6936–6945 *Nucleic Acids Research, 2012, Vol. 40, No. 14* *Published online 11 April 2012*
doi:10.1093/nar/gks320

Non-specific protein–DNA interactions control I-CreI target binding and cleavage

Rafael Molina[1], Pilar Redondo[1], Stefano Stella[1], Marco Marenchino[2], Marco D'Abramo[3], Francesco Luigi Gervasio[3], Jean Charles Epinat[4], Julien Valton[4], Silvestre Grizot[4], Phillipe Duchateau[4], Jesús Prieto[1,*] and Guillermo Montoya[1,*]

[1]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Macromolecular Crystallography Group, [2]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), NMR Unit, [3]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Computational Biophysics Group, c/Melchor Fdez. Almagro 3, 28029 Madrid, Spain and [4]CELLECTIS S.A., 8 rue de la croix Jarry, 75013 Paris, France

Received January 18, 2012; Revised March 26, 2012; Accepted March 27, 2012

ABSTRACT

Homing endonucleases represent protein scaffolds that provide powerful tools for genome manipulation, as these enzymes possess a very low frequency of DNA cleavage in eukaryotic genomes due to their high specificity. The basis of protein–DNA recognition must be understood to generate tailored enzymes that target the DNA at sites of interest. Protein–DNA interaction engineering of homing endonucleases has demonstrated the potential of these approaches to create new specific instruments to target genes for inactivation or repair. Protein–DNA interface studies have been focused mostly on specific contacts between amino acid side chains and bases to redesign the binding interface. However, it has been shown that 4 bp in the central DNA sequence of the 22-bp substrate of a homing endonuclease (I-CreI), which do not show specific protein–DNA interactions, is not devoid of content information. Here, we analyze the mechanism of target discrimination in this substrate region by the I-CreI protein, determining how it can occur independently of the specific protein–DNA interactions. Our data suggest the important role of indirect readout in this substrate region, opening the possibility for a fully rational search of new target sequences, thus improving the development of redesigned enzymes for therapeutic and biotechnological applications.

INTRODUCTION

The processes of transcription, recombination or DNA replication require that the DNA be untwisted prior to the initiation of any of these processes. This untwisting is often initiated at particularly thermodynamically labile sequences; thereby bending flexibility is also an essential aspect of biological function. In addition, DNA bending has been shown to be a critical feature of catalysis and target recognition in some restriction enzymes (1–4). This phenomenon known as indirect readout is defined as sequence specificity occurring in the absence of hydrogen bonds and van der Waals interactions between the protein and the DNA base functional groups. Indirect readout has been proposed to involve contacts mediated by water or other small molecules, as well as distortions of DNA that can distinguish different sequences energetically (5).

LAGLIDADG homing endonucleases are sequence-specific enzymes that recognize and cleave long DNA targets (12–45 bp) generating a double-strand break (DSB). Structure analysis revealed that the central DNA target region displays a strong bending, thus resulting in base twisting and un-stacking near the scissile phosphate groups, which allows the proper binding and positioning in the active site (6–9).

I-CreI is a homodimeric LAGLIDADG family member, which recognizes and cleaves a 22-bp pseudo-palindromic target. Each monomer contains its own DNA-binding region and the catalytic center is formed at the dimer interface. The structure of non-digested substrate complexes (determined in the presence of non-activating Ca^{2+} ions) shows the presence of two Ca^{2+} ions at the active site (6) and the cleaved substrate structures had three Mg^{2+}/Mn^{2+}

*To whom correspondence should be addressed. Tel: +34 91 2246900; Fax: +34 91 2246976; Email: gmontoya@cnio.es
Correspondence may also be addressed to Jesús Prieto. Tel: +34 91 2246900; Fax: +34 91 2246976; Email: jprieto@cnio.es

ions in the active site (10), resembling the canonical two-metal-ion catalytic mechanism (11,12). This configuration requires two acidic residues at the carboxyl-termini of the LAGLIDADG helices in the active site and the coordination of divalent metal ions. In the case of I-CreI, these acidic residues correspond to the D20 residue in each monomer, which participate in the cleavage of the DNA strands along the minor groove, resulting in the hydrolysis of specific phosphodiester bonds (10).

The analysis of the I-CreI crystal structure bound to its natural target shows that in each monomer nine residues establish direct interactions with seven bases (6), mostly grouped in two boxes that previously were called 5NNN, located at positions ±3, ±4, ±5 and 10NNN, located at positions ±8, ±9, ±10 (13,14) (Figure 1a). Recently, another amino acid region involved in DNA binding has been described in I-CreI, the 7NN, located at positions ±6, ±7 (15). The 4 bp (±1 and ±2), called 2NN in the center of the homing site (Figure 1a), only show a backbone contact between the base at position −1 (both strands) and K139 (of each I-CreI monomer) (16). However, changes in this region produce a strong impact on substrate binding and cleavage. This observation is supported by data which show that methylation of the central +1 base (both strands) reduces protein binding (17). There is no obvious explanation for this behavior as there is enough room for a methyl group at the N7 position of this base. In accordance with previous reports (14), the four central base pairs have a crucial role in determining overall substrate specificity. The influence of the central sequence has been explained by its topology. However, the role of these four bases, which do not make direct contacts with the protein, seems to be fundamental to determine enzyme binding and cleavage activity.

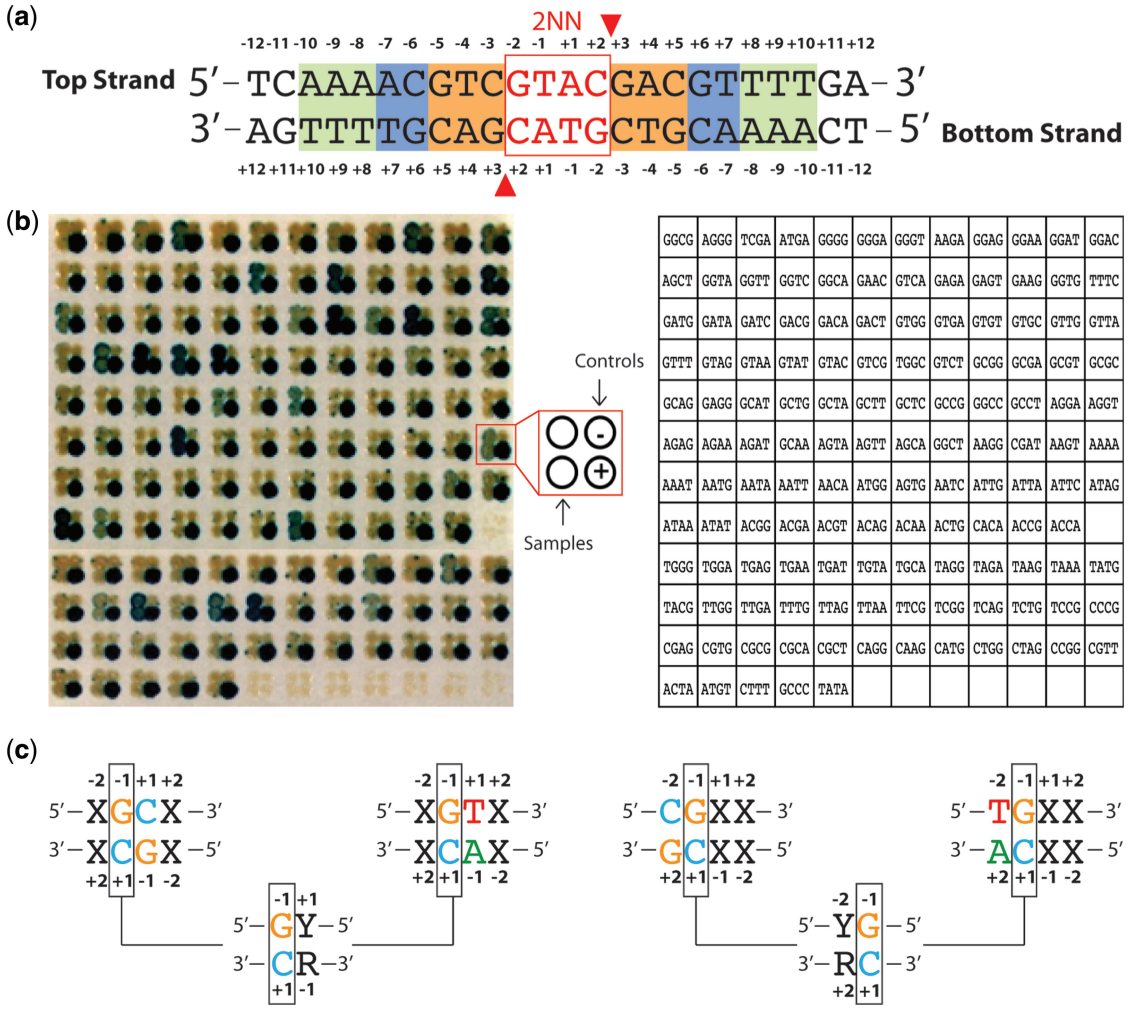

Figure 1. I-CreI *in vivo* cleavage patterns. All the sequences are described in 5′–3′ (**a**) Scheme of the DNA target regions involved in homing endonuclease binding or catalysis. The protein–DNA contacts are focused in three nucleotide regions: 10NNN (±8, ±9, ±10 boxed in green), 7NN (±6, ±7 boxed in blue) and 5NNN (±3, ±4, ±5 boxed in orange). The 2NN region (±2, ±1) is located at the cleavage site (red rectangle, red \triangles indicate the scissile +3PO). (**b**) I-CreI cleavage activity using different 2NN DNA targets. From 256 possible 2NN targets, 136 were analyzed *in vivo* (left panel) discarding 120 sequences because they represent the complementary targets (Supplementary Figure S1). Each target contains the experimental result duplicate, two left dots and the controls, two right dots (see schematic distribution of each target between panels). The right panel display the 5′–3′ top strands 2NN sequences. (**c**) Non-cleaved targets display a common sequence pattern at −1 position, where a G is flanked by pyrimidines.

Here, we conduct a comprehensive study of the impact of the 2NN target bases in binding and catalysis using an *in vivo* cleavage/recombinational screening, fluorescence anisotropy and a structural analysis, suggesting a mechanism governing target discrimination not based on specific protein–DNA contacts. This finding will aid to further rationalize the search of target sequences in the development of new engineered homing endonucleases, also called meganucleases, for therapeutic and biotechnological applications.

MATERIALS AND METHODS

I-CreI *in vivo* cleavage experiments

The I-CreI gene was cloned and expressed as in (18,19). A palindromic 24-bp DNA target (Figure 1a), derived from the pseudo-palindromic I-CreI wild-type target (17), was used as a template for the *in vivo* cleavage experiments. Since this palindromic scaffold is cleaved by I-CreI with the same efficiency as the wild-type pseudo-palindromic target (14), this template allowed us to assess binding and cleavage of the 256 possible four base combinations by changing only the four central bases. All the sequences are described in 5′–3′. I-CreI was screened against 136 2NN derived targets (120 non-palindromic, 16 palindromic). These 136 targets represent the total 256 possible four central base-pair combinations (Figure 1). The other 120 sequences are the complementary targets of the non-palindromic sequences (Supplementary Figure S1). Two different yeast cell lines are transformed with the meganuclease expression vector and the reporter plasmid containing the target site. The strain harboring the expression vector encoding the I-CreI gene is mated with another strain bearing the reporter plasmid. In the reporter plasmid, a LacZ gene is interrupted with an insert containing the I-CreI 24-bp target sequence differing only in the 2NN region. The insert is flanked by two direct repeats needed for single-strand annealing (SSA) repair. Upon mating, if the enzyme recognizes the DNA target, it will generate a DSB on the site of interest, allowing restoration of a functional LacZ gene by SSA between the two flanking direct repeats. Expression of the LacZ gene can be visualized by X-Gal staining (13). The restoration of the beta-galactosidase activity is directly associated with the homologous recombination efficiency. However, experiments using several purified I-CreI mutants with various recombination activities in yeast have shown that the recombination efficiency quantified in yeast directly correlate with the cleavage activity *in vitro* (Grizot,S. and Valton,J., unpublished data). The I-CreI scaffold used in this work contains the D75N mutation. The D75N mutation does not affect protein structure and facilitates the enzyme purification. I-CreI and its D75N variant display similar *in vitro* activities and levels of specificity (13).

Fluorescence polarization binding assays

The dissociation constants for I-CreI DNA binding (Figure 2) were determined from the change in anisotropy of the different 6-FAM-labeled DNAs when the protein–DNA complex is formed. The optimal concentration of

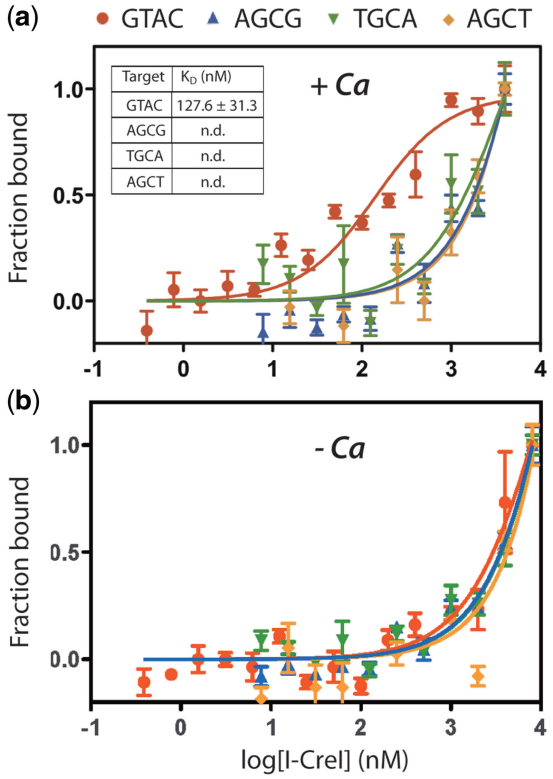

Figure 2. Target binding by fluorescence polarization assays. I-CreI logarithmical representation of DNA binding using different 2NN target sequences in presence (**a**) and absence (**b**) of 2 mM $CaCl_2$. The targets include the wild-type (GTAC) and three non-cleavable (AGCG, TGCA and AGCT) 2NN sequences. The wild-type K_d was determined for comparison with the other targets, which did not display a detectable K_d under the same experimental conditions.

6-FAM-DNA that gives a stable polarization signal avoiding measurement artifact was determined empirically by measuring the fluorescence polarization of a serial diluted 6-FAM-DNA sample (20). Thus, a solution including 25 nM 6-FAM-DNA and various concentrations (0–400 nM) of I-CreI were prepared in a total volume of 50 µl binding buffer (10 mM Tris pH 8.0, 300 mM NaCl, 10 mM $CaCl_2$). After incubation at 30°C for 15 min, the fluorescence intensities parallel (I_{VV}) and perpendicular (I_{VH}) to the plane of excitation were measured in a black 96-well assay plate (Corning) on a Wallac Victor^{2}V 1420 multilabel counter (PerkinElmer) using a 485-nM excitation filter, a 535-nM emission filter, a 1-s per well reading time and a Z height set at 4 mm. The steady-state fluorescence anisotropy (r) was calculated using the equation:

$$r = \frac{I_{VV} - GI_{VH}}{I_{VV} + 2GI_{VH}}$$

where the correction factor G ($G = 1.12$ determined empirically using a fluorescein standard) was introduced to correct the polarized light transmission efficiency of the instrument optics (21). The fluorescence quantum yield of the 6-FAM-DNA did not change upon binding to the I-CreI and the molar fraction of bound (FSB) of

6-FAM-DNA was calculated following the principle of additivity of anisotropies:

$$F_{SB} = \frac{r - r_F}{r_B - r_F}$$

where r is the observed anisotropy, r_F and r_B are the anisotropies of the free and bound 6-FAM-DNA, respectively (22). After normalization, the dissociation constant (K_d) was determined by data fitting (Prism, GraphPad) to the following equation:

$$F_{SB} = \frac{1}{2L_T}\left|(L_T+P_T+K_d) - \sqrt{(L_T+P_T+K_d)^2-(4L_TP_T)}\right|$$

where L_T is the total 6-FAM-DNA concentration and P_T is the total protein concentration (23).

Crystallization

The 24-bp DNA targets were purchased from Proligo and consisted of two strands containing the following sequences: 5′-TCAAAACGTCGTACGACGTTTTGA-3′ (GTAC, wild-type), 5′-TCAAAACGTCAGCGGACGTTTTGA-3′ and 5′ TCAAAACGTCCGCTGACGTTTTGA-3′ (AGCG, no cleavage) and 5′-TCAAAACGTCTGCAGACGTTTTGA-3′ (TGCA, no cleavage). All of them form a 24-bp blunt-end duplex after incubation. The protein–DNA complexes were obtained in the absence of metal ions and in the presence of non-catalytic Ca^{2+} (for I-CreI: GTAC and I-CreI: GTAC-Ca^{2+}), to obtain the uncleaved states and in the presence of 2 mM $MgCl_2$ (for I-CreI: GTAC, I-CreI: AGCG and I-CreI: TGCA), to obtain the cleaved states of the target DNA, respectively. Complexes were formed by pre-warming the meganuclease and the oligonucleotide samples at 37°C and mixing them in a 0.75:1 molar ratio (DNA:protein). The mixture was incubated for 50 min at this temperature and then spun down for 5 min to remove insoluble material. The final concentration of protein in the DNA–protein complex solution was 4 mg/ml. I-CreI: GTAC-Mg^{2+} crystals were grown using the hanging-drop method at 290 K, in 2 µl droplets formed by 1 µl of the DNA–protein complex (containing 2 mM of Mg^{2+}) and 1 µl of precipitant solution consisting of 30% (v/v) 1,2-propanediol, 20% PEG-400 in 0.1 M HEPES pH 7.5.

I-CreI: GTAC crystals were also grown from 2 µl droplets formed by 1 µl of the DNA–protein complex (in absence metal ions) and 1 µl of precipitant solution consisting of 20% PEG-300, 5% PEG-8000, 10% Glycerol in 0.1 M Tris pH 8.5.

I-CreI: GTAC-Ca^{2+} crystals were also grown from 2 µl droplets formed by 1 µl of the DNA–protein complex (containing 2 mM of Ca^{2+}) and 1 µl of precipitant solution consisting of 30% PEG400, 0.1 M NaCl in 0.1 M HEPES pH 7.5. The crystallization conditions for trying to obtain the I-CreI: AGCG-Mg^{2+} and I-CreI: TGCA-Mg^{2+} were 50% PEG-400, 0.2 M NaCl in 0.1 M CHES pH 9.5 and 35% methanol in 0.1 M sodium cacodilate pH 6.5, respectively, both in the presence of Mg^{2+} (2 mM for I-CreI: AGCG-Mg^{2+} and up to 30 mM I-CreI: TGCA-Mg^{2+}). This complexes were cryo-protected by adding 20% (v/v) ethylenglycol to the mother liquor. I-CreI: GTAC-Ca^{2+} was also cryo-protected by adding 10% (v/v) PEG400 to the mother liquor. The rest were directly flash-frozen in liquid nitrogen.

Data collection, structure solution, model building and refinement

All data were collected at 100 K, using synchrotron radiation at the PX beam line (SLS, Villigen, Switzerland). The diffraction pattern was recorded on Pilatus detectors. Data processing and scaling were accomplished with XDS (24) or MOSFLM (25) (see Supplementary Table S4). The structures were solved by molecular replacement as implemented in the programs MOLREP (26) or PHASER (27). The search model was based on a poly-alanine backbone derived from the PDB entry 1G9Z (I-CreI–DNA–Mg^{2+}). The structures were then subjected to iterative cycles of model building and refinement with O (28), Coot (29) and PHENIX (30). The identification and analysis of the protein–DNA hydrogen Bonds and van der Waals contacts was done with the Protein Interfaces, Surfaces and Assemblies service PISA at the European Bioinformatics Institute (http://www.ebi.ac.uk/msdsrv/prot_int/pistart.html). DNA structures were analyzed using 3DNA (31).

Molecular dynamics

Three molecular dynamics simulations of the I-CreI bound to the wild-type DNA and to the DNA oligomers (TGCA and AGCG) were performed starting from the corresponding crystallographic structures. The version 4.5 of the molecular dynamics program GROMACS (32) with Amber99sb-ildn* (33) force field were used. The protein–DNA complexes were solvated and thermalized. In the solvation step, two additional calcium ions were added in the proximity of the cleavage site, using as a reference the crystal structure of the I-CreI bound to the GTAC in absence of Mg^{2+}. It is worth noting that the DNA oligomers are in a conformation that is not compatible with the cleavage, i.e. the +2PO (metal positioning phosphates) are far away from the LAGLIDADG motif. After 10 ns of equilibration, three runs of 100 ns each were performed, using a time step of 2 fs. The temperature was kept constant at 310 K using the velocity rescaling algorithm (34).

RESULTS

The DNA bases located in the 2NN region affect substrate cleavage

We have previously described an assay to monitor meganuclease-induced recombination in yeast cells (35). This method allows for the screening of the same meganuclease with different targets. In a previous study, we used this procedure to show that not all 2NN targets can be cleaved by I-CreI and its engineered variants (14). To further analyze the role of the 2NN sequence combination in meganuclease function, we tested all possible base combination of the 2NN region allowing binding and cleavage by I-CreI in yeast.

I-CreI was screened *in vivo* against the 136 2NN targets derived from this template (Figure 1). This 136 target

screening represents the total 256 possible four central base-pair combinations—120 non-palindromic and 16 palindromic (see 'Materials and Methods' section and Supplementary Figure S1). Only 41 2NN targets were cleaved out of 136 tested (Supplementary Figure S1). A detailed analysis of the combinatorial 2NN panel cleavage pattern revealed preferences depending on base position. Thus, G and T are the most frequent bases at position −2, T is the preferred one at position −1, A at position +1 and A and C at position +2 (Supplementary Table S1). When the analysis takes into account the adjacent bases and different base-pair combinations are evaluated, we observed that the GC and GT pairs at positions −1 and +1 were not cleaved. Similarly TG and CG pairs at positions −2 and −1 do not show any cleavage (Supplementary Table S2). The analysis shows that a G in position −1 prevents cleavage of most targets and the phosphodiester hydrolysis is only allowed when a G and/or A is located in the −2 and +1 positions (Supplementary Table S2). Therefore, the G in position −1 flanked by pyrimidines (Y) avoids cleavage, however, when it is flanked by purines (R) the phosphodiester digestion could be allowed (Figure 1c and Supplementary Table S2). To compare these tendencies with other members of the LALIDADG family, we examined the 2NN sequences for 30 LAGLIDADG enzymes whose targets are known (Supplementary Table S3). Only one of them, I-CpaI, contains the unfavorable CG at position −2 and −1. The rest do not contain the non-cleavable 2NN sequences. Furthermore, they display a high frequency of T at position −1, similarly to I-CreI, supporting our observations. The analysis of the 2NN region using the base-pair combination provides a better description of the central region (Figure 1b and c; Supplementary Tables S1 and S2). Using this information from our two base position analysis, we suggest a non-cleavable sequence preferences based on the base stacking energy from the different dinucleotide steps. Pyrimidine–purine (Y–R) base-pair steps tend to be more prone to slide and shift changes, along with lower propeller twisting. Hence they are more susceptible of bending than R–Y steps. These correlations are particularly evident for certain Y–R steps like TA and CA that tend to exhibit the highest ranges of slide (36).

The 2NN non-cleavable targets hinder endonuclease binding

To analyze whether differences in cleavage could arise from changes in DNA-binding to different 2NN DNA targets, we monitored I-CreI binding to different 24 bp targets, containing different types of sequence combinations in the central four bases containing the XGCX pattern by fluorescence anisotropy. Only the wild-type sequence was recognized in the presence of Ca^{2+} (a non-catalytic metal) and a K_d was calculated for comparison purposes (Figure 2a). The other three targets containing identical DNA sequence in the 5NNN and 10NNN regions, which make direct amino acids side chains to base contacts, displayed a non-detectable K_d. Our data indicate that beyond its role in catalysis, the metal ion is key for binding at the protein concentrations tested. Even the wild-type target binding was not detected in the absence of Ca^{2+} (Figure 2b). The interaction was only observed at higher protein concentrations, suggesting that the cation presence is essential for binding at physiological conditions.

The 2NN non-cleavable sequences do not allow metal positioning in the catalytic site

To understand at the molecular level the protein–DNA interaction between the wild-type 2NN sequence and the XGCX targets, we shifted the equilibrium toward complex formation to crystallize and solve the structures of different protein–2NN DNA complexes in the presence and absence of metal (Supplementary Table S4). This includes I-CreI in complex with the wild-type DNA target in the presence of Ca^{2+}, Mg^{2+} or without metal and the same enzyme bound to two different XGCX targets, the AGCG and TGCA, including 2 and 30 mM Mg^{2+} in the crystallization conditions (see Materials and Methods). The protein–DNA complex containing the wild-type DNA sequence (2NN GTAC) in presence of Mg^{2+} showed the corresponding phosphodiester bonds hydrolyzed, generating a DSB and displaying the catalytic ions in the previously reported positions (10). In the absence of Mg^{2+}, the substrate was not cleaved and the cation positions were unoccupied. However, when the complex was assembled in the presence of Ca^{2+}, a non-catalytic metal, two ions were observed in the catalytic center and both DNA strands exhibited a conformation similar to the cleaved structure but the phosphodiester bonds were not digested (Supplementary Figure S2). Interestingly, when the protein–DNA complex was formed without any cation present the structure revealed a very different DNA conformation in the catalytic center (Figure 3a). While the distance between the scissile phosphates (+3PO) in the absence of cation is 10.5 Å, in the presence of Ca^{2+} the distance is almost 5 Å shorter, indicating that the presence of the metal rearranges the 2NN region conformation. The entrance of Ca^{2+} or Mg^{2+} ions generates the same rearrangement in the DNA conformation supporting the structural data obtained with the Ca^{2+} (Supplementary Table S2). The large compression of the minor groove in the 2NN region shows that the flexibility of the nucleic acid in this region is important to adopt the proficient conformation for cleavage in I-CreI (Figure 3b). Moreover, the analysis of the wild-type GTAC structure in the presence of Ca^{2+} exhibits a DNA helix overwinding with an average twist angle of 36.4°. The distribution of twist angle—the DNA parameter that describes the angle between 1 bp rotated about the double helix long axis relative to the adjacent base pair—in the presence and in the absence of Ca^{2+} did not display significant variations except for positions −3C/−2G, −1T/+1A and +2C/+3G, where the twist angle changes beyond 10° (Figure 3c and Supplementary Figure S3a).

However, when the crystal structures were solved in complex with the DNA containing the non-cleavable AGCG and TGCA sequences in the 2NN region, neither

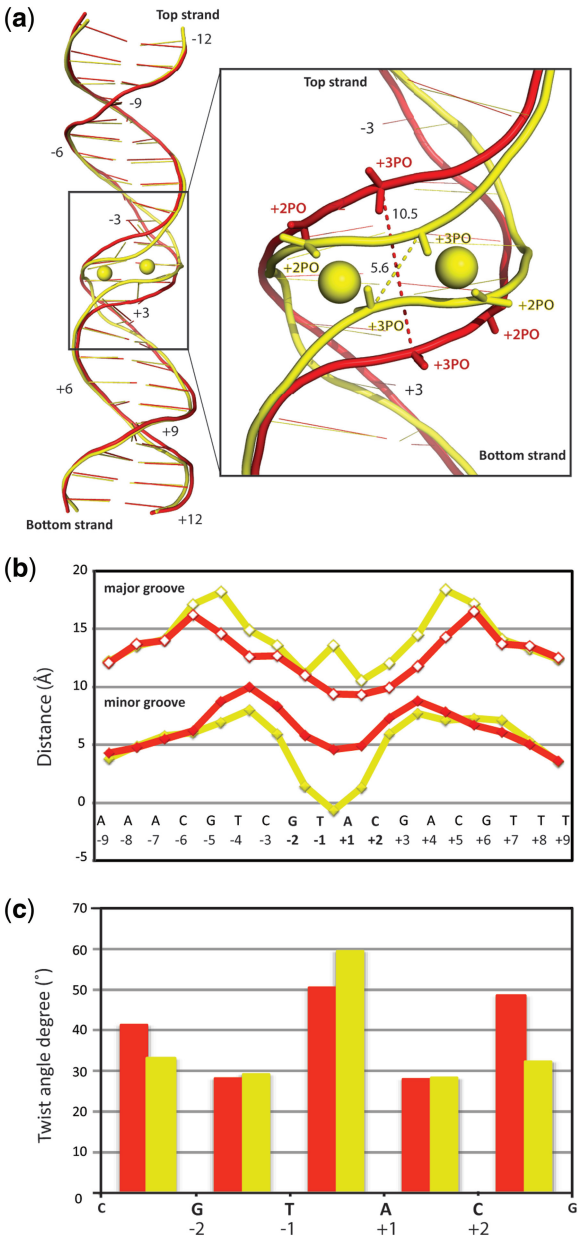

Figure 3. Comparison of the DNA structure in the wild-type target with non-catalytic ions and without active site ions. (**a**) Superimposition of the protein-GTAC target structures in presence of non-catalytic Ca^{2+} ion (yellow) and in absence of active site ions (red). The protein moiety has been omitted for clarity. (**b**) Major and minor groove width along DNA in both GTAC structures. Values were calculated subtracting the van der Waals surfaces. (**c**) Twist parameter distribution at 2NN region.

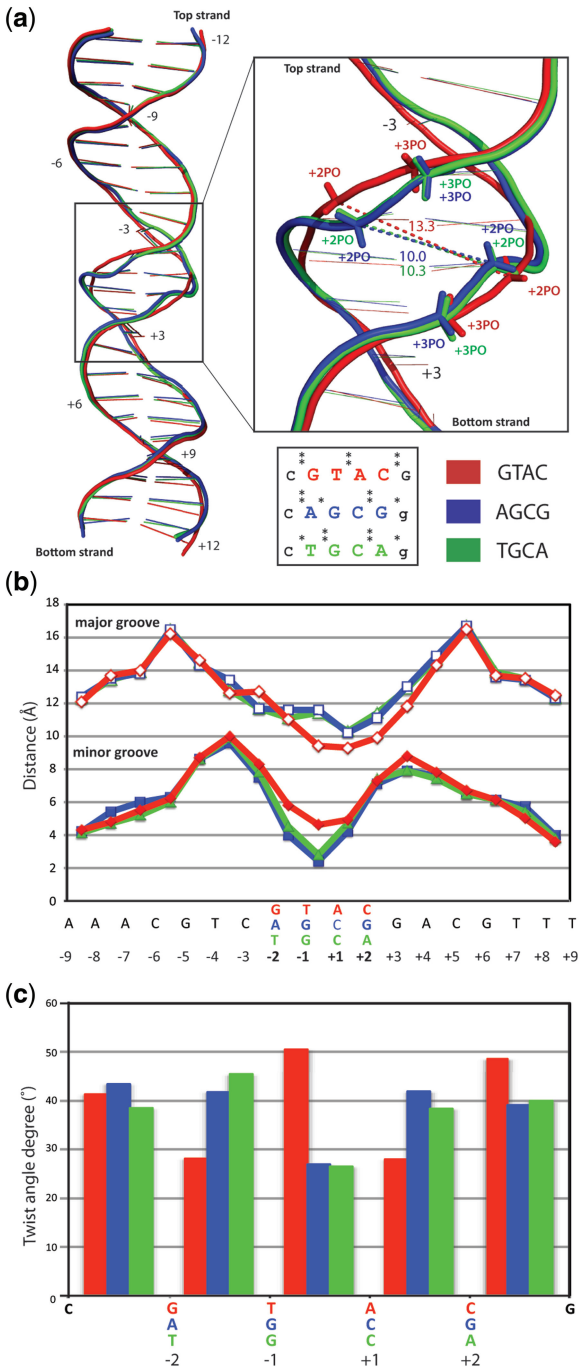

Figure 4. Comparison of the non-cleaved DNA wild-type target and the XGCX pattern structures. (**a**) Superimposition of the non-cleaved GTAC, AGCG and TGCA target structures. The protein moiety has been omitted for clarity. The asterisk indicates the degree of flexibility, ranging from 0 to **, based on the overlapping surface between dinucleotide steps (36). (**b**) Major and minor groove width along DNA in the structures. Values were calculated subtracting the van der Waals surfaces. (**c**) Twist parameter distribution at 2NN region.

the ions nor phosphodiester bond hydrolysis were observed in the catalytic site (Figure 4a and Supplementary Figure S4) although high concentrations of Mg^{2+} were present in the crystallization solution (see Materials and Methods), thus supporting the *in vivo* cleavage data. The superimposition of the non-cleavable 2NN structures shows differences between the putative metal positioning phosphates (+2PO) in the XGCX structures and the wild-type (Figure 4a). The distance between the positioning phosphates in position +2 in GTAC is 13.3 Å; however, in the XGCX targets is around 3 Å shorter. Thus, we assessed whether the active site conformation in the XGCX targets could hinder the metal entrance modeling the cations in the positions where they are located in the wild-type bound and

cleaved structures. However, no clashes between the cations and the DNA or protein were detected using molprobity (37), suggesting that the cations can enter the active site but they are not properly positioned, and thus phosphodiester bond hydrolysis is not accomplished. The analysis of wild-type GTAC, the AGCG and TGCA structures shows that the minor groove width in the GTAC 2NN region is significantly larger than those of AGCG and TGCA (Figure 4b). In addition, it also reveals two DNA underwinding regions from position −8 to −3 and from 4 to 8 with average twist values of 31 ± 0.3° and 32 ± 0.1°, respectively. These regions correspond to the two expanded minor groove areas (Supplementary Figure S3b). In contrast, in all the structures the 2NN region (comprising positions −2 to 3) displays a compression of the minor groove associated with a DNA helix overwinding with an average twist of 38.5 ± 0.7° (Figure 4b and c). Although the average twist in the 2NN region is very similar, the distribution of the single twist angles depends on the base-pairs steps and is different depending on the 2NN region sequence (Figure 4c).

The wild-type GTAC shows a twist value of 54.48° at position −1T/+1A while the same position −1G/+1C both in AGCG and TGCA, the twist value is 27° and 26.5°, respectively. Remarkably, the twist values at the adjacent positions in the GTAC structure are around 28° at both −2G/−1T and the +1A/+2C positions while for the −2A/−1G and +1C/+2G in the AGCG structure and −2T/−1G and +1C/+2A in the TGCA structure the average twist value is around 42°. The neighboring base positions −3/−2 and +2/+3 in the AGCG and TGCA structures recover the twist values found in the wild-type GTAC. Therefore, TA dinucleotide steps found in the wild-type sequence, which are the most flexible ones allowing higher twist angle values (36,38), permit the proper conformation to achieve function (Figure 4c).

Molecular dynamics simulations

To provide an animated view of the process, we modeled the DNA conformational behavior by molecular dynamics (MD) (39–41) using the structures containing the wild-type DNA sequence (GTAC) without metal and the two structures with the XGCX sequences (Figure 5). We started our MD simulations with two Ca^{2+} ions at the cleavage site and the DNA conformation not compatible with cleavage (obtained from the X-ray structures crystallized without Mg^{2+} ions). Interestingly, in the cleaved GTAC sequence, the distance between the +2PO and the LAGLIDADG motif suddenly changes to lower values with respect to the initial structure, while in the case of the TGCA and AGCG sequences, this distance is always larger, ranging between 6 and 10 ± Å (Figure 5a). Also, we detected that in the case where phosphodiester hydrolysis occurs, the two central bases (TA) of the DNA oligomer, controlling recognition and cleavage, show a relevant distortion. In fact, the most probable value of the roll angle is ∼−25°, indicating a remarkable difference from the ideal B-DNA and the other two non-cleavable sequences (average roll angle −10°) (Figure 5b, central panels). Moreover, these 2 bp show large and negative values of the opening angles

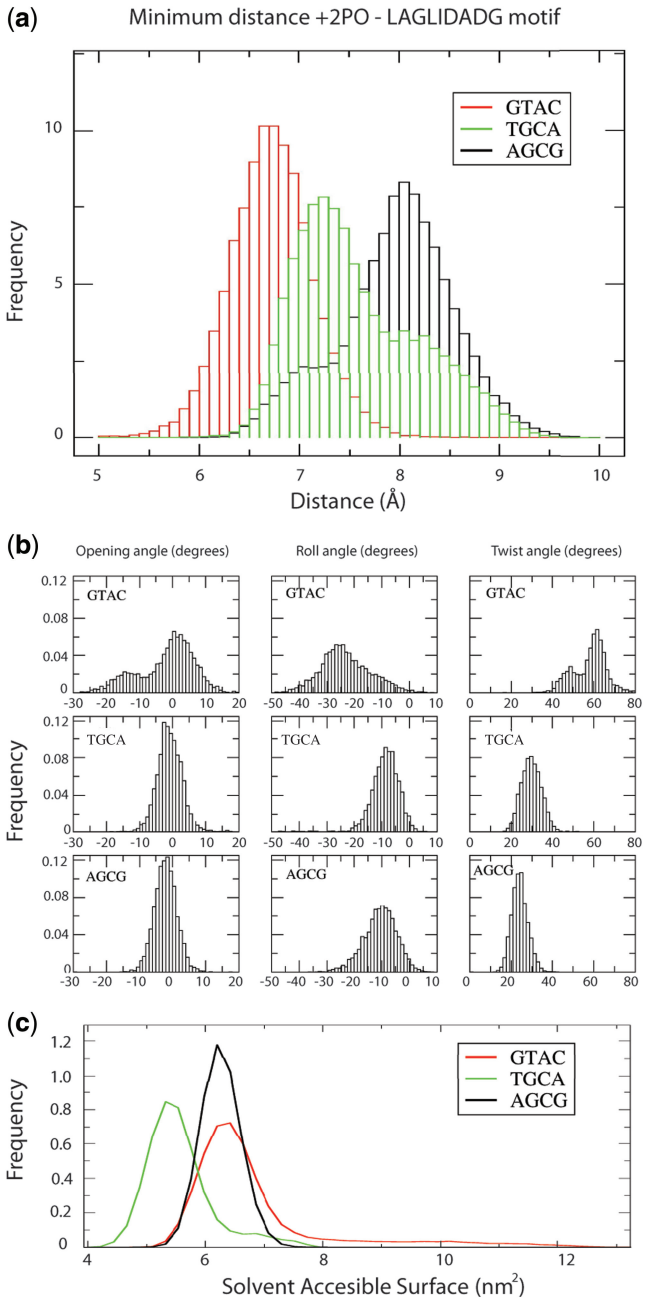

Figure 5. Molecular dynamics simulations. (**a**) Distribution of the minimum distances between the positioning phosphates (+2PO) of the three non-cleaved DNA sequences reported in Figure 4 and the LAGLIDADG motif extracted from the MD trajectories. (**b**) Distribution of the opening, roll and twist angles in the three target sequences. (**c**) Solvent accessible surface distributions of the 4 bp forming the cleavage site.

with respect to the corresponding base pairs of the TGCA and AGCG sequences (Figure 5b, left panels). The analysis of the twist angle shows that the two central bases of the GTAC target adopt average values significantly larger than the XGCX sequences, supporting the structural analysis (Figure 5b, right panels). Remarkably, the dinucleotide steps at the cleavage site of GTAC are more exposed to the solvent (Figure 5c).

DISCUSSION

Protein–DNA recognition is fundamental for the fate of a cell. Although we have progressed in our understanding of these interactions, we still need to improve our knowledge to be able to generate new protein–DNA interactions that would allow the manipulation of a genome. LAGLIDADG homing endonucleases are one of the scaffolds widely used to create protein 'cutters' able to produce a DSB in a desired genome region once their DNA specificity has been redesigned. However, the central 2NN region, which does not display specific protein–DNA contacts, can drastically impact DNA cleavage, as illustrated previously for I-CreI (14) and engineered I-CreI derivatives. *In vivo* experiments (Figure 1; Supplementary Figure S1, Supplementary Tables S1 and S2) clearly show that certain combinations of bases in these positions are not compatible with the enzyme function. Especially the XGCX pattern is strongly avoided when the scaffold is redesigned to create new variants. Moreover, these bases are absent in the 2NN positions in members of the LAGLIDADG family with well-characterized DNA targets (Supplementary Table S3).

The DNA double helix is a highly flexible structure that can bend and twist and the biological consequences of its physicochemical properties depend on sequence-dependent differences affecting the energetics of bending or twisting (38). Base composition and sequence have an essential role in DNA flexibility and thus in adaptability to its partner. DNA flexibility is mainly controlled by the energetics of base stacking. Purine–Purine (R–R) or Pyrimidine–Pyrimidine (Y–Y) base-pair steps are relatively well stacked, particularly over the purine rings and the oligonucleotide crystallographic data show them to be intermediate in flexibility. R–Y dinucleotide steps share the greatest surface area of the three classes and thus are not associated with large bends in protein complexes (36). However, Y–R steps provide a minimum overlap between them allowing a wider conformational sequence-dependent landscape (38,42). Our dinucleotide analysis suggests that sequence alterations in the 2NN region increasing the stacking energy and decreasing DNA flexibility (i.e. TA to GC) disturb cleavage. Furthermore, the 2NN sequence pattern that always abrogates cleavage contains a G in position −1 and a C in +1 in the 5′–3′ direction (Figure 1c), as observed *in vivo* and *in vitro* cleavage assays (see Supplementary Figure S5).

To examine whether the sequence of the 2NN region may affect DNA target recognition we studied its binding affinity to DNA sequences whose only variations were restricted to this area (Figure 2). We observed that the enzyme displayed a decrease in the affinity for its target in the absence of cation. The comparison of the measurements were performed at protein concentrations close to the cellular level in the presence and absence of Ca^{2+} indicating that the enzyme target recognition is enhanced when the metal is present. Although the specific contacts between the protein and the DNA 5NNN and 10NNN regions are conserved, the enzyme did not cleave the target sequences that contained the GC pair in the 2NN region, due to the lack of metal positioning. Hence the cellular assembly of the protein–DNA complex seems to be dependent on the DNA sequence in this area, regardless of the specific contacts arising from the interaction between other protein and DNA sections. This suggests the need of a restrictive sequence induced conformation at the active site to allow protein–DNA–metal complex formation to allow recognition and cleavage.

The analysis of the different crystal structures indicates that to recognize and then cleave a target sequence, the conformation adopted by the DNA in the 2NN region to form the protein–DNA–metal complex, which is not mediated by any specific protein–DNA interaction, is important. Therefore, the 2NN region would control the active site rearrangement, permitting binding by allowing cation positioning and then cleavage once all the elements are properly positioned. Consequently changes in DNA sequence in this region affect the dinucleotide steps and thus the precise conformation to cut the target DNA is not allowed. Our hypothesis is supported by the MD simulations, which shows that the TA dinucleotide can sample larger regions of the conformational space resulting in high values of solvent accessible surface, thus providing more room to obtain the conformation proficient in target binding and DNA cleavage (Figure 5).

Our results show that the specificity of I-CreI, a widely redesigned enzyme with therapeutic and biotechnological purposes, depends on indirect readout to recognize and cleave its target sequence. So far the redesign of the protein–DNA binding properties in this scaffold has been performed taking into account the specific protein–DNA contacts. The understanding of the 2NN region role in target recognition and cleavage has a strong impact in meganuclease engineering targeting new DNA sequences that would avoid the presence of the non-preferred bases in the central region, thus optimizing meganuclease tailoring. The fact that I-CreI variants can promote homologous recombination inducing DSBs at positions relatively far away from the DNA region that should be corrected (43,44), enables us to find a suitable target to engineer meganuclease variants with improved functional properties.

ACCESSION NUMBERS

The coordinates and structure factors have been deposited in the PDB (4AAB, 4AAD, 4AAE, 4AAF, 4AAG).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–4 and Supplementary Figures 1–5.

ACKNOWLEDGEMENTS

We thank the Swiss Light Source and the European Synchrotron Radiation Facility beamline staff for their support. We thank Dr S. Ramón-Maiques and Dr D. Lietha for discussion and helpful comments.

FUNDING

Ministerio de Ciencia e Innovación [JCI-2011-09308 to R.M., BFU2008-01344/BMC, BFU2011-23815/BMC and CSD2006-20642 to G.M.]; Comunidad Autónoma de Madrid [CAM-P2006/Gen-0166 to G.M.]; EU Marie Curie 'SMARTBREAKER' [2010-276953, to S.S.] and Ministerio de Educación [SB2010-0105, to S.S.]. Funding for open access charge: Spanish Ministry of Science; Ministerio de Ciencia e Innovación.

Conflict of interest statement. None declared.

REFERENCES

1. Etzkorn,C. and Horton,N.C. (2004) Ca2+ binding in the active site of HincII: implications for the catalytic mechanism. *Biochemistry*, **43**, 13256–13270.
2. Etzkorn,C. and Horton,N.C. (2004) Mechanistic insights from the structures of HincII bound to cognate DNA cleaved from addition of Mg2+ and Mn2+. *J. Mol. Biol.*, **343**, 833–849.
3. Little,E.J. and Horton,N.C. (2005) DNA-induced conformational changes in type II restriction endonucleases: the structure of unliganded HincII. *J. Mol. Biol.*, **351**, 76–88.
4. Zahran,M., Daidone,I., Smith,J.C. and Imhof,P. (2010) Mechanism of DNA recognition by the restriction enzyme EcoRV. *J. Mol. Biol.*, **401**, 415–432.
5. Garvie,C.W. and Wolberger,C. (2001) Recognition of specific DNA sequences. *Mol. Cell*, **8**, 937–946.
6. Jurica,M.S., Monnat,R.J. Jr and Stoddard,B.L. (1998) DNA recognition and cleavage by the LAGLIDADG homing endonuclease I-CreI. *Mol. Cell*, **2**, 469–476.
7. Marcaida,M.J., Munoz,I.G., Blanco,F.J., Prieto,J. and Montoya,G. (2010) Homing endonucleases: from basics to therapeutic applications. *Cell. Mol. Life Sci.*, **67**, 727–748.
8. Marcaida,M.J., Prieto,J., Redondo,P., Nadra,A.D., Alibes,A., Serrano,L., Grizot,S., Duchateau,P., Paques,F., Blanco,F.J. *et al.* (2008) Crystal structure of I-DmoI in complex with its target DNA provides new insights into meganuclease engineering. *Proc. Natl Acad. Sci. USA*, **105**, 16888–16893.
9. Stoddard,B.L. (2005) Homing endonuclease structure and function. *Q. Rev. Biophys.*, **38**, 49–95.
10. Chevalier,B., Sussman,D., Otis,C., Noel,A.J., Turmel,M., Lemieux,C., Stephens,K., Monnat,R.J. Jr and Stoddard,B.L. (2004) Metal-dependent DNA cleavage mechanism of the I-CreI LAGLIDADG homing endonuclease. *Biochemistry*, **43**, 14015–14026.
11. Pingoud,A. and Jeltsch,A. (2001) Structure and function of type II restriction endonucleases. *Nucleic Acids Res.*, **29**, 3705–3727.
12. Roberts,R.J. and Halford,S.E. (1993) *Type II Restriction Endonucleases*, 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
13. Arnould,S., Chames,P., Perez,C., Lacroix,E., Duclert,A., Epinat,J.C., Stricher,F., Petit,A.S., Patin,A., Guillier,S. *et al.* (2006) Engineering of large numbers of highly specific homing endonucleases that induce recombination on novel DNA targets. *J. Mol. Biol.*, **355**, 443–458.
14. Arnould,S., Perez,C., Cabaniols,J.P., Smith,J., Gouble,A., Grizot,S., Epinat,J.C., Duclert,A., Duchateau,P. and Paques,F. (2007) Engineered I-CreI derivatives cleaving sequences from the human XPC gene can induce highly efficient gene correction in mammalian cells. *J. Mol. Biol.*, **371**, 49–65.
15. Grizot,S., Duclert,A., Thomas,S., Duchateau,P. and Paques,F. (2011) Context dependence between subdomains in the DNA binding interface of the I-CreI homing endonuclease. *Nucleic Acids Res.*, **39**, 6124–6136.
16. Ulge,U.Y., Baker,D.A. and Monnat,R.J. Jr (2011) Comprehensive computational design of mCreI homing endonuclease cleavage specificity for genome engineering. *Nucleic Acids Res.*, **39**, 4330–4339.
17. Wang,J., Kim,H.H., Yuan,X. and Herrin,D.L. (1997) Purification, biochemical characterization and protein-DNA interactions of the I-CreI endonuclease produced in Escherichia coli. *Nucleic Acids Res.*, **25**, 3767–3776.
18. Prieto,J., Redondo,P., Padro,D., Arnould,S., Epinat,J.C., Paques,F., Blanco,F.J. and Montoya,G. (2007) The C-terminal loop of the homing endonuclease I-CreI is essential for site recognition, DNA binding and cleavage. *Nucleic Acids Res.*, **35**, 3262–3271.
19. Redondo,P., Prieto,J., Munoz,I.G., Alibes,A., Stricher,F., Serrano,L., Cabaniols,J.P., Daboussi,F., Arnould,S., Perez,C. *et al.* (2008) Molecular basis of xeroderma pigmentosum group C DNA recognition by engineered meganucleases. *Nature*, **456**, 107–111.
20. Hafner,M., Vianini,E., Albertoni,B., Marchetti,L., Grune,I., Gloeckner,C. and Famulok,M. (2008) Displacement of protein-bound aptamers with small molecules screened by fluorescence polarization. *Nat. Protoc.*, **3**, 579–587.
21. Lakowicz. (2006) *Principles of Fluorescence Spectroscopy*. Springer, New York.
22. Tetin,S.Y. and Hazlett,T.L. (2000) Optical spectroscopy in studies of antibody-hapten interactions. *Methods*, **20**, 341–361.
23. Roehrl,M.H., Wang,J.Y. and Wagner,G. (2004) A general framework for development and data analysis of competitive high-throughput screens for small-molecule inhibitors of protein-protein interactions by fluorescence polarization. *Biochemistry*, **43**, 16056–16066.
24. Kabsch,W. (2010) Xds. *Acta Crystallogr. D Biol. Crystallogr.*, **66**, 125–132.
25. Leslie,A.G.W. (1992) Recent changes to the MOSFLM package for processing film and image plate data. *Joint CCP4 + ESF-EAMCB Newsletter on Protein Crystallography*, 26.
26. Vagin,A. and Teplyakov,A. (2010) Molecular replacement with MOLREP. *Acta Crystallogr. D Biol. Crystallogr.*, **66**, 22–25.
27. McCoy,A.J., Grosse-Kunstleve,R.W., Adams,P.D., Winn,M.D., Storoni,L.C. and Read,R.J. (2007) Phaser crystallographic software. *J. Appl. Crystallogr.*, **40**, 658–674.
28. Jones,T.A., Zou,J.Y., Cowan,S.W. and Kjeldgaard,M. (1991) Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A*, **47(Pt 2)**, 110–119.
29. Emsley,P., Lohkamp,B., Scott,W.G. and Cowtan,K. (2010) Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.*, **66**, 486–501.
30. Adams,P.D., Afonine,P.V., Bunkoczi,G., Chen,V.B., Davis,I.W., Echols,N., Headd,J.J., Hung,L.W., Kapral,G.J., Grosse-Kunstleve,R.W. *et al.* (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.*, **66**, 213–221.
31. Lu,X.J. and Olson,W.K. (2008) 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat. Protoc.*, **3**, 1213–1227.
32. Hess,B., Kutzner,C., van der Spoel,D. and Lindahl,E. (2008) GROMACS 4: algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.*, **4**, 12.
33. Lindorff-Larsen,K., Piana,S., Palmo,K., Maragakis,P., Klepeis,J.L., Dror,R.O. and Shaw,D.E. (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins*, **78**, 1950–1958.
34. Bussi,G., Donadio,D. and Parrinello,M. (2007) Canonical sampling through velocity rescaling. *J. Chem. Phys.*, **126**, 014101.
35. Epinat,J.C., Arnould,S., Chames,P., Rochaix,P., Desfontaines,D., Puzin,C., Patin,A., Zanghellini,A., Paques,F. and Lacroix,E. (2003) A novel engineered meganuclease induces homologous recombination in yeast and mammalian cells. *Nucleic Acids Res.*, **31**, 2952–2962.
36. Johnson,R.C., Stella,S. and Heiss,J.K. (2008) *Bending and Compaction of DNA by Proteins*. RSC Press, Cambridge.
37. Chen,V.B., Arendall,W.B. III, Headd,J.J., Keedy,D.A., Immormino,R.M., Kapral,G.J., Murray,L.W., Richardson,J.S. and Richardson,D.C. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.*, **66**, 12–21.

38. Travers,A.A. (2004) The structural basis of DNA flexibility. *Philos. Transact. A Math. Phys. Eng. Sci.*, **362**, 1423–1438.
39. Mackerell,A.D. Jr and Nilsson,L. (2008) Molecular dynamics simulations of nucleic acid-protein complexes. *Curr. Opin. Struct. Biol.*, **18**, 194–199.
40. McCammon,J.A., Gelin,B.R. and Karplus,M. (1977) Dynamics of folded proteins. *Nature*, **267**, 585–590.
41. Orozco,M., Noy,A. and Perez,A. (2008) Recent advances in the study of nucleic acid flexibility by molecular dynamics. *Curr. Opin. Struct. Biol.*, **18**, 185–193.
42. Olson,W.K., Gorin,A.A., Lu,X.J., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl Acad. Sci. USA*, **95**, 11163–11168.
43. Grizot,S., Smith,J., Daboussi,F., Prieto,J., Redondo,P., Merino,N., Villate,M., Thomas,S., Lemaire,L., Montoya,G. *et al.* (2009) Efficient targeting of a SCID gene by an engineered single-chain homing endonuclease. *Nucleic Acids Res.*, **37**, 5405–5419.
44. Munoz,I.G., Prieto,J., Subramanian,S., Coloma,J., Redondo,P., Villate,M., Merino,N., Marenchino,M., D'Abramo,M., Gervasio,F.L. *et al.* (2011) Molecular basis of engineered meganuclease targeting of the endogenous human RAG1 locus. *Nucleic Acids Res.*, **39**, 729–743.