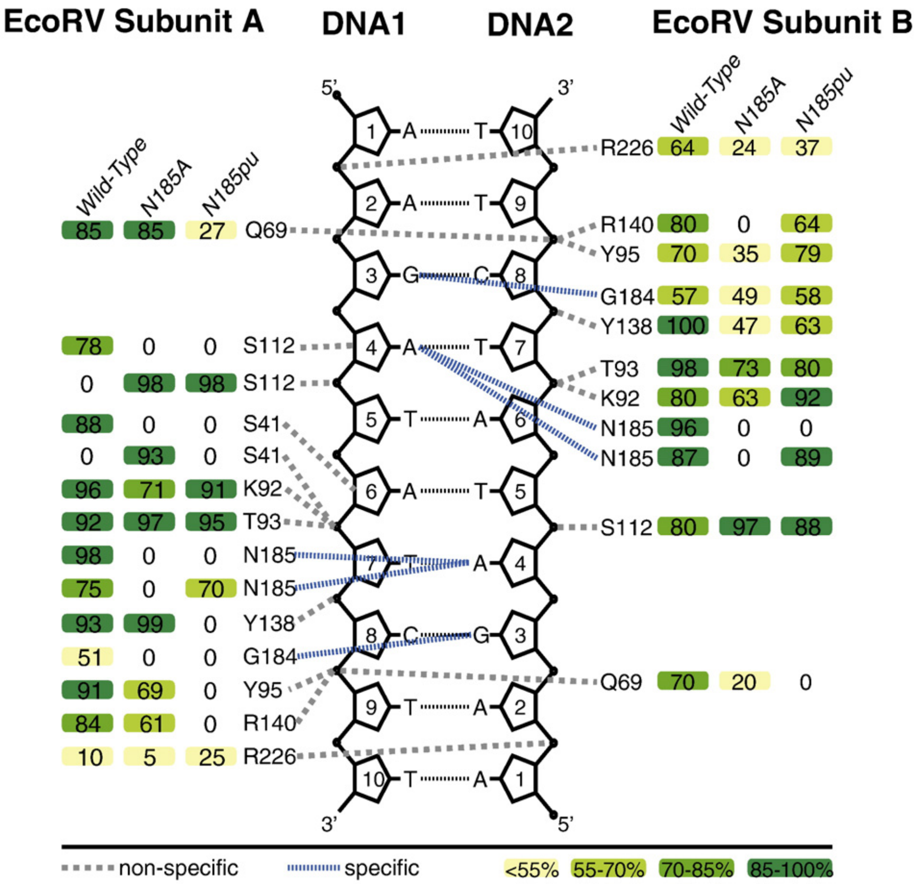

Fig. 11. Schematic view of the hydrogen-bond interactions of the DNA with each subunit of the protein for the wild-type, N185A and N185pu complexes. Specific interactions are represented as blue dotted lines and interactions with the backbone are shown with dashed gray lines. The occupancy of the hydrogen bonds is highlighted by a color gradient from pale yellow for low occupancy to dark green for high occupancy.

suggest that the free TA sequence is more flexible than the free AT sequence, and EcoRV would need significantly more energy to kink AT than it does for TA. A similar argument has been proposed in the specific recognition of damaged DNA by repair enzymes.[64] It was shown that higher flexibility of the damaged DNA leads to the DNA being more susceptible to distortions induced by the enzyme, thus lowering the barrier to base flipping. This intrinsic propensity has a clear effect on the interaction with the protein: the present calculations suggest that the protein and the DNA interact less closely in the case of the non-cognate sequence AT than for the cognate TA, and that the C-terminal protein domains are further apart in the presence of the AT sequence than with the cognate one. This longer distance leads to weaker and more asymmetric interactions between the non-cognate DNA and the two subunits of the protein, preventing cleavage.

It has been shown experimentally that EcoRV does not cleave the cognate sequence when it is methylated on the first adenine of the recognition sequence, that is, $GA_{CH_3}TATC$.[27] However, no crystal structure of the complex of EcoRV with the methylated sequence ($GA_{CH_3}TATC$) has been reported. The MD results suggest that, although the intrinsic propensity of the unbound TA-CH_3 to bend is unchanged relative to that of the cognate TA sequence, in the EcoRV–DNA complex the addition of the CH_3 results in a loss of specific hydrogen-bond interactions with Asn185, leading to a reduction of the roll angle ρ at the center step. This less bent DNA does not bind as tightly to the protein as the fully bent one and, as a result, many protein–DNA hydrogen-bond interactions are perturbed, as manifested by lower hydrogen-bond occupancies. The TA-CH_3 and the cognate TA sequences require approximately the same energy to bend, but when the sequence is methylated, this energy cannot be supplied by complex formation due to the loss of the strong interactions.

Mutation of the adenine hydrogen-bonding partner Asn185 in TA has a similar effect. The N185A mutant loses both specific wild-type hydrogen-bond interactions with the DNA, whereas N185pu removes one of these two specific interactions. As a result, in both these mutants, ρ at the central step does not exceed 20°, impeding the final step of the recognition mechanism, that is, the formation of a tight, compact, cleavage-ready complex. Asn185 plays an important role in the stability of the bend. However, the results for the interaction between the AT sequence and EcoRV suggest that the presence of the two stable bonds involving Asn185 is not sufficient and that an intrinsic propensity to bend is also required.

On the basis of the crystal structures of different forms of EcoRV free and bound to cognate and non-cognate DNA, the study of the conformational flexibility of the C-terminal subdomains of EcoRV, and the absence of experimental evidence for an open state of the unbound protein, Schulze *et al.* suggested a mechanism of DNA association to EcoRV based on two steps.[28] In this model, the free EcoRV is in a closed conformation. The DNA first binds to the outer parts of the C-terminal subdomains of EcoRV at the surface. Then, the two subdomains open up and the DNA can enter the DNA binding cleft. In Ref. 28, the authors explicitly state that "the existence of an open conformation cannot be ruled out," although there is no experimental evidence for an open state of free EcoRV in solution. The present simulations, however, suggest that in solution the unbound protein is in an open state with flexible C-terminal domains, the average distance between the two domains being larger than in the X-ray structures of the protein either free or complexed with non-cognate sequence. Moreover, the results show that the R loops play a crucial role, being very flexible when the protein is unbound, but more ordered, and forming very stable hydrogen-bond interactions when complexed to both the cognate (TA) and non-cognate (AT) sequences.

In Ref. 35 it was shown that the association rates for binding of AT and TA sequences to EcoRV are nearly identical, suggesting that the initial stages of the induced-fit transition are similar at both specific and non-cognate sites. However, the authors observed that the AT complex is unstable, allowing the protein to leave the DNA. The present results confirm their suggestions; recognition of nearly identical sequences with cognate outer base pairs leads to the formation of a partially bound, closed complex with bent DNA. Moreover, we observe that the non-cognate AT complex is less tight and less stable than the cognate TA complex.

Proposed mechanism

Although force field inaccuracies and sampling limitations lead to errors in MD simulation, the accuracy of this computational technique is sufficient to enable us to propose the following sequence-recognition mechanism. Combination of the information obtained from the present analysis of different EcoRV–DNA complexes and from experimental studies suggests a recognition mechanism consisting of three steps (Fig. 12). In the first step, the protein, in an open conformation, interacts loosely with the DNA and stochastically moves along it. The X-ray structure of EcoRV bound to a full, non-cognate sequence shows that the C-terminal subdomains are in an open conformation [Protein Data Bank (PDB) identifier 2RVE[26]]. This is labeled as the "loosely bound, open-state" in Fig. 12. In the second step, the two outer base pairs, GAxxTC, are recognized and the R loops become more ordered, forming strong hydrogen-bonding interactions. In particular, specific hydrogen bonds between the base of the first adenine and Asn185 play a crucial role. Some residues form very strong hydrogen bonds in the tight cognate complex but looser hydrogen bonds in the non-cognate complexes, notably Gly184, Ser112, Lys92, Thr92 and Gln69. An asymmetry is observed in the case of non-cognate complexes: residues Tyr138, Arg140 and Arg226 form strong hydrogen bonds with only one subunit, while they behave symmetrically in the cognate complex. These interactions are consistent with the simulation of the EcoRV–AT complex, which shows that the C-terminal subdomains come closer together than in the free protein or when bound to a fully non-cognate sequence. Consequently, a "partially bound, closed state" EcoRV–DNA complex is formed. The presence of the partially bound, closed state can be inferred from the AT–EcoRV complex simulations that show that (i) the DNA has a favored roll angle of ~10°; (ii) the protein complexed to the AT-sequence is in a closed state, with the Glu220–Glu220 distance being shorter than when the protein is complexed to a full non-cognate sequence (as seen in the X-ray structure 2RVE), but longer than in the "tightly bound, closed state" (Fig. 6); and (iii) the protein is more loosely bound to the DNA, as evidenced by the larger T7–Lys92 distance *D* (Fig. 7). Therefore, the simulations indicate the presence of a partially bound, closed state in the GAxxTC–EcoRV complex, where the structures of both DNA and protein and the DNA–protein distance are calculated to be different from that of the tightly bound, closed state of the TA–EcoRV complex or the full non-cognate complex. The partially bound, closed state was not accessible to the TA–EcoRV complex simulations probably because the starting configuration had the DNA in the thermodynamically favored 50° roll angle configuration. In the case of the cognate TA sequence, this state can be regarded as transitory rather than a stable intermediate. Due to the intrinsically much lower probability of non-cognate sequences to bend, the energy gained from complex formation is not sufficient to increase the roll angle. An example is the non-cognate AT sequence studied here for which the angle is ~10° when complexed to EcoRV. Consequently, the protein–DNA hydrogen-bond interactions remain weaker and the DNA does not bind deeply enough in the protein for cleavage. The stabilities of the loose non-cognate complexes are lower than the stability of the compact cognate complex, thus permitting rapid unbinding and further search for the recognition sequence along the DNA.

Materials and Methods

Systems setup and solvation

Standard B-DNA starting structures were generated with the program NAB.[65] Three 14-bp B-DNA molecules were examined with different central nucleotides but the same flanking sequences and same nucleotide content (the TA sequence, 5′-dAGAAGATATCTTGA-3′; the AT sequence, 5′-dAGAAGAATTCTTGA-3′; and the TA-CH_3 sequence, 5′-dAGAAGA$_{CH_3}$TATCTTGA-3′). Simulations of the free TA sequence in aqueous solution were also started from a bent form by taking the coordinates of the DNA from the 1SX8 PDB structure.

Three corresponding EcoRV–DNA complexes were prepared. The X-ray crystallographic coordinates of the EcoRV–DNA complex (PDB identifier 1SX8) solved at 2.5 Å resolution[58] were used as the starting model structure with an all-atom representation. Ala92 was back-mutated to Lys as described in Ref. 66. This structure consists of the dimeric protein associated with a dodecamer DNA duplex of base sequence 5′-dAAGATATCTT-3′ containing the recognition sequence (GATATC) as the central part of the DNA. Three MD simulations of the unbound protein were performed in three different conditions: in crystal, in solution, and in solution with restraints on the first 8 ns. The unbound protein simulations were done starting from the structure of the protein in the complexed form, using the 1SX8 PDB structure. The structure of EcoRV complexed with the non-cognate AT DNA has been solved (PDB identifier 2B0D), but the protein was not entirely resolved, with some regions missing. To correct this, we modeled a hybrid complex with the EcoRV protein from the X-ray structures from the PDB identifier 1SX8[58] and the DNA duplex structure of the base sequence 5′-dAAGAATTCTT-3′ containing the recognition sequence (GAATTC) from the PDB identifier 2B0D.[35] The root-mean-square deviation (RMSD) calculated between the C^α atoms of the protein of the two X-ray structures after superimposition is 0.61 Å. The third complex modeled involves the methylated TA sequence, in which the methyl group was added on the fourth adenine of the 5′-dAAGATATCTT-3′ DNA duplex of the protein–DNA complex from the PDB

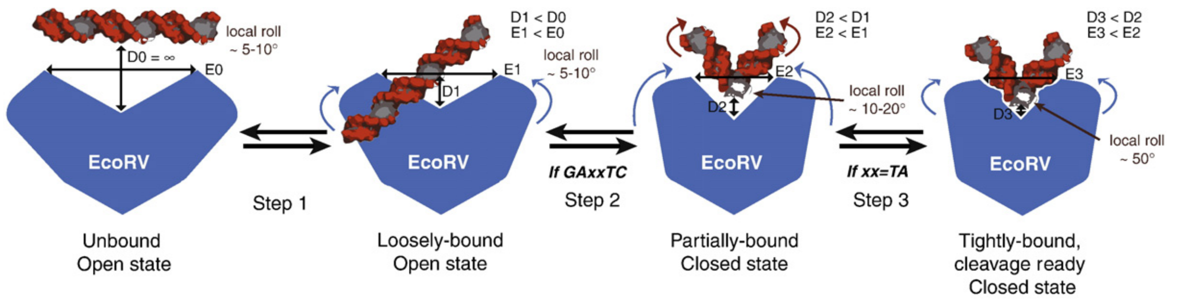

Fig. 12. Schematic representation of the proposed DNA-sequence recognition mechanism by EcoRV in which only the most important differences between the states are shown. First, EcoRV in an open state binds to a random DNA sequence and forms a loosely bound open complex (step 1). Then, upon recognition of the outer base pairs, GAxxTC, a partially bound closed complex with bent DNA is formed (step 2). Finally, recognition of the center base pairs, xx = TA, results in a tightly bound cleavage-ready complex, with a full kink of 50° (step 3). The distance between the "arms" of the protein, *E*, decreases from step 1 to step3 ($E0 < E1 < E2 < E\ 3$). Simultaneously, the distance between the protein and the DNA, *D*, becomes smaller ($D0 < D1 < D2 < D3$).

identifier 1SX8.[58] A control simulation in which the system was restrained for the first 2 ns, except for a region of 5 Å around the methyl, was done. Two simulations of mutated EcoRV complexed to the cognate DNA sequence were also performed. In the first mutation, the partial charges of the atoms of the NH_2 group of Asn185 were set to zero (N185pu) to prevent the formation of hydrogen bonds, and in the second, Asn185 was mutated to alanine.

The simulations were run with the program NAMD[67] using the CHARMM27 force field.[68] The free DNA systems were solvated with the explicit TIP3P water model[69] extending to at least 10 Å beyond the DNA in each direction in a cubic box ($x=y=z=70$ Å). Twenty-eight Na^+ counterions were added to neutralize the system and a further excess of Na^+ and Cl^- ions to obtain a physiological concentration of 150 mM NaCl. The addition of the ions was carried out by random substitution of water oxygen atoms.

Systems of complexed EcoRV–DNA were also solvated with the explicit TIP3P water model[69] extending to at least 10 Å beyond the protein–DNA complex in each direction in a cubic box ($x=y=z=110$ Å). Two Cl^- counterions were added to neutralize the system, and an excess of Na^+ and Cl^- ions as described above. The crystallographic water molecules were included and the crystallized divalent Mn ions were changed to Mg^{2+} as in the native enzyme.

A summary of the simulations is given in Table 3.

MD simulation protocol

Simulations were performed using periodic boundary conditions, and the long-range electrostatic interactions were treated using the particle mesh Ewald method[70] on a 72×72×72 charge grid for the free DNA systems, and on a 112×112×112 charge grid for the protein–DNA complex systems, with a nonbonded cutoff of 12 Å. The short-range electrostatics and van der Waals interactions were truncated at 12 Å using a switch function starting at 10 Å.

The solvated structures were minimized using 5000 steps of steepest descent, followed by minimization with the conjugate gradient algorithm, with solute atoms harmonically constrained until an energy gradient of 0.01 kcal/(mol Å) was reached. The system was then gradually heated for 30 ps to 300 K with 1 K temperature steps with harmonic restraints on the solute atoms.

Table 3. Overview of the MD simulations

		Simulation time (ns)
Free protein (1sx8 without DNA)	In solution	50
	In crystal	50
	Solution, with initial restraints	50
Free DNA	AAGATATCTT	50
	AAGAATTCTT	50
	AAGA$_{CH_3}$TATCTT	50
Protein–DNA	EcoRV–TAseq (1sx8)	50
	EcoRV–ATseq (2bod–1sx8)	50
	EcoRV-mTAseq (1sx8)	50
	EcoRV-mTAseq (1sx8, with initial restraints)	16
	EcoRV N185A-TAseq (1sx8)	50
	EcoRV N185pu-TAseq (1sx8)	50

The systems were equilibrated in three different stages with the numbers of particles, pressure (1 bar) and temperature kept constant (NPT ensemble) during 75 ps. In the first 25 ps, velocities were rescaled every 0.1 ps, and in the second 25 ps, Langevin dynamics were used to maintain constant temperature. Pressure control was introduced in the third 25 ps and in the production run using the Nosé–Hoover Langevin piston with a decay period of 500 fs. The harmonic restraints were gradually lifted [to 0.5, 0.25 and 0.05 kcal/(mol Å^2)] in the three equilibration stages.

After equilibration, the NPT production runs were performed for 50 ns. The integration time step was 2 fs and coordinates were saved with a sampling interval of 2 ps. All covalent bond lengths involving hydrogen were fixed using SHAKE algorithm.[71]

MD simulation on the crystal unit cell of EcoRV endonuclease (1SX8) without DNA was performed in explicit solvent with the NAMD program (using the CHARMM27 force field[68]). This triclinic crystal structure had unit cell dimensions $a\times b\times c$ of 47.8×49.1×63.7 Å,[58] according to the experimental space group symmetry, $P1$. The system was solvated in 7563 TIP3P[69] water molecules and 12 chloride counterions were added, leading to an electrically neutral system of 15,604 atoms. Periodic boundary conditions were applied to mimic the full crystalline environment.

Simulations were performed on the HELICS (Heidelberg Linux Cluster System) computer and using TeraGrid resources provided by NICS (National Institute for Computational Studies).

Analysis of trajectories

For the analysis, the first 2 ns of each trajectory were not included. The conformations of the complexed cognate and non-cognate DNA were characterized by calculating all six helical parameters (the three rotational parameters: roll, tilt and twist, and the three translational parameters: slide, rise and shift) that define the DNA geometry. In what follows, we present only the roll angle ρ because this was found to exhibit by far the largest difference between the complexed cognate and non-cognate DNA (see Fig. A, Supplementary Material). ρ is a rotational helical parameter measuring the angle between the planes formed between two consecutive base pairs. It is the primary mode of DNA bending and is especially important in protein–DNA interaction.[31,72–74]

Two atoms are considered here to form a hydrogen bond if the acceptor–hydrogen distance is <2.4 Å and the acceptor–hydrogen donor angle is >135°. Hydrogen-bond occupancy is calculated as the ratio of the time when the hydrogen bond is formed to the total time of the trajectory. We consider hydrogen bonds that have an occupancy of more than 80% to be very stable.

The deformation in the DNA is quantified by measuring a conformational variable ρ, here chosen to be the roll angle. For a system in thermodynamic equilibrium at constant temperature and pressure, the change in free energy on going from a reference state, defined by $\rho=\rho_{\mathrm{ref}}$, to a generic state, $\rho=\rho_i$, $\Delta\ G_{\mathrm{ref}\to i}$ is given by;

$$\Delta G_{\mathrm{ref}\to i} = -RT\ln\left[\frac{P(\rho_i)}{P(\rho_{\mathrm{ref}})}\right]$$

where R is the ideal gas constant, T is the temperature and $P(\rho_i)$ and $P(\rho_{\mathrm{ref}})$ are the probabilities of finding the system in states $\rho=\rho_i$ and $\rho=\rho_{\mathrm{ref}}$, respectively. $P(\rho)$ is obtained directly from the unbiased MD simulation, as has been performed in previous DNA simulation work.[64] Due to limited sampling, the method described here is strictly applicable only close to local minima of the free-energy surface.

All molecular images were generated with the molecular graphics program PyMOL.[75] Structural analysis and calculation of the free energy were performed using standard programs: Curves5.3,[76–78] Gromacs[79] tools and homemade scripts.

Acknowledgements

M.Z. thanks the Computational Molecular Biophysics group, as well as, Karine Voltz, Nicolas Calimet, Loukas Petridis, Peter J. Winn and Thomas Splettstoesser, for useful discussions. P.I. is grateful for funding from the Heidelberg BIOMS (Center for Modelling and Simulation in the Biosciences) initiative. J.C.S. was funded by a laboratory-directed research and development grant from the U.S. Department of Energy. This research was supported in part by the National Science Foundation through TeraGrid resources provided by National Institute for Computational Sciences.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2010.06.026

References

1. de Vargas, L. M., Kim, S. & Landy, A. (1989). DNA looping generated by DNA bending protein IHF and the two domains of lambda integrase. *Science*, **244**, 1457–1461.
2. Huang, H., Zhu, L., Reid, B. R., Drobny, G. P. & Hopkins, P. B. (1995). Solution structure of a cisplatin-induced DNA interstrand cross-link. *Science*, **270**, 1842–1845.
3. Mysiak, M. E., Bleijenberg, M. H., Wyman, C., Holthuizen, P. E. & van der Vliet, P. C. (2004). Bending of adenovirus origin DNA by nuclear factor I as shown by scanning force microscopy is required for optimal DNA replication. *J. Virol.* **78**, 1928–1935.
4. Shi, Q., Thresher, R., Sancar, A. & Griffith, J. (1992). Electron microscopic study of (A)BC excinuclease. DNA is sharply bent in the UvrB–DNA complex. *J. Mol. Biol.* **226**, 425–432.
5. Snyder, U. K., Thompson, J. F. & Landy, A. (1989). Phasing of protein-induced DNA bends in a recombination complex. *Nature*, **341**, 255–257.
6. Stenzel, T. T., Patel, P. & Bastia, D. (1987). The integration host factor of *Escherichia coli* binds to bent DNA at the origin of replication of the plasmid pSC101. *Cell*, **49**, 709–717.
7. Wang, H., Yang, Y., Schofield, M. J., Du, C., Fridman, Y., Lee, S. D. *et al.* (2003). DNA bending and unbending by MutS govern mismatch recognition and specificity. *Proc. Natl Acad. Sci. USA*, **100**, 14822–14827.
8. Parkinson, G., Gunasekera, A., Vojtechovsky, J., Zhang, X., Kunkel, T. A., Berman, H. & Ebright, R. H. (1996). Aromatic hydrogen bond in sequence-specific protein DNA recognition. *Nat. Struct. Biol.* **3**, 837–841.
9. Porschke, D., Hillen, W. & Takahashi, M. (1984). The change of DNA structure by specific binding of the cAMP receptor protein from rotation diffusion and dichroism measurements. *EMBO J.* **3**, 2873–2878.
10. Schultz, S. C., Shields, G. C. & Steitz, T. A. (1991). Crystal structure of a CAP-DNA complex: the DNA is bent by 90 degrees. *Science*, **253**, 1001–1007.
11. Seeman, N. C., Rosenberg, J. M. & Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl Acad. Sci. USA*, **73**, 804–808.
12. Lawson, C. L. & Carey, J. (1993). Tandem binding in crystals of a *trp* repressor/operator half-site complex. *Nature*, **366**, 178–182.
13. Luisi, B. F., Xu, W. X., Otwinowski, Z., Freedman, L. P., Yamamoto, K. R. & Sigler, P. B. (1991). Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. *Nature*, **352**, 497–505.
14. Otwinowski, Z., Schevitz, R. W., Zhang, R. G., Lawson, C. L., Joachimiak, A., Marmorstein, R. Q. *et al.* (1988). Crystal structure of *trp* repressor/operator complex at atomic resolution. *Nature*, **335**, 321–329.
15. Rhodes, D., Schwabe, J. W., Chapman, L. & Fairall, L. (1996). Towards an understanding of protein-DNA recognition. *Philos. Trans. R. Soc. London, Ser. B*, **351**, 501–509.
16. Roberts, R. J., Vincze, T., Posfai, J. & Macelis, D. (2007). REBASE—enzymes and genes for DNA restriction and modification. *Nucleic Acids Res.* **35**, D269–D270.
17. Kovall, R. A. & Matthews, B. W. (1999). Type II restriction endonucleases: structural, functional and evolutionary relationships. *Curr. Opin. Chem. Biol.* **3**, 578–583.
18. Parry, D., Moon, S. A., Liu, H. H., Heslop, P. & Connolly, B. A. (2003). DNA recognition by the EcoRV restriction endonuclease probed using base analogues. *J. Mol. Biol.* **331**, 1005–1016.
19. Radzicka, A. & Wolfenden, R. (1995). A proficient enzyme. *Science*, **267**, 90–93.
20. Lesser, D. R., Kurpiewski, M. R. & Jen-Jacobson, L. (1990). The energetic basis of specificity in the Eco RI endonuclease–DNA interaction. *Science*, **250**, 776–786.
21. Taylor, J. D. & Halford, S. E. (1989). Discrimination between DNA sequences by the EcoRV restriction endonuclease. *Biochemistry*, **28**, 6198–6207.
22. Thielking, V., Alves, J., Fliess, A., Maass, G. & Pingoud, A. (1990). Accuracy of the EcoRI restriction endonuclease: binding and cleavage studies with oligodeoxynucleotide substrates containing degenerate recognition sequences. *Biochemistry*, **29**, 4682–4691.
23. von Hippel, P. H. (1994). Protein-DNA recognition: new perspectives and underlying themes. *Science*, **263**, 769–770.
24. von Hippel, P. H. & Berg, O. G. (1986). On the specificity of DNA–protein interactions. *Proc. Natl Acad. Sci. USA*, **83**, 1608–1612.
25. Arber, W. (1979). Promotion and limitation of genetic exchange. *Science*, **205**, 361–365.
26. Winkler, F. K., Banner, D. W., Oefner, C., Tsernoglou, D., Brown, R. S., Heathman, S. P. *et al.* (1993). The

crystal structure of EcoRV endonuclease and of its complexes with cognate and non-cognate DNA fragments. *EMBO J.* **12**, 1781–1795.

27. Nwosu, V. U., Connolly, B. A., Halford, S. E. & Garnett, J. (1988). The cloning, purification and characterization of the Eco RV modification methylase. *Nucleic Acids Res.* **16**, 3705–3720.
28. Schulze, C., Jeltsch, A., Franke, I., Urbanke, C. & Pingoud, A. (1998). Crosslinking the EcoRV restriction endonuclease across the DNA-binding site reveals transient intermediates and conformational changes of the enzyme during DNA binding and catalytic turnover. *EMBO J.* **17**, 6757–6766.
29. Martin, A. M., Sam, M. D., Reich, N. O. & Perona, J. J. (1999). Structural and energetic origins of indirect readout in site-specific DNA cleavage by a restriction endonuclease. *Nat. Struct. Biol.* **6**, 269–277.
30. Kostrewa, D. & Winkler, F. K. (1995). Mg^{2+} binding to the active site of EcoRV endonuclease: a crystallographic study of complexes with substrate and product DNA at 2 Å resolution. *Biochemistry*, **34**, 683–696.
31. Dickerson, R. E. (1998). DNA bending: the prevalence of kinkiness and the virtues of normality. *Nucleic Acids Res.* **26**, 1906–1926.
32. Olson, W. K., Gorin, A. A., Lu, X. J., Hock, L. M. & Zhurkin, V. B. (1998). DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proc. Natl Acad. Sci. USA*, **95**, 11163–11168.
33. Packer, M. J., Dauncey, M. P. & Hunter, C. A. (2000). Sequence-dependent DNA structure: tetranucleotide conformational maps. *J. Mol. Biol.* **295**, 85–103.
34. Perona, J. J. & Martin, A. M. (1997). Conformational transitions and structural deformability of EcoRV endonuclease revealed by crystallographic analysis. *J. Mol. Biol.* **273**, 207–225.
35. Hiller, D. A., Rodriguez, A. M. & Perona, J. J. (2005). Non-cognate enzyme-DNA complex: structural and kinetic analysis of EcoRV endonuclease bound to the EcoRI recognition site GAATTC. *J. Mol. Biol.* **354**, 121–136.
36. Bareket-Samish, A., Cohen, I. & Haran, T. E. (1998). Direct versus indirect readout in the interaction of the trp repressor with non-canonical binding sites. *J. Mol. Biol.* **277**, 1071–1080.
37. Bareket-Samish, A., Cohen, I. & Haran, T. E. (2000). Signals for TBP/TATA box recognition. *J. Mol. Biol.* **299**, 965–977.
38. Lankas, F. (2004). DNA sequence-dependent deformability—insights from computer simulations. *Biopolymers*, **73**, 327–339.
39. Heddi, B., Oguey, C., Lavelle, C., Foloppe, N. & Hartmann, B. (2010). Intrinsic flexibility of B-DNA: the experimental TRX scale. *Nucleic Acids Res.* **38**, 1034–1047.
40. Arauzo-Bravo, M. J., Fujii, S., Kono, H., Ahmad, S. & Sarai, A. (2005). Sequence-dependent conformational energy of DNA derived from molecular dynamics simulations: toward understanding the indirect readout mechanism in protein–DNA recognition. *J. Am. Chem. Soc.* **127**, 16074–16089.
41. Sarai, A. & Kono, H. (2005). Protein–DNA recognition patterns and predictions. *Annu. Rev. Biophys. Biomol. Struct.* **34**, 379–398.
42. Rohs, R., West, S. M., Liu, P. & Honig, B. (2009). Nuance in the double-helix and its role in protein–DNA recognition. *Curr. Opin. Struct. Biol.* **19**, 171–177.
43. Perez, A., Lankas, F., Luque, F. J. & Orozco, M. (2008). Towards a molecular dynamics consensus view of B-DNA flexibility. *Nucleic Acids Res.* **36**, 2379–2394.
44. Pan, Y. P. & Nussinov, R. (2007). Structural basis for p53 binding-induced DNA bending. *J. Biol. Chem.* **282**, 691–699.
45. Duan, J. X. & Nilsson, L. (2006). Effect of Zn^{2+} on DNA recognition and stability of the p53 DNA-binding domain. *Biochemistry*, **45**, 7483–7492.
46. Ma, B. & Levine, A. J. (2007). Probing potential binding modes of the p53 tetramer to DNA based on the symmetries encoded in p53 response elements. *Nucleic Acids Res.* **35**, 7733–7747.
47. Paillard, G. & Lavery, R. (2004). Analyzing protein-DNA recognition mechanisms. *Structure*, **12**, 113–122.
48. Roy, S. & Sen, S. (2006). Exploring the potential of complex formation between a mutant DNA and the wild type protein counterpart: A MM and MD simulation approach. *J. Mol. Graphics Modell.* **25**, 158–168.
49. Falconi, M., Oteri, F., Eliseo, T., Cicero, D. O. & Desideri, A. (2008). MD simulations of papillomavirus DNA-E2 protein complexes hints at a protein structural code for DNA deformation. *Biophys. J.* **95**, 1108–1117.
50. Price, D. J. & Brooks, C. L., III (2002). Modern protein force fields behave comparably in molecular dynamics simulations. *J. Comput. Chem.* **23**, 1045–1057.
51. Villa, A., Fan, H., Wassenaar, T. & Mark, A. E. (2007). How sensitive are nanosecond molecular dynamics simulations of proteins to changes in the force field? *J. Phys. Chem. B*, **111**, 6015–6025.
52. Fujii, S., Kono, H., Takenaka, S., Go, N. & Sarai, A. (2007). Sequence-dependent DNA deformability studied using molecular dynamics simulations. *Nucleic Acids Res.* **35**, 6063–6074.
53. MacKerell, A. D. & Nilsson, L. (2008). Molecular dynamics simulations of nucleic acid-protein complexes. *Curr. Opin. Struct. Biol.* **18**, 194–199.
54. Matsumoto, A. & Lankas, F. (2002). Sequence-dependent motions of DNA: a normal mode analysis at the base-pair level. *Biophys. J.* **83**, 22–41.
55. Lankas, F., Sponer, J., Langowski, J. & Cheatham, T. E. (2003). DNA basepair step deformability inferred from molecular dynamics simulations. *Biophys. J.* **85**, 2872–2883.
56. Chandrasekaran, R. & Arnott, S. (1996). The structure of B-DNA in oriented fibers. *J. Biomol. Struct. Dyn.* **13**, 1015–1027.
57. Frederick, C. A., Quigley, G. J., van der Marel, G. A., van Boom, J. H., Wang, A. H. & Rich, A. (1988). Methylation of the EcoRI recognition site does not alter DNA conformation: the crystal structure of d (CGCGAm6ATTCGCG) at 2.0-Å resolution. *J. Biol. Chem.* **263**, 17872–17879.
58. Horton, N. C. & Perona, J. J. (2004). DNA cleavage by EcoRV endonuclease: two metal ions in three metal ion binding sites. *Biochemistry*, **43**, 6841–6857.
59. Selent, U., Rüter, T., Köhler, E., Liedtke, M., Thielking, V., Alves, J. *et al.* (1992). A site-directed mutagenesis study to identify amino acid residues involved in the catalytic function of the restriction endonuclease EcoRV. *Biochemistry*, **31**, 4808–4815.
60. Wenz, C., Jeltsch, A. & Pingoud, A. (1996). Probing the indirect readout of the restriction enzyme EcoRV. Mutational analysis of contacts to the DNA backbone. *J. Biol. Chem.* **271**, 5565–5573.
61. Selvaraj, S., Kono, H. & Sarai, A. (2002). Specificity of protein–DNA recognition revealed by structure-based potentials: symmetric/asymmetric and cognate/noncognate binding. *J. Mol. Biol.* **322**, 907–915.
62. Thielking, V., Selent, U., Köhler, E., Wolfes, H., Pieper, U., Geiger, R. *et al.* (1991). Site-directed mutagenesis

studies with EcoRV restriction endonuclease to identify regions involved in recognition and catalysis. *Biochemistry*, **30**, 6416–6422.

63. Vermote, C. L., Vipond, I. B. & Halford, S. E. (1992). EcoRV restriction endonuclease: communication between DNA recognition and catalysis. *Biochemistry*, **31**, 6089–6097.
64. Fuxreiter, M., Luo, M., Jedlovszky, P., Simon, I. & Osman, R. (2002). Role of base flipping in specific recognition of damaged DNA by repair enzymes. *J. Mol. Biol.* **323**, 823–834.
65. Macke, T. & Case, D. A. (1998). Modeling unusual nucleic acid structures. *Mol. Model. Nucleic Acids, 379–393.*
66. Imhof, P., Fischer, S. & Smith, J. C. (2009). Catalytic mechanism of DNA backbone cleavage by the restriction enzyme EcoRV: a quantum mechanical/molecular mechanical analysis. *Biochemistry*, **48**, 9061–9075.
67. Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E. *et al.* (2005). Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26**, 1781–1802.
68. MacKerell, A. D., Banavali, N. & Foloppe, N. (2000). Development and current status of the CHARMM force field for nucleic acids. *Biopolymers*, **56**, 257–265.
69. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935.
70. Darden, T., York, D. & Pedersen, L. (1993). Particle mesh Ewald: an *N* log(*N*) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092.
71. Evans, D. J. & Holian, B. L. (1985). The Nosé–Hoover thermostat. *J. Chem. Phys.* **83**, 4069–4074.
72. el Hassan, M. A. & Calladine, C. R. (1996). Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. *J. Mol. Biol.* **259**, 95–103.
73. Horton, N. C. & Perona, J. J. (2000). Crystallographic snapshots along a protein-induced DNA-bending pathway. *Proc. Natl Acad. Sci. USA*, **97**, 5729–5734.
74. Zhurkin, V. B., Lysov, Y. P. & Ivanov, V. I. (1979). Anisotropic flexibility of DNA and the nucleosomal structure. *Nucleic Acids Res.* **6**, 1081–1096.
75. DeLano, W. L. (2002). *The PyMol Molecular Graphics System.* DeLano Scientific, Palo Alto, CA.
76. Lavery, R. & Skelnar, H. (1998). *Curves v.5.3.* Institut de Biologie Physico-Chimique, Paris, France.
77. Lavery, R. & Sklenar, H. (1988). The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. *J. Biomol. Struct. Dyn.* **6**, 63–91.
78. Lavery, R. & Sklenar, H. (1989). Defining the structure of irregular nucleic acids: conventions and principles. *J. Biomol. Struct. Dyn.* **6**, 655–667.
79. Spoel, D. V. D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E. & Berendsen, H. J. C. (2005). GROMACS: fast, flexible, and free. *J. Comput. Chem.* **26**, 1701–1718.