doi:10.1016/S0022-2836(03)00447-9 J. Mol. Biol. (2003) 329, 253–269

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Flexible DNA Target Site Recognition by Divergent Homing Endonuclease Isoschizomers *I-CreI* and I-*Mso*I

Brett Chevalier[1], Monique Turmel[2], Claude Lemieux[2] Raymond J. Monnat Jr.[3] and Barry L. Stoddard[1]*

[1]*Division of Basic Sciences Graduate Program in Molecular and Cellular Biology University of Washington and the Fred Hutchinson Cancer Research Center 1100 Fairview Avenue North A3-025, Seattle WA 98109 USA*

[2]*Centre de recherche sur la Fonction, la structure et l'Ingenierie des Proteines Pavillon Charles-Eugene Marchand, Universite Laval Quebec, Canada G1K 7P4*

[3]*Departments of Pathology and Genome Sciences Box 357705 University of Washington, Seattle WA 98195 USA*

**Corresponding author*

Homing endonucleases are highly specific catalysts of DNA strand breaks that induce the transposition of mobile intervening sequences containing the endonuclease open reading frame. These enzymes recognize long DNA targets while tolerating individual sequence polymorphisms within those sites. Sequences of the homing endonucleases themselves diversify to a great extent after founding intron invasion events, generating highly divergent enzymes that recognize similar target sequences. Here, we visualize the mechanism of flexible DNA recognition and the pattern of structural divergence displayed by two homing endonuclease isoschizomers. We determined structures of *I-CreI* bound to two DNA target sites that differ at eight of 22 base-pairs, and the structure of an isoschizomer, I-*Mso*I, bound to a nearly identical DNA target site. This study illustrates several principles governing promiscuous base-pair recognition by DNA-binding proteins, and demonstrates that the isoschizomers display strikingly different protein/DNA contacts. The structures allow us to determine the information content at individual positions in the binding site as a function of the distribution of direct and water-mediated contacts to nucleotide bases, and provide an evolutionary snapshot of endonucleases at an early stage of divergence in their target specificity.

Keywords: endonuclease; structure; isoschizomer; homing; specificity

Introduction

Homing is the process by which mobile intervening genetic sequences, either introns or inteins, are duplicated into cognate recipient alleles that lack such a sequence.[1–5] The process is induced by an endonuclease encoded by an open reading frame (ORF) harbored within the intervening sequence.[6] The endonuclease specifically recognizes a DNA target site corresponding to the intron insertion site, generates a DNA double-strand break, and induces cellular mechanisms to repair the break. If the intron-containing allele is used as a template for repair, the endonuclease ORF is duplicated into the target site and the homing cycle is complete. Transfer of mobile introns can be extremely efficient. It can occur between different subcellular compartments of unrelated organisms,[7] and sometimes allowing introns to overrun diverse lineages within entire biological families, as demonstrated for a mobile fungal intron found throughout angiosperm plants.[8]

Homing endonucleases are found in all biological super-kingdoms. On the basis of primary sequence homology, four homing enzyme families have been identified: the LAGLIDADG, GIY-YIG, HNH, and His-Cys Box endonucleases.[5] The largest family, LAGLIDADG, contains several hundred identified members, many of which have been shown to be functional endonucleases.[5,9] The conserved LAGLIDADG sequence motif, from which the family draws its name, was shown in initial crystallographic analyses to form the core of the structural interface between endonuclease domains or subunits, and to contribute conserved acidic residues to the enzyme active sites.[10,11] Endonucleases containing a single LAGLIDADG motif per polypeptide chain form homodimers that

Abbreviations used: ORF, open reading frame.

E-mail address of the corresponding author: bstoddar@fhcrc.org

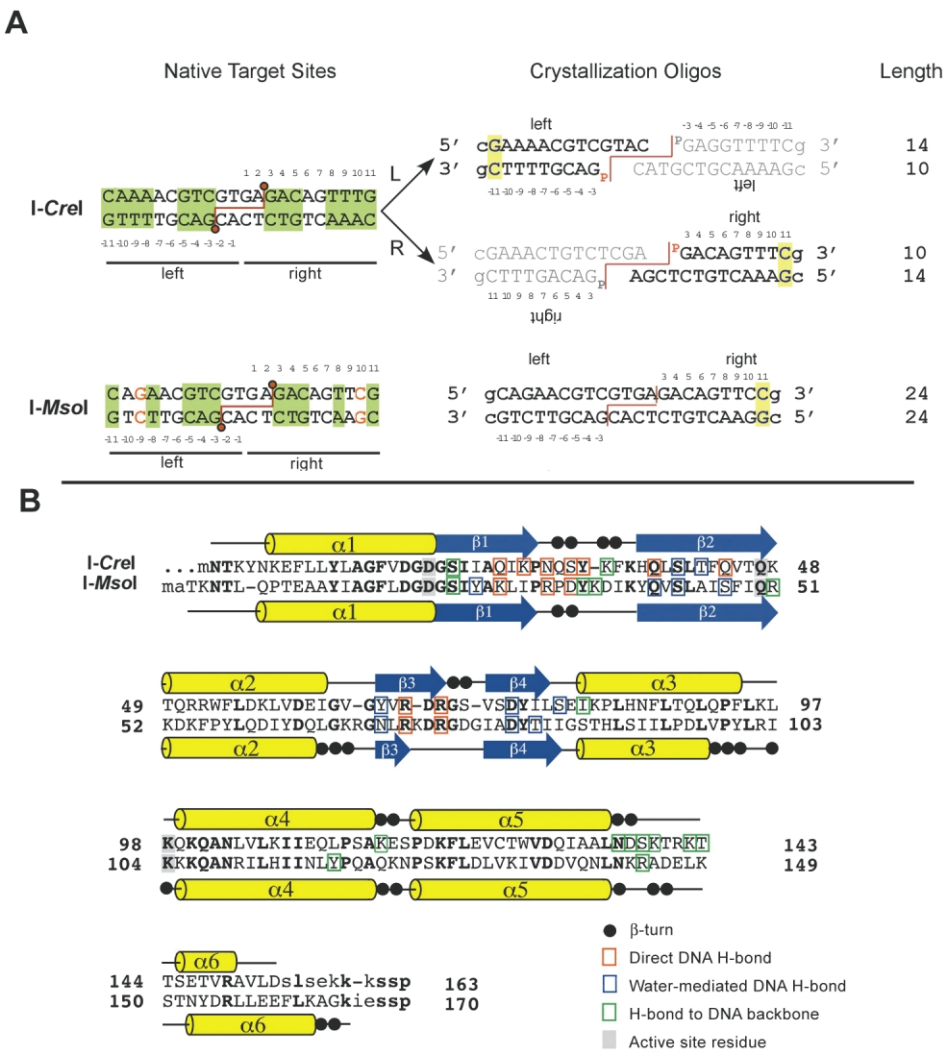

Figure 1. Physiological DNA target sites, crystallization constructs and structure-based sequence alignment of I-*Mso*I and *I-CreI*. (A) A comparison of the native DNA homing sites to the synthetic crystallization oligonucleotides. Each subunit recognizes one half of the 22 base-pair DNA target site, marked left and right, and the nuclease cleavage pattern is indicated by red lines. Palindromic base-pairs are shaded green and the two positions that differ between the I-*Mso*I and *I-CreI* targets are shown in red. *I-CreI* was crystallized with two different, pre-cleaved palindromic target sites, Cre left (CreL) and Cre right (CreR), representing perfect palindromes of either the left or right half target site. Each half-site was synthesized as a 14 base oligonucleotide and a complementary 5′ phosphorylated ten base oligonucleotide that annealed to form symmetric target sites with central, four base 3′ overhangs (black/light grey sites). I-*Mso*I was crystallized with the 24 base-pair blunt duplex shown. For both the *I-CreI* and I-*Mso*I crystallization oligonucleotides, bases flanking the 22 base-pair target site are lowercase and bases that differ from the wild-type homing site are blue with yellow shading. (B) Structure-based sequence alignment of *I-CreI* and I-*Mso*I. These proteins are 33% identical; conserved residues are bold. Lowercase letters show residues for which no density was visible in the crystal structures. A variety of residues that appear to align based on sequence alone (such as K28 in each ORF) are functionally and structurally distinct in the DNA cocrystal structures, leading to a revised alignment (for example, Q26 of I-*Cre*I is aligned more properly with K28 of I-*Mso*I).

recognize palindromic DNA target sites and pseudo-palindromic variants; members with two motifs fold to form pseudo-symmetric monomers capable of recognizing DNA target sites with significant asymmetry.[12]

Despite their relatively small size, all LAGLIDADG nucleases recognize long DNA target sites (14–30 bp), cleaving across the minor groove within the site to generate cohesive four base, 3′ overhangs.[13–16] The enzymes typically bind their physiological target sites with dissociation constants ranging from 1 to 5 nM, and exhibit chemical rates of DNA strand cleavage of approximately 0.3 min^{-1} (R.J.M., M.T. and C.L., unpublished results). These cleavage rates accelerated by at least six orders of magnitude over the rate of cleavage of non-specific DNA sequences. The rate of actual non-enzymatic cleavage of DNA under physiological conditions is much slower, to the point of not being measurable or directly comparable to the enzyme reaction. Of the intron-encoded LAGLIDADG homing endonucleases, the

I-CreI enzyme has perhaps been the best characterized in terms of recognition specificity and flexibility. The native DNA target site (or homing site) for the enzyme is a pseudopalindromic 22 bp site in which symmetry is broken at base-pairs ±1, 2, 6 and 7 between the target half-sites (Figure 1).[14,17] Palindromic variants of this site, consisting of inverted repeats of the left or right half-sites from the native target, are recognized and cleaved with similar affinities and activities (R.J.M. Jr, unpublished results). *In vitro* selection experiments, in which variant DNA target sites that can be cleaved by the wild-type enzyme were recovered and sequenced, indicate that most nucleotide positions in the site may be mutated to at least one alternative base-pair without loss of binding or cleavage sensitivity.[18] In some instances, several base-pairs may be altered simultaneously. The positions of polymorphisms that are tolerated by the enzyme most readily generally correspond to base-pairs that are not palindromically conserved between native half-sites, which in turn generally correspond to base-pairs that display fewer direct contacts with enzyme side-chains in the protein/DNA complex.[19] Many of the observed polymorphisms recovered in these studies increase the palindromic symmetry of the full-length site by converting a base-pair on one side to the corresponding sequence from the opposite side.

Estimates of overall site specificity for LAGLIDADG homing endonucleases such as *I-CreI* are difficult to calculate, but appear to approach 1 in 10^9 on the basis of the observed recovery of alternate viable substrates in cleavage-based selection experiments (i.e. at physiological concentrations the enzyme will recognize and cleave only one sequence out of any 10^9 random sequences of length equivalent to the target site, so that that it recognizes approximately one out of a billion random sequences of length 22 bp).[18] The balance of long substrate recognition coupled with tolerance of individual nucleotide polymorphisms may function to minimize host toxicity, while allowing recognition of a well-defined population of target sequences that are highly homologous to the original intron invasion site. This is likely a prerequisite for persistence of a mobile intron at a specific target site, because repeated horizontal transmission of introns appears to be necessary to maintain the introns in a rapidly diversifying and speciating host population.[20,21]

The persistence and diversification of homing endonucleases subsequent to a founding intron invasion event has been described elegantly for the *I-CreI* endonuclease family. *I-CreI* is encoded within a group I intron present in the chloroplast large subunit (LSU) rDNA of the green alga *Chlamydomonas reinhardtii*; the insertion site of this intron corresponds to position 2593 in the *Escherichia coli* 23 S rDNA. Sequence analysis of chloroplast and mitochondrial LSU rDNAs from numerous other green algae disclosed 15 new single-LAGLIDADG ORFs within identically positioned introns. Three of these genes were shown to encode active endonucleases that are isoschizomers of *I-CreI*, including I-*Mso*I from *Monomastix*.[22] Although the native target sites of *I-CreI* and I-*Mso*I differ at two out of 22 base-pair positions, each endonuclease cleaves both target sites efficiently. Threading the I-*Mso*I sequence onto the *I-CreI* structure suggests significant protein sequence divergence, especially at residues involved in DNA binding.[22] This divergence implies that homing endonucleases are under significant selective pressure to maintain similar cleavage specificities, but likely employ different arrangements of protein side-chains and collections of binding contacts to recognize nearly identical DNA target sites.

In this study, we use X-ray crystallography to determine the structures of three separate homing endonuclease/DNA complexes, in order to visualize directly the features of target site recognition and of divergent evolution as described above. In the first pair of experiments, the *I-CreI* endonuclease was cocrystallized with two separate palindromic variants of its pseudo-palindromic target site, allowing us to describe the similarities and differences in protein/DNA contacts and neighboring solvent positions for the base-pairs that differ between the two DNA targets. In a third crystallographic experiment, the structure of isoschizomeric I-*Mso*I was determined in complex with its native target site from *Monomastix* and compared to the structures of *I-CreI* described above. The results reveal the protein/DNA interfaces of *I-CreI* and I-*Mso*I to be even more divergent than anticipated on the basis of primary sequence alone, and provide a detailed visualization of the principles governing DNA site recognition for the LAGLIDADG endonuclease family.

Results

Recognition of degenerate DNA targets by *I-CreI*

The structures of four LAGLIDADG endonucleases in the absence of bound DNA have been described, including the group I intron-encoded *I-CreI*,[10] the intein-associated enzymes PI-*Pfu*I,[23] and PI-*Sce*I,[11] and the archaeal intron-encoded I-*Dmo*I.[24] *I-CreI* is a homodimer, whereas the latter three enzymes are pseudo-symmetric monomers with two domains that are each similar in topology to a single *I-CreI* subunit. The structure of PI-*Sce*I and *I-CreI* have been determined in complex with their physiological DNA target sites;[5,19,25] the former to 3.5 Å resolution and the latter to 1.9 Å resolution. In the high-resolution cocrystal structure of *I-CreI*,[5] the asymmetric DNA target site was bound in a 50/50 mixture of two orientations, leading to crystallographic averaging of side-chain conformations at non-palindromic base-pairs, as well as a mixture of water peaks simultaneously

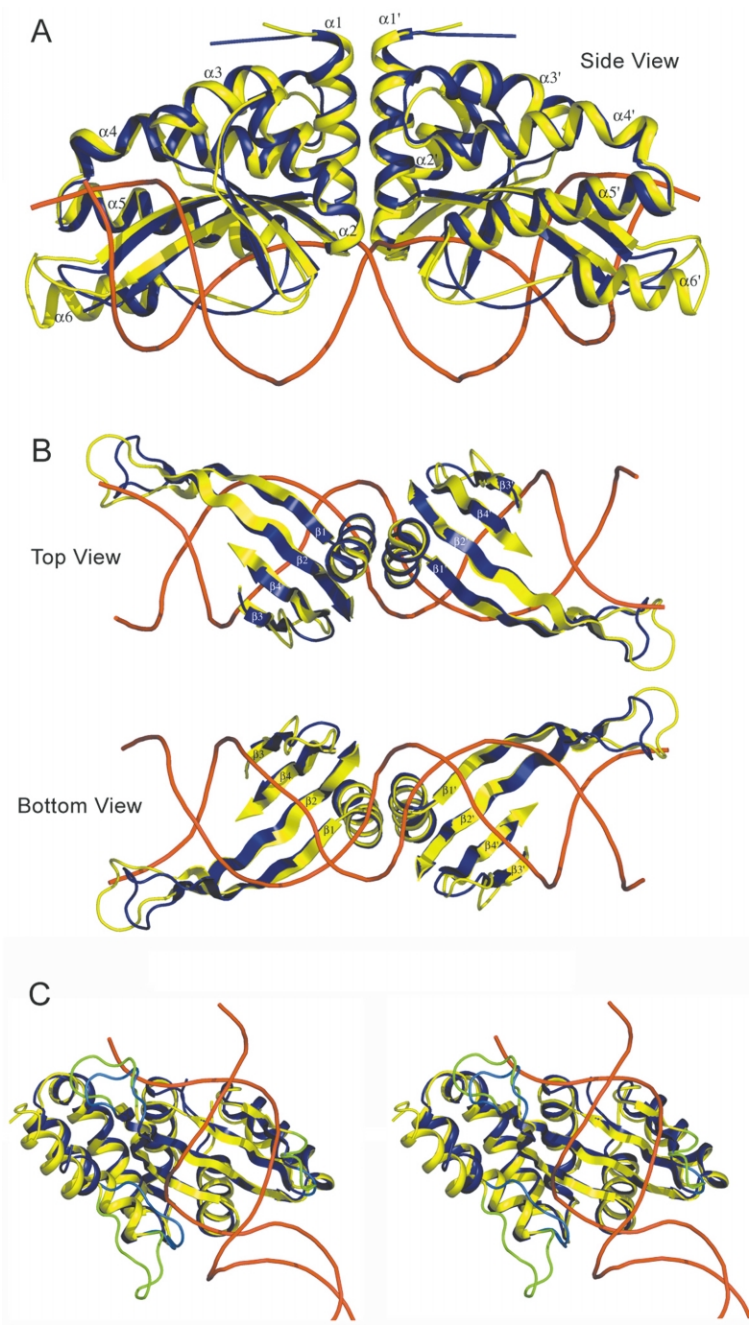

Figure 2. Superposition of the *I-CreI*:DNA and I-*Mso*I:DNA cocrystal structures. In this and all following Figures, *I-CreI* protein is blue, and I-*Mso*I protein is yellow For clarity, only the DNA from the I-*Mso*I structure is shown. (A) Side-view of the complete structures. Each protein is a homodimer. (B) Top and bottom views of the LAGLIDADG helices and the β-strands and associated loops that form the DNA-binding saddle. (C) Stereo-view of single subunits showing the relative positions of loops at the DNA interface. Loops in *I-CreI* are marine and loops in I-*Mso*I are green. The loop with the greatest change is between α5 and α6 of each protein. Figures 2–4 were made using PYMOL (W. DeLano (2002). The PYMOL molecular graphics system 0.83 edit. DeLano Scientific, San Carlos CA).

representing the solvation of both base-pairs. In this study, we determined cocrystal structures of *I-CreI* bound to DNA target sites using crystal forms that do not suffer from non-crystallographically averaged DNA orientations. This allowed us to visualize directly the unique interactions of protein, DNA and solvent molecules at individual nucleotides corresponding to non-palindromic target site base-pairs.

Each DNA target site for *I-CreI*, designed to avoid DNA hairpins, consisted of a 5′ phosphorylated ten base oligonucleotide and a complementary 14 base oligonucleotide that, when annealed, created a duplex with a four base, 3′ overhang characteristic of LAGLIDADG cleavage products (Figure 1(a)). The overhang was designed to anneal to a second identical DNA duplex, and thus form a precleaved, palindromic target site. These constructs were generated from the left and the right half-sites of the asymmetric *I-CreI* homing site. The protein/DNA complex with the palindromic repeat of the left *I-CreI* target half-site was designated CreL, whereas the complex with the palindromic repeat of the right half-site was termed CreR. Each blunt-ended DNA target site oligonucleotide was 24 bp long, a design that contains the target site flanked by one additional pair on each end. Throughout the text and in all Figures, base-pairs of the target site are numbered and denoted by ±1 to ±11 for each half-site, as shown in Figure 1.

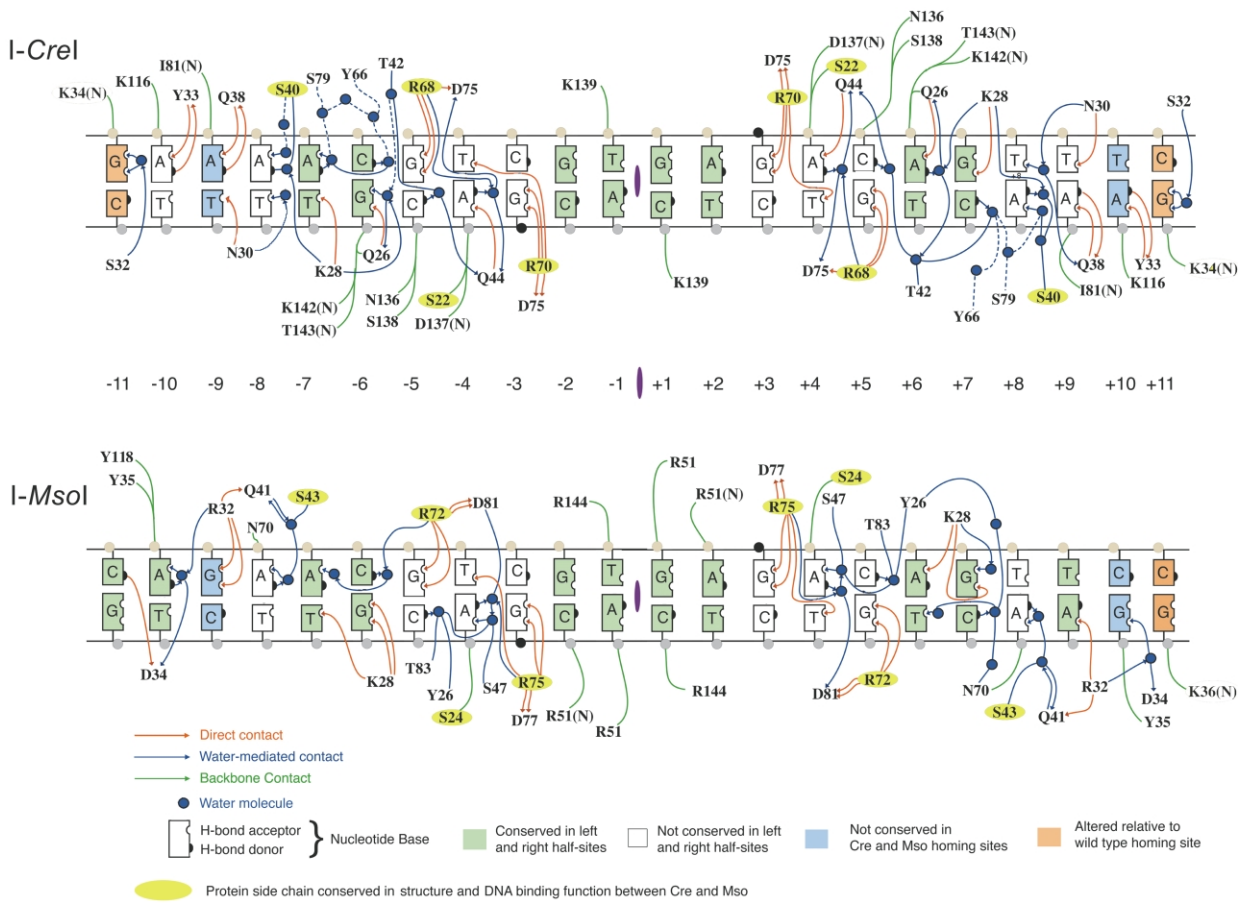

Figure 3. Summary of the DNA interfaces for *I-CreI* (top) and I-*Mso* I (bottom). Water molecules and water-mediated H-bonds are blue; direct protein:DNA H-bonds are red; H-bonds to the phosphate backbone are green. The scissile phosphate groups are black. Palindromic bases are white; non-palindromic bases are green. Bases that differ between the *I-CreI* and I-*Mso* I target sites are blue; bases that differ from those found in the native target sites are orange. The colors of the backbone circles (yellow *versus* grey) correlate to the colors of the bases in Figures 4–6. Hydrogen bond donors and acceptors on nucleotide bases are denoted by protruding and recessed ovals, respectively.

The DNA constructs in *Cre* L and *Cre* R differ at positions ±1, ±2, ±6 and ±7, corresponding to eight out of 22 (or 36%) of the base-pairs within the full-length complex (Figure 1(a)). The *Cre* L and *Cre* R complexes crystallized in different space groups ($P2_1$ and $P2_12_12_1$, respectively). Two complete *I-CreI*/DNA complexes are found in the asymmetric unit in both *Cre* L and *Cre* R, allowing redundant visualization of four *I-CreI*/DNA complexes for each equivalent DNA half-site. The protein conformations in *Cre* L and *Cre* R are essentially indistinguishable from previously reported structures. Except for the N-terminal methionine residue and eight C-terminal residues, all amino acid residues are readily visible in each structure. Three calcium ions are present within the active sites of all complexes in *Cre* L and *Cre* R, with one metal ion shared by the two enzyme subunits and their active sites, as previously reported.[5]

In the *I-CreI*/DNA interface, a set of four anti-parallel β-strands make direct and water-mediated contacts between residue side-chains and nucleotide atoms in the major groove of each half-site of both complexes, extending from base-pairs ±3 to ±11. Strands β1 and β2 extend the entire length of this interface in each half-site, while strands β3 and β4 provide additional contacts to base-pairs 3, 4 and 5 in each complex (Figure 2). The most striking characteristic of the interface, apart from its length, is that unlike restriction endonucleases,[26] potential hydrogen-bond (H-bond) contacts to individual nucleotide bases are remarkably under-saturated. *I-CreI* makes contacts appropriate for H-bonds to 26 of the 66 H-bond acceptors/donors within the major groove (13 out of 33 in each half-site). These direct contacts are made by eight equivalent protein side-chains in each half-site. A total of 28 water molecules (14 per half-site) mediate additional contacts between nucleotides and protein side-chains in the protein/DNA interface for *Cre* L; a subset of these (24 total, 12 per half-site) are visible in *Cre* R. Finally, 14 out of 44 phosphate groups in the DNA target site display direct contacts with protein side-chains. No additional contact is made to DNA bases within the minor groove. In sum, *I-CreI* utilizes 78% of possible major groove (and 47% of all major and minor groove) contacts available to its DNA target

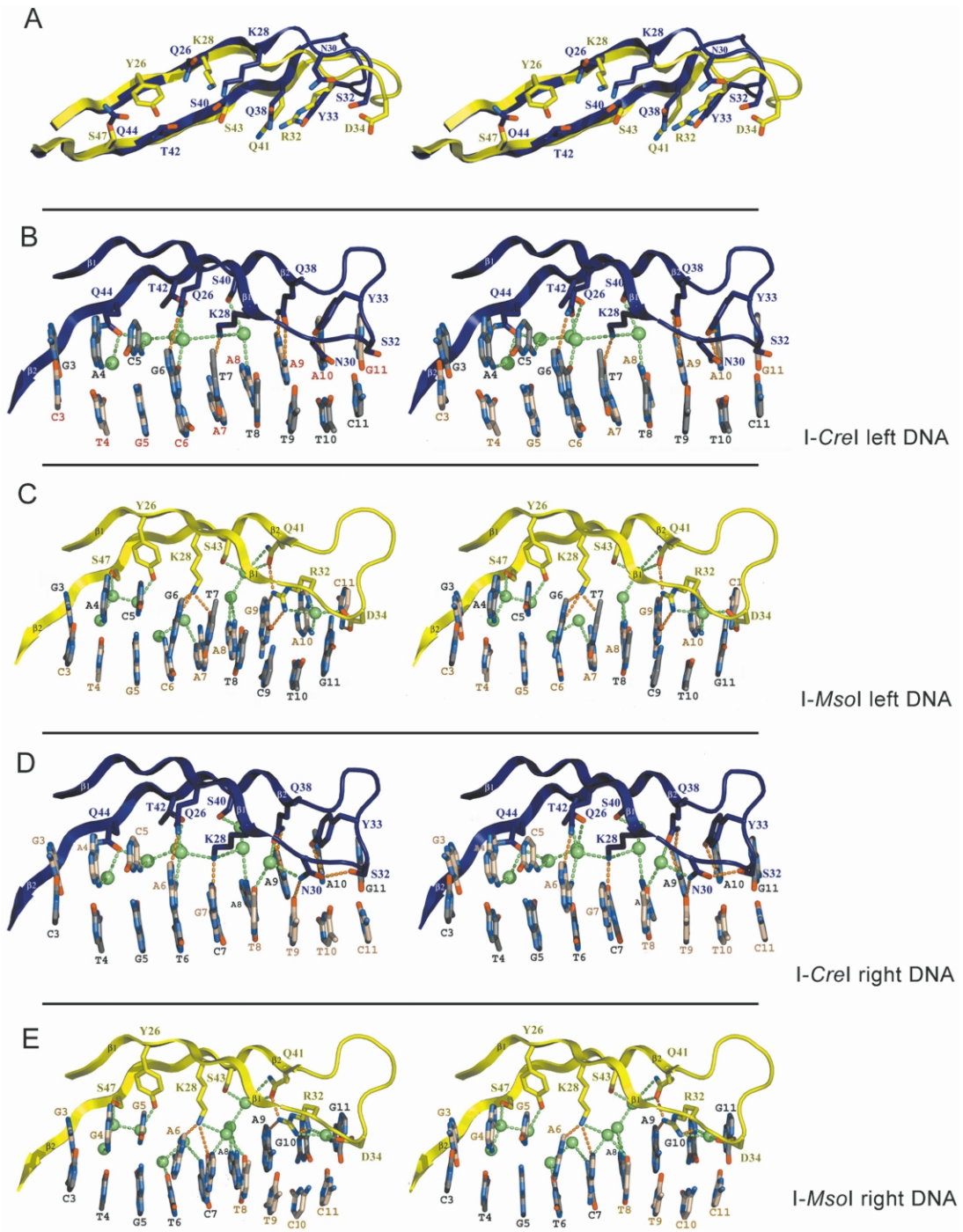

Figure 4. Stereo-views of the DNA interface formed by the extended antiparallel strands β1 and β2, which form contacts to base-pairs ±3–11. (A) Superposition of the I-*Mso*I and the *I-CreI* strands reveals little side-chain homology. (B)–(E) These strands extend over non-palindromic bases in their target sites; the interface between each protein and both half-sites is shown. Water molecules are shown as green spheres. Water-mediated H-bonds are represented by green dotted lines, direct H-bonds between side-chains and bases are orange. For clarity, the DNA backbone is omitted.

site; the contacts that are made are split evenly between direct and water-mediated contacts. A cartoon of all potential H-bond contacts for the physiological target site, representing a composite of *Cre*L and *Cre*R half-sites, is shown in Figure 3.

The extent of degeneracy of base-pair recognition at individual positions in the DNA target (i.e. whether different nucleotides can be recognized at those positions) was assayed previously using a site recovery screen;[18] many of the results of that study correlate well with the observed asymmetry of the native homing site for *I-CreI*, as well as with the total number of contacts made by the protein at individual base-pairs within the target site. Eight base-pairs (±1, 2, 6 and 7) are different between CreL and CreR, and display

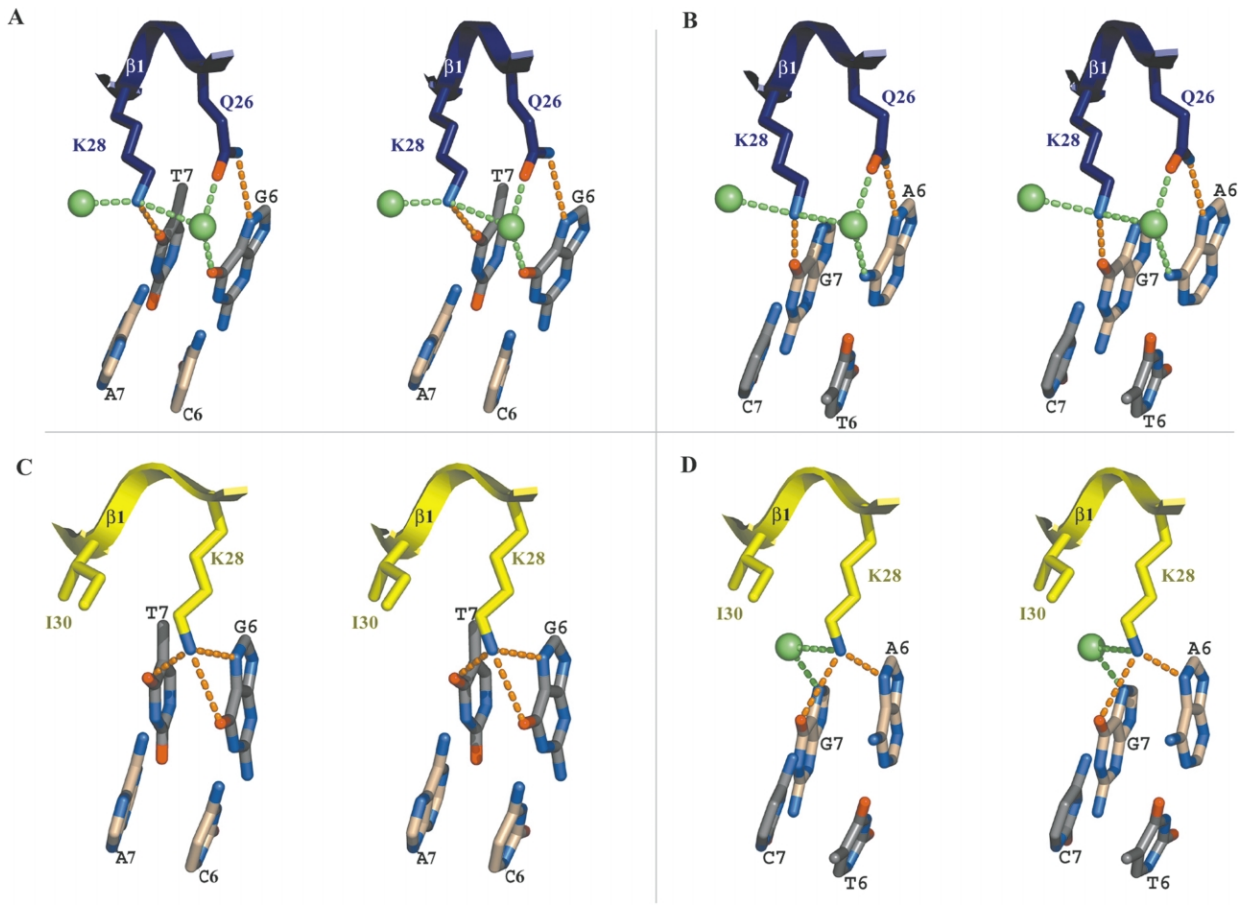

Figure 5. Stereo-views of contacts made by *I-CreI* and by I-*Mso*I to non-conserved base-pairs at positions ±6 and 7. Note that *I-CreI* uses a set of contacts that are structurally and chemically similar for the two different dinucleotide base-pair sequences, whereas I-*Mso*I uses a set of contacts in which a water molecule is used in place of a thymine methyl group.

significant degeneracy in site recovery experiments. Four of these base-pairs (±1 and 2) make no direct or water-mediated contact to protein side-chains and are located between between the cleaved phosphate groups in the substrate. The other four non-conserved base-pairs (±6 and ±7) exhibit one direct side-chain contact and one additional water-mediated contact. In contrast, the remaining base-pairs are palindromically conserved between CreL and CreR, and display significantly higher recognition specificity in site recovery experiments. Of these, two base-pairs (±3 and ±10) display direct contacts to two protein side-chain atoms; two (±4 and ±5) display direct contacts to two protein side-chain atoms and one additional water-mediated contact; one base-pair (± 9) displays three direct contacts to the protein, and one base-pair (±8) displays three water-mediated contacts. The correlation between the structures of the DNA complexes and the sequence specificity displayed by these enzymes is discussed in detail below.

The contacts to base-pairs in *Cre*L and *Cre*R are nearly identical, both at positions that are conserved, and at positions that differ between the two DNA target sites. Protein strands β1 and β2, comprising residues 21–48, form the most extensive part of the complex interface, extending from the central bases ±3 to ±11 (Figure 4(b) and (d)). From these strands, 16 direct contacts are made to the DNA target (eight to each half-site) in either complex. In each structure, Y33 and Q38 make identical bipartite contacts to conserved bases (to Ade ±10 and Ade ±9, respectively) and Q44 makes a single direct contact to Ade ±4. The contacts made to the non-palindromic bases at positions ±6 and ±7 are structurally conserved (Figure 5(a) and (b)). In the *Cre*L structure, Q26 contacts Gua ±6 and K28 contacts Thy ±7; in the *Cre*R structure, these same residues make similar contacts to Ade ±6 and Gua ±7. These contacts are appropriate for this particular nucleotide polymorphism: alternate base-pairs at either position would fail to present suitable H-bond acceptors to these same side-chain atoms. The positions and interactions of solvent molecules to these non-conserved base-pairs are virtually identical.

The only observable, and quite subtle, difference in the protein/DNA interface between *Cre*L and *Cre*R appears to be a contact made by N30 to Thy ±9. Though strong density is evident in *Cre*R for this residue, the corresponding density is not as

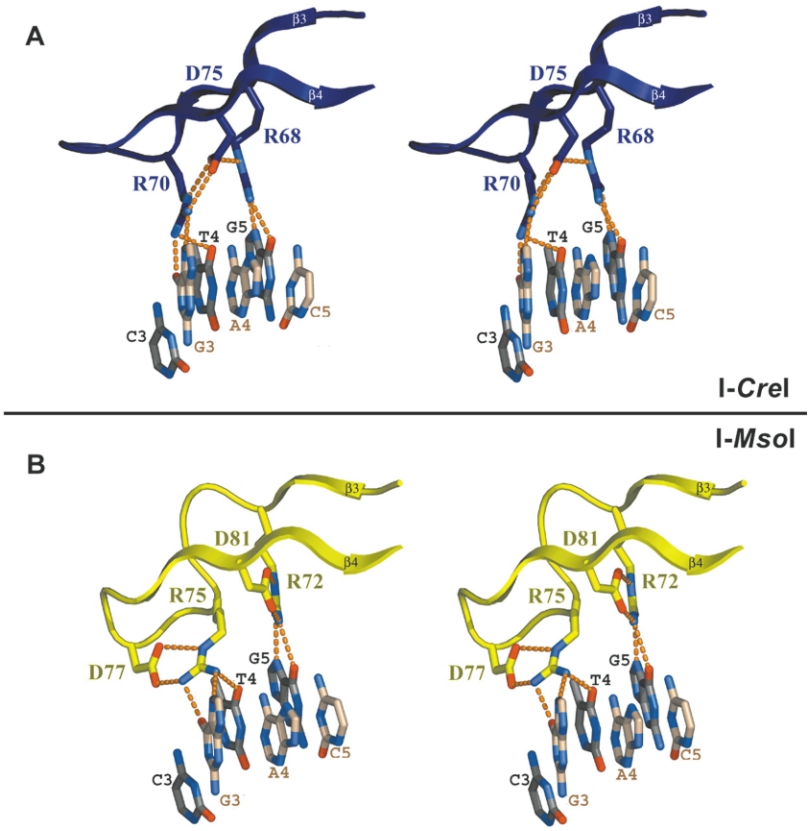

Figure 6. Stereo-views of the DNA interface formed by the β3 and β4 protein strands. *I-CreI* with DNA bases is on the top; I-*Mso* I with bases is on the bottom. The interface of only one half-site is shown, since these bases are palindromic and the observed interactions are identical in both half-sites. Direct H-bonds between side-chains and bases are orange. For clarity, the DNA backbone is omitted.

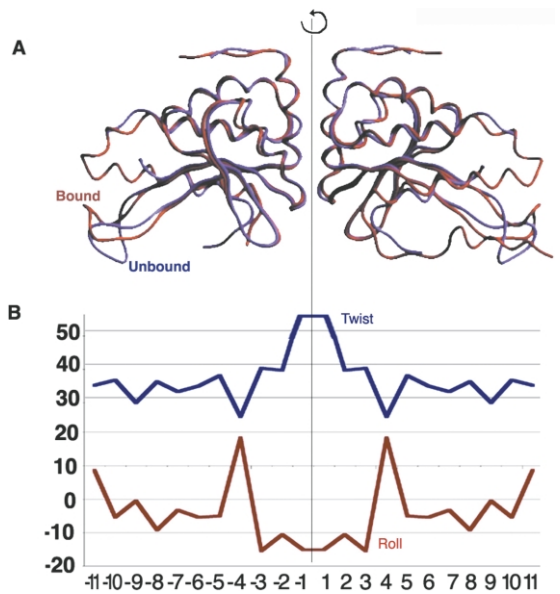

Figure 7. Conformational changes associated with I-CreI binding to DNA. (A) A superposition of bound and unbound protein backbones. The largest, and only significant, motions are confined to loops connecting β-strands 1 and 2 at the distal ends of each subunit and their DNA half-sites, and related loops preceding β-strand 3 in each subunit. (B) The local overwinding (twist) and resulting reduction in roll values for DNA base-pairs across the four base center of the cleavage target. These motions result in closure of the minor groove at the sites of cleavage and a shallow bend in the DNA around the enzyme surface. DNA bend parameters were calculated with program 3DNA.[47]

strong in *Cre* L and the residue appears to exhibit higher torsional mobility in that complex. This residue is located directly adjacent to the region of the protein (a loop connecting strands β1 and β2) that undergoes the largest conformational change upon DNA-binding.

Residues from the short strands β3 and β4 make additional contacts across the three most central palindromically conserved base-pairs (±3–5) (Figures 2, 3 and 6) and appear to greatly increase specificity at these base-pairs. Bipartite contacts are made by R68 to Gua ±5 and R70 to Gua ±3. R70 forms an additional bond with Thy ±4 (across from Ade ±4, contacted by Q44). Furthermore, Q44, R68 and D75 are all in H-bond distance to a single common water molecule that, in turn, makes a second specific contact to Ade ±4. This is the most obvious sequence-specific, water-mediated contact in the protein/DNA interface. Identical contacts are made by each of these residues to the palindromic bases in either half-site.

Binding of I-CreI to its target DNA involves significant conformational changes in both molecules (Figure 7). Superposition of the bound and unbound enzyme structures reveals differences primarily in two loop regions. Residues 29–37, which connect strands β1 and β2 in the protein/DNA interface, participate in DNA binding and adopt a significantly altered conformation. This loop donates three side-chains (Asn30, Ser32, and Tyr33) to contacts to nucleotides ± 10 and 11 at the ends of the homing site and provides a

distinctive twist of the β-ribbon that allows it to maintain contact over a long DNA site. A second loop preceding strand β3 does not place side-chains in contact with nuceotide base-pairs, but does optimize contacts to the phosphate backbone in the region of basepairs ±3. The target DNA is gradually bent around the endonuclease binding surface, giving an overall curvature across the entire length of the site of approximately 45°. The DNA is locally overwound between bases −3 to +3 (twist rising to ~50°), with a corresponding deformation in the base-pair propeller twist and buckle angles for those same bases. The bending of the DNA is symmetric, resulting from the combination of two identical kinks in the DNA caused by positive roll values of 18° located three bases distant from each cleavage site.[27] As observed for the DNA complex with the endonuclease domain of PI-SceI,[25] the central 4 bp region of the cleavage site displays negative roll values, which translate into a narrowing of the minor groove. As a result, the scissile phosphate groups are positioned 5 Å apart and are located near the conserved acidic residues and bound magnesium ions in the active sites.

Recognition of a similar DNA target by the divergent isoschizomer I-*Mso*I

The I-*Mso*I/DNA complex contains the native, asymmetric DNA target site crystallized in space group *P*1 with one complex per unit cell. The sequences of the I-*Mso*I DNA target half-sites used in his study are identical with the half-sites of *Cre*L and *Cre*R with the exception of G:C at −9 and G:C at +10 (which are A:T base-pairs in *Cre*L and *Cre*R, respectively). The uncleaved, 24 bp DNA duplex in the crystal structure contains the 22 bp I-*Mso*I target site flanked by one additional base-pair at either end (Figure 1(a)). Fortuitously, the orientation of the DNA in this structure was not averaged (i.e. the DNA was found in a single unique orientation). This allowed a direct comparison of unique nucleotide–protein–solvent contacts between half-sites for this enzyme to the contacts made in the *Cre*L and *Cre*R structures described above.

All but the first two and last five residues in each protein chain are visible in the structure of I-*Mso*I. Overall, the structures display 33% sequence identity[22] (Figure 1(b)) and an r.m.s.d. between all conserved atoms of 1.29 Å (Figure 2(a)). As observed for *I-CreI*,[5] three divalent calcium ions are present in the active sites and there is a similar enzyme-imposed 10° bend of the DNA substrate. The three most critical active-site residues found in *I-CreI* are conserved: D20, Q47 and K98 in *I-CreI* are D22, Q50, and K104 in I-*Mso*I. The DNA-binding saddle formed by the antiparallel β-sheets in I-*Mso*I superimposes well with the β-sheets of *I-CreI*, except for variation in the peripheral loops connecting β1 to β2 and β3 to β4 (Figure 2(b) and (c)). The greatest structural difference between the two proteins is in the loop connecting α5 to α6; in I-*Mso*I, this loop does not fold as tightly to the body of the protein and makes fewer contacts to the DNA backbone (Figure 2(c)). I-*Mso*I is seven residues longer than *I-CreI* (170 *versus* 163). Two of these additional residues are inserted near the N-terminal tail and one near the C-terminal tail. An additional residue is inserted into the loop between β1 and β2, and the remaining three are inserted into or near the sequence spanning β3 and β4 (Figure 1(b)). The latter strands fold into part of the protein/DNA interface and the insertions in this region disrupt the secondary structure as discussed below (Figures 1(b) and 2(b)).

The structures of *I-CreI* and I-*Mso*I revealed remarkable and unexpected differences in their DNA interfaces (Figures 3–6). Despite nearly identical DNA substrates, only five of the approximately 25 residues of each subunit of *I-CreI* contacting the DNA substrate are conserved in both identity and function in I-*Mso*I. The most striking differences are found throughout the β1 and β2 strands that extend over 8 bp of each half-site (Figure 4). Except for S22 and S40 in *I-CreI* relative to S24 and S43 in I-*Mso*I, respectively, which contact similar solvent and backbone atoms in the interface, no residue from these β-strands is conserved in structure or in binding function. For example, K28 of each protein appear to align in an initial sequence alignment created prior to the structure-determination of I-*Mso*I; however, these side-chains make different contacts to the DNA: K28 in I-*Mso*I is actually functionally homologous to Q26 in *I-CreI* (Figures 3–5). Similarly, while the I-*Mso*I residue Q41 is equivalent to Q38 in *I-CreI* and shares a nearly identical C^{α} position, their side-chains make entirely different contacts to the DNA. Q38 in *I-CreI* makes a pair of direct bipartite bonds to Ade ±9, whereas Q41 in I-*Mso*I is not directly involved in DNA binding. Yet again, Y33 in *I-CreI* makes a bipartite contact to Ade ±10, while the homologous residue in I-*Mso*I, Y35, contacts a nearby backbone phosphate group. Thus, these enzymes bound to their respective DNA target sites reveal a much greater degree of structural divergence at the DNA interface than was predicted from primary sequence alignments and site similarities.

The direct protein/DNA contacts are more conserved between the final two strands, β3 and β4, that make up the rest of the DNA-binding saddle for both enzymes (Figure 6). This region includes three conserved residues: R68, R70 and D75 in *I-CreI* are R72, R75, and D81 in I-*Mso*I. In each protein, both arginine residues make highly specific bipartite bonds to conserved guanine bases. The conservation of contacts for these residues is surprising, as this region of the protein backbone displays the greatest amount of structural divergence between the two enzymes. I-*Mso*I has two additional residues inserted in this region, one of which is between R72 and R75, that severely

disrupt the local secondary structure. Indeed, this region in *I-CreI* forms a tight strand-turn-strand topology with both arginine residues positioned inside β3. This tight turn is lost in I-*Mso*I, and R75 reaches the DNA interface from an extended loop. Furthermore, the double H-bond made by D75 to R70 (second arginine residue) in *I-CreI* is instead made by D81 to R72 (first arginine residue) in I-*Mso*I. A second aspartate residue, D77, is present in the loop that makes a similar contact to R75.

Recognition of degenerate DNA target sequences by I-*Mso*I

As observed for *I-CreI*, the I-*Mso*I/DNA interface is undersaturated; far fewer base-specific contacts are made by the protein to its DNA substrate than are actually possible given the 66 unique, base-identifying H-bond donors and acceptors present in the major groove. Again, all contacts to bases are made within the major groove, at positions that flank the central cleavage sites of the DNA substrate. Direct protein/DNA contacts are made to fewer DNA bases (11 contacts in the left half-site, eight contacts in the right) than is observed for *I-CreI*; only half of these contacts are made to the same nucleotide atoms as in the *I-CreI* complex. In addition, nine water-mediated contacts are made to each half-site; these are quite different contacts than those employed by *I-CreI*. In total, I-*Mso*I exploits only 34% of the potential H-bond donors and acceptors found in its DNA substrate major groove, even lower than the 47% seen in the *I-CreI*/DNA complex.

Degeneracy in base-pair recognition by I-*Mso*I at individual positions in its DNA target can again be correlated to the total number of H-bond contacts made by the protein at those positions, although not as simply as for *I-CreI*. Within the 22 bp target site, ten positions (±3, 4 5, 8 and 11) are identical in the left and right half-sites; the remainder of the positions (±1, 2 6, 7, 9 and 10) are not palindromically conserved (Figure 3). As in *I-CreI*, the central conserved base-pairs (±3, 4 and5) are contacted by residues from all four β-strands of the DNA-binding sheets. Two of these positions (±4 and 5) exhibit a complete complement of direct and water-mediated contacts across their potential H-bond donors and acceptors, while base-pair ±3 exhibits two direct, bifurcated H-bonds to a single arginine residue. Of the non-conserved base-pairs, two (±1 and 2) exhibit no direct or water-mediated contacts, one (±10) makes a single water-mediated contact, and three (±6, 7 and 9) make a single direct contact and up to two additional water-mediated contacts.

The structure and geometry of the protein and solvent contacts to base-pairs in the left and right half-sites of I-*Mso*I are nearly identical at many positions, as described above for *I-CreI*. In the I-*Mso*I/DNA complex, however, degeneracy at base-pairs 6 and 7 in the two half-sites is facilitated by significant differences in solvation and side-chain contacts (Figure 5(b)). Degenerate recognition of Gua 6 and Thy 7 in the left site *versus* Ade 6 and Gua 7 in the right site is accomplished in both half-sites by multiple contacts between K28 and the two base-pairs. However, a single direct contact between K28 and Thy 7 of the left half-site is replaced by a single direct contact plus a water-mediated contact to Gua 7 in the right half-site. In addition, two direct contacts to Gua 6 in the left half-site are reduced to a single direct contact to Ade 6 in the right half-site. These contacts provide the clearest example in this study of degenerate sequence recognition corresponding to alterations in side-chain contacts and differences in the presence and position of solvent molecules at non-identical base-pairs.

Discussion

As catalysts of the genetic transposition of mobile introns and inteins, homing endonucleases need to balance two seemingly contradictory requirements: they need to be highly sequence-specific in order to promote precise intron transfer and avoid deleterious cleavage of their host genomes, yet must retain sufficient site recognition flexibility to allow successful lateral transfer in the face of sequence variation in genetically diverging hosts. The structures described in this study provide new insight into how homing endonucleases use a flexible site-recognition strategy, in which a well defined, limited number of individual polymorphisms are tolerated by the enzyme without significant loss of binding affinity, to ensure that both requirements are met efficiently. The basis of this strategy is to make undersaturating contacts across long DNA target sites. The length of the interface provides overall high specificity, while formation of a broadly distributed set of phased, sub-saturating contacts across the interface facilitates the recognition and accommodation of specific polymorphisms at individual target site positions.

Recognition of multiple DNA targets by a single homing endonuclease

The determination of high-resolution structures of *I-CreI*/DNA complexes, with the explicit visualization of direct and water mediated contacts for each potential H-bond donor and acceptor throughout the target site, allows us to correlate the relationship between the number and type of inter-molecular contacts made to each base-pair to the calculated information content (i.e. the specificity of recognition) exhibited by the bound endonuclease at each of these positions. Three general conclusions from this analysis, discussed below, are (i) that the specificity of base-pair recognition to structurally unperturbed DNA sequence appears to be proportional to the number of H-bond contacts to each base-pair; (ii) that the

degree of specificity is not attenuated significantly by the use of solvent molecules as direct bridges between nucleotide atoms and protein side-chains; and (iii) information content (sequence specificity) is increased at individual base-pairs, particularly near the center of the cleavage site, by indirect recognition of DNA conformational preferences.

The distribution of related target site sequences that are recognized and cleaved by *I-CreI* has been described using a site preference screen.[18] In those experiments, variant target site variants that were cleaved by the native enzyme were recovered from a randomized homing site library. Using these data, the information content at each base-pair of the target site can be calculated using a computational method that accounts for the probability of each possible base being found at each position across the site, relative to the background base content expected at each position on the basis of genome composition or library design.[28] The distribution and recovery of sequence variations or polymorphisms among DNA sites recognized by a single protein allows us to calculate the specificity of the enzyme towards each individual base-pair. This form of analysis, which can be directed against collections of either naturally occurring sites that are recognized by a single protein, or sites recovered from a library screen, is the accepted standard for describing site degeneracy for a DNA-binding protein and for quantifying the relative specificity observed at individual base-pairs at each position in those sites. The information content or specificity exhibited towards base-pairs within the site are quantified as "bits" representing the probability of a specific base-pair at any given position. The units of bits as a measure of specificity or information content corresponds to the units often shown on sequence logo figures throughout the literature for aligned, homologous sites or of aligned homologous protein signatures. Information content at individual base-pairs in a collection of DNA recognition sites range from 0 bits, corresponding to complete degeneracy (25% probability of any base at a position) to two bits, corresponding to no degeneracy (100% chance of a unique base at that position).

The results of an information content analysis for I-*CreI* (Figure 8) demonstrate several important features of site-recognition by that enzyme that are presumably generalizable to the LAGLIDADG enzyme family. Information content (sequence specificity) varies dramatically across the target site (Figure 8(a)). No single position in the site is determined absolutely, and no single position is devoid of information content. With the exception of base-pairs ± 1 and ± 3 (discussed below), plots of the information content *versus* the total number of contacts made at each base-pair (Figure 8(b)) give reasonable linear correlations of 0.84, 0.89 and 0.86, corresponding to water-mediated contacts being counted as one-half, two-thirds or as full equivalents of direct protein/DNA contacts, respectively. When only direct protein/DNA contacts are counted for each base-pair (i.e. water-mediated contacts are not counted), the correlation is much lower ($R^2 = 0.57$). When analyzed together, the structural and genetic analyses summarized above imply that water-mediated contacts impart nearly the same information content to DNA target site recognition as direct protein/DNA contacts. However, sequence degeneracy can be facilitated at some nucleotide positions by the ability of water to reposition and/or reorient its dipole in response to sequence changes. This role for bound water is visualized directly at base-pairs $\pm$ 6 and 7 in the I-MsoI/DNA complex (Figure 5).

The observation that water-mediated contacts contribute to information content in protein/DNA binding to a degree that roughly matches the contribution of direct contacts is in agreement with a variety of published studies, especially those describing site recognition by the Trp repressor and by Hox transcriptional regulators. Crystallographic analysis of the Trp repressor bound to its DNA operator sequence[29] showed that sequence recognition is accomplished entirely through a combination of water-mediated contacts to nucleotide bases and indirect readout of DNA geometry. This observation was subsequently validated by mutagenic[30] and thermodynamic[31] analyses of protein binding. Similarly, a crystallographic analysis[32] and subsequent mutagenesis study[33] of a paired class homeodomain to its DNA site indicated that sequence-specific DNA recognition is driven largely by water-mediated contacts.

As mentioned above, base-pairs ± 1 and ± 3 both deviate from the linear correlation described above for the rest of the site (Figure 8(b)); in both cases these positions display twice the information content predicted by intermolecular contacts alone. One simple explanation for this is that these bases are found in the region of maximum DNA bending, base twisting and unstacking in the complex near the scissile phosphate groups (Figure 7). They may thus display conformational preferences that contribute favorably to binding energy and could therefore be recognized indirectly as an additional source of sequence specificity.

The overall information content of the entire 22 bp *I-CreI* target site, consisting of a summation of individual contributions to specificity for each base-pair position (assuming independent recognition of individual base-pairs) is approximately 24 bits, corresponding to an average of 1.09 bits per base-pair. Information content is spread unevenly over the target site, with no one base-pair displaying absolute sequence specificity for binding (Figure 8). In contrast, the *Hin*cII restriction endonuclease, which displays strict recognition of sequences (G-T-(T/C)-(A/G)-A-C), has a total information content of approximately 11, corresponding to an average of 1.83 bits per base-pair, and a predicted frequency of recognition every 1024 bases in a random genome sequence.[28] The greater specificity per base-pair for a restriction

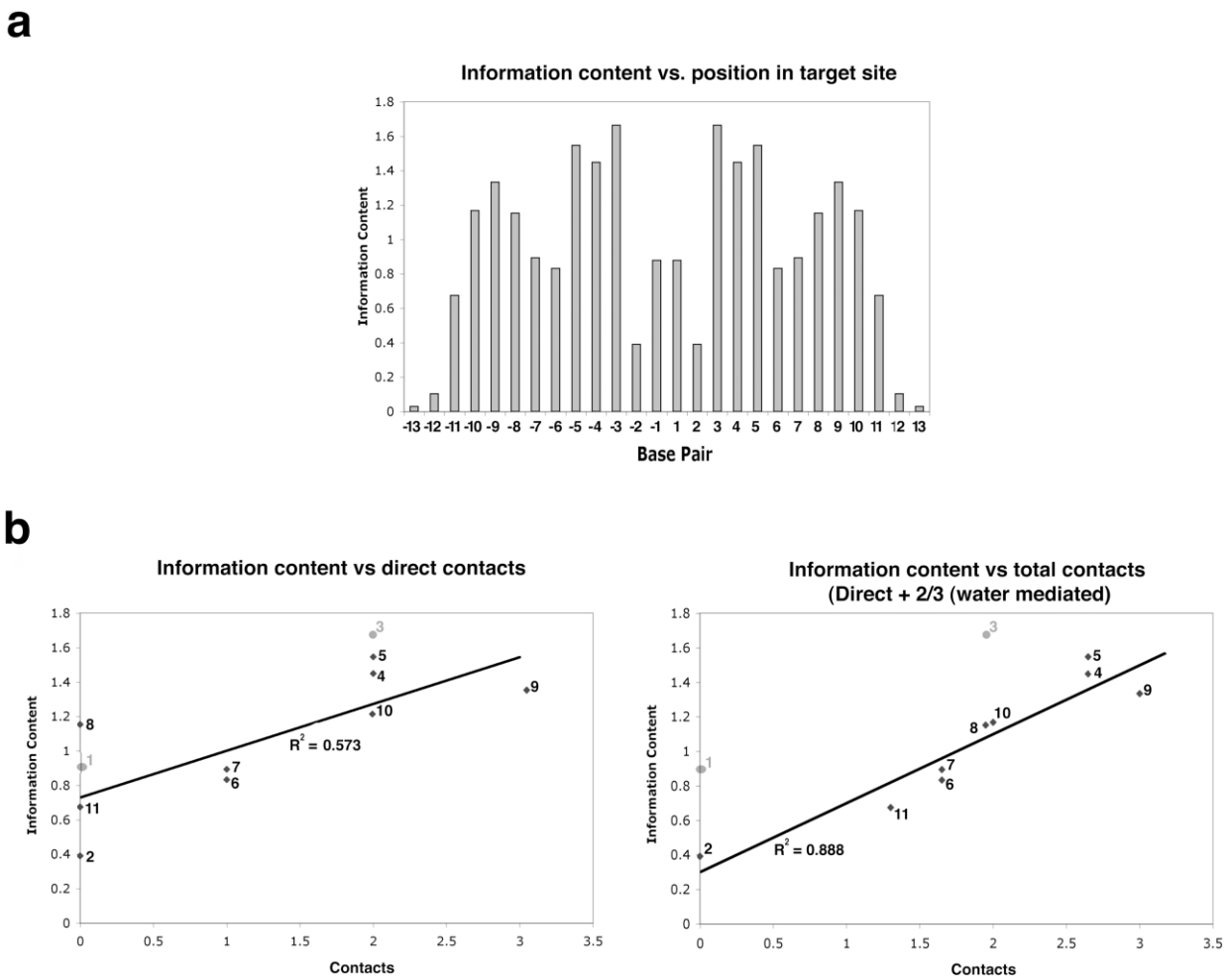

Figure 8. Information content in the *I-CreI*/DNA complex. (a) Information content calculated at individual base-pair positions across the target site. The positions with highest information content (±3, 4, 5, 8, 9 and 10) correspond to positions that are palindromically conserved in the native I-*Cre*I homing site. The average information content is 1.09 bits per base-pair. Information content falls off to background beyond positions ±11. (b) Information content for individual positions in the *I-CreI* target half-site plotted against the number of contacts made to each base-pair. The linear correlation coefficient is calculated for bases 4–11 (filled diamonds), corresponding to positions outside the active sites and the region of significant DNA distortion. The plot and its linear correlation is shown for direct protein contacts only (left; $R^2 = 0.57$) and for total direct and water-mediated contacts, with water-mediated contacts counted as two-thirds of a direct contact (right; $R^2 = 0.89$). The correlation is slightly lower if water-mediated contacts are counted as either half of a direct contact ($R^2 = 0.86$) or as the equivalent of a full contact ($R^2 = 0.84$). Note that base-pairs 1 and 3 (grey ovals), which are located near the position of maximal DNA distortion and cleavage, display higher information content than expected solely on the basis of sequence-specific contact only (possibly indicating indirect readout of conformational preferences at those positions), but that the overall trend is similar to bases outside this region.

enzyme is correlated directly to an extraordinary number of contacts between each target site nucleotide and the protein. For example, the *Mun*I enzyme, which recognizes a six base target site, forms 22 contacts to nucleotide bases in the major and minor groove, and 18 contacts to the 14 phosphate groups across the site;[34] this corresponds to over double the number of contacts per base-pair formed by I-*Cre*I or I-*Mso*I.

The individual contacts between side-chains and nucleotide bases, in addition to exhibiting additive contributions to information content, can be correlated with their contribution to the overall free energy of binding and the resulting association constants for the complex. For example, individual DNA site mutations at position ±10 in the target site (corresponding to an inversion of an A:T base-pair that confounds the ability of the enzyme to make its usual side-chain contacts at those positions) causes an increase in K_D of 100-fold (5 nM–560 nM), corresponding to an approximate $\Delta\Delta G$ for binding of 3–4 kcal/mol (1 cal = 4.184 J).[35] In contrast, a similar base-pair inversion at base-pairs ±11 causes a smaller tenfold increase in K_D, corresponding to a more moderate $\Delta\Delta G$ for binding of 1–2 kcal/mol.[35] The relative magnitude of these effects are in agreement with the corresponding values of information content in the protein–DNA complex for these two base-pairs (1.15 and 0.65, respectively). Furthermore, the individual

polymorphisms corresponding to the sequence differences between creL and creR palindromic sites (at base-pairs ±1, 2, 6, and 7) do not cause measurable differences in target site affinity (R.J.M., unpublished results), because these mutations do not confound individual H-bond contacts with the protein, as described here.

The heterogeneous pattern of target site information content exhibited by *I-CreI* makes it difficult to calculate an absolute site recognition specificity, but upper and lower limits can be estimated for this value. The most liberal interpretation of the previously published target site degeneracy data[18] and the calculated information content for the site is that any number of base-pairs in the recognition sequence can be altered simultaneously and independently, with each position exhibiting an average informational content of 1.09 bits (Figure 8(a)). This model gives a predicted site specificity of 1 in 3×10^7 random sequences. However, the calculation described above systematically underestimates specificity by assuming that contacts made to individual base-pairs are simply additive and independent of one another, and do not influence the probability of additional site polymorphisms. The target site recovery data indicate that sequences with minimal numbers of polymorphisms relative to the native *I-CreI* target site are enriched selectively from the starting site library, and that many polymorphisms convert asymmetric bases in the left half-site to the corresponding residue in the right half-site. A calculation of site frequency that accounts for relative recoveries of individual sites, corrected for the nucleotide ratios in the starting library, provides a more conservative estimate of binding specificity of 1 site per 6×10^9 random sequences. The correlations between individual contacts (direct *versus* water-mediated) and information content described above are still observed regardless of the assumptions used to calculate site specificity.

Divergence of homing endonuclease isoschizomers

Endonucleases that recognize and cleave similar or identical DNA target sequences have been identified both for restriction enzymes and for homing enzymes. For the former, at least two pairs of DNA-cocrystal structure analyses have been performed between enzymes that recognize similar sites (*Mun*I and *Eco*RI;[34] *Bgl*II and *Bam*HI[36]). These studies show that *Mun*I and *Eco*RI use a conserved structural mechanism for the interaction with their common core AATT target but differ in the recognition of peripheral nucleotides, whereas, *Bgl*II and *Bam*HI display quite different protein/DNA contacts even at their common GATC core target sequence. An additional structural comparison of true isoschizomers *Bse*634I and *Cfr*10I[37] in the absence of bound DNA demonstrate interesting structural similarities at both the tertiary and quaternary structural levels; detailed comparisons await cocrystal structures for these and similar isoschizomer pairs.

The differences in the contacts displayed by I-*Cre*I and I-*Mso*I for the recognition of identical or very similar target sites are greater than predicted by comparative sequence analysis. This unexpected finding illustrates clearly the hazards in trying to model or predict DNA-binding interactions between divergent DNA-binding proteins solely on the basis of sequence data, even when the proteins retain nearly identical specificities. The diversity of protein side-chains and contacts employed by I-*Mso*I and I-*Cre*I to recognize extremely similar sites suggests that homing endonucleases are under dual pressures of diversification and preserved recognition of the endogenous target site. An analysis of the patterns of synonymous and non-synonymous substitutions[38] for the I-*Cre*I and I-*Mso*I genes and homologous site-2593 intron ORFs from green algal chloroplasts encoding putative homing endonucleases failed to reveal any coding region that is evolving under positive selection[39] (synonymous substitution rate was found to greatly exceed non-synonymous rate; data not shown), indicating that the structural diversification of target site recognition exhibited by these endonucleases reflects an remarkable ability of compensatory mutations to achieve recognition of similar target sites.

Analysis of the protein/DNA complexes and their contacts for I-*Cre*I and I-*Mso*I suggests that the divergence of homing endonucleases involves significant changes in information content and sequence recognition degeneracy. Although site recovery studies have not been carried out for I-*Mso*I, we predict that this enzyme is less sequence-specific than I-*Cre*I because of the smaller number of contacts made to the target site. This should, in turn, increase the chances that the intron encoding I-*Mso*I will successfully invade site 2593 of divergent rRNA genes if it encounters new hosts or gains access to a different genetic compartment within the same cell. The success rate of mobile introns in colonizing and persisting at a given genomic site within a community of diverse organisms is likely to depend on a combination of access to genomic sites and the level of sequence recognition degeneracy exhibited by the homologous homing endonucleases encoded by the introns at this site. The different strategies employed by homing endonucleases to recognize nearly identical target sites implies that not all enzymes that recognize a single target site, e.g. site 2593, will recognize and cleave the same variant target sites. However, the diversity of sites recognized by members of the same "site family" may be pivotal in ensuring the potential for dissemination of mobile introns to new genetic loci, and thus their persistence and spread in natural populations.[22]

Implication for engineering of homing endonucleases with novel specificities

Homing endonucleases (in particular the LAGLIDADG enzyme family) are highly sequence-specific, with DNA-binding domains that are highly modular, separable, and intimately associated with catalytic active sites. As such, these endonucleases appear to be highly attractive targets for the creation of gene-specific reagents, using a variety of tools for the engineering and selection of novel DNA-recognition activities. The importance of solvent-mediated contacts for DNA site recognition by these enzymes has clear implications for the engineering variants with altered specificity, in that any structure-based computational approaches to such a problem must incorporate an accurate strategy for the explicit modeling of water molecules in the protein/DNA interface. At this time, successful strategies have been reported both for the selection of homing endonuclease point mutants with minor alterations in site specificity[35] and for the recombination and fusion of independent LAGLIDADG domains to create artificial chimeric enzymes with drastically altered site-specificity.[40] Future attempts to engineer artificial homing endonucleases with completely novel sequence specificities will presumably exploit both of these types of strategies, but will be dependent also on the concerted randomization and selection of large numbers of residues in the DNA-binding interface as part of an overall redesign strategy. The results reported here seem to indicate that the accurate modeling and prediction of the structural and energetic features of direct and water-mediated protein/DNA contacts, in combination with continued development of powerful selection methods for enzyme/DNA specificity, will be critical for the success of such experiments.

In addition, the comparative analysis of binding by I-*Mso*I and *I-CreI* to similar DNA target sites implies that a large number of diverse side-chain and solvent packing Scheme may allow the recognition of a given DNA target site, and that variations within selected interfaces may be associated with dramatic differences in overall sequence specificity. It should be possible to use additional analyses of DNA-binding properties of LAGLIDADG homing endonucleases to clarify the rules governing sequence-specific DNA recognition and thus facilitate the design of novel, gene-specific proteins.

Materials and Methods

Protein expression and purification

The I-*Mso*I ORF was PCR amplified from pET-I-*Mso*I[22] and subcloned into the *Nde*I/*Xho*I sites of the p*I-CreI* vector[41] to make pI-*Mso*I. *I-CreI* and I-*Mso*I were purified in identical fashion. Briefly, protein was overexpressed by induction with 0.5 mM IPTG in BL21[DE3] *E. coli* cells overnight at 15 °C. Cells were harvested by centrifugation and lysed by sonication in 50 mM Tris (pH 8.0), 100 mM NaCl. Cell debris was removed by centrifugation at 40,000*g* for 45 minutes at 4 °C, then forced through a 0.2 μm syringe filter and applied to a heparin column (Pharmacia). Protein eluted with an increasing salt gradient as a single peak that was collected, diluted with an equal volume of 50 mM Tris (pH 8.0), and loaded onto the heparin column. After the second elution, protein was >95% pure as determined by SDS-PAGE analysis. The protein solution was dialyzed overnight against storage buffer (50 mM Tris (pH 8.0), 100 mM NaCl, 1 mM $CaCl_2$) and concentrated to ~4 mg/ml by centrifugation (Centriprep, Millipore), flash-frozen in liquid nitrogen and stored at −80 °C.

Crystallization

The design of DNA oligonucleotides for *I-CreI* and I-*Mso*I crystallizations is summarized in Results. Despite crystallizing in different space groups, all *I-CreI*/DNA crystals grew in previously described conditions (21–27% PEG 400, 100 mM Mes (pH 6.2–6.8), 10 mM $CaCl_2$, 20 mM NaCl, 22 °C) at ~4 mg/ml protein with a 1.2–1.7 molar excess of DNA target substrate. *Cre*R crystals grew readily in three to six days; *Cre*L crystals grew as thick needles in one to two weeks. Diffraction-quality I-*Mso*I/DNA crystals grew at 18 °C in 19–22% PEG 400, 100 mM Tris (pH 7.3–7.9), 10 mM $CaCl_2$, 20 mM NaCl at 3.5 mg/ml of protein with a 1.5–2.0 molar excess of DNA substrate. Crystals appeared within one hour, and diffraction quality crystals were obtained after two to six days.

Data collection

Crystals were removed directly from the drops in which they grew, suspended in a fiber loop, frozen in liquid nitrogen and maintained at 100 K during data collection. Data from crystals of *Cre*R and the I-*Mso*I/DNA complex were collected at the Advanced Light Source (beamline 5.0.2). Data from the crystals of *Cre*L were collected at the Advanced Photon Source (beamline BM-19) with $2\theta = 8°$. Data were reduced using the DENZO/SCALEPACK crystallographic data reduction package (Table 1).

The structures were solved *via* molecular replacement using EPMR† with the low-resolution *I-CreI*/DNA complex structures as an initial search model for *Cre*L and *Cre*R and a polyalanine *I-CreI* model (lacking loops and α6) for the I-*Mso*I structure. In both *Cre*L and *Cre*R, two *I-CreI*/DNA complexes were found in each asymmetric unit. Only one I-*Mso*I structure was found within its *P*1 cell. All structures were modeled in XtalView[42] and O[43], then refined using CNS[44] with 5% of the data set aside for cross-validation[45]. The final refinement statistics (Table 1) for *Cre*L were $R_{work}/R_{free} = 0.219/0.246$, for *Cre*R $R_{work}/R_{free} = 0.195/0.229$, and for the I-*Mso*I complex $R_{work}/R_{free} = 0.218/0.253$. Geometric analysis of the structures using PROCHECK[46] indicates no residue

† Kissinger, C. R. & Gehlhaar, D. K. (1997). EPMR: A program for crystallographic molecular replacement by evolutionary search. Agouron Pharmaceuticals, La Jolla, CA.

Table 1. Data and refinement statistics

A. *Data*			
Structure	I-*Cre*I (left; CreL)	I-*Cre*I (right; CreR)	I-*Mso*I
Source	APS	5.0.2 ALS	5.0.2 ALS
Resolution limit (Å)	2.50	2.00	2.25
Wavelength (Å)	1.00	1.10	1.10
Space group	$P2_12_12_1$	$P2_1$	$P1$
Unit cell parameters			
a, *b*, *c* (Å)	46.7, 68.4, 301.9	78.4, 76.4, 81.1	41.5, 42.2, 71.3
α, β, γ (deg.)	90, 90, 90	90, 108.8, 90	73.3, 73.2, 71.1
Measured reflections	77,280	193,008	38,547
Unique reflections	32,363	60,035	19,849
R_{merge}[a]	5.4 (32.4)	4.0 (20.8)	3.0 (7.6)
Competeness (%)[1]	93.5 (71.1)	99.1 (98.1)	97.5 (96.9)
B. *Refinement*			
R_{work} (%)	21.9	19.5	21.8
R_{free} (%)[b]	24.6	22.9	25.3
Resolution(Å)	50–2.50	50–2.00	50–2.25
No. atoms	7102	7331	3824
No. water molecules	233	513	229
r.m.s. deviations			
Bond length (Å)	0.008	0.006	0.006
Bond angles (deg.)	1.61	1.26	1.22
Ramachandran plot			
Favorable (%)	88.9	88.8	92.3
Allowed (%)	11.1	11.2	7.7
Generous (%)	0	0	0
Unfavorable (%)	0	0	0
Mean *B* value (Å^2)			
Overall	41.5	39.9	38.3
Protein	41.2	37.6	34.0
DNA	42.7	41.4	48.9
Solvent	38.9	42.3	40.6
Cations	31.0	38.1	53.7

[a] The numbers in parentheses are statistics from the highest-resolution shell.
[b] R_{free} was calculated with 5% of the data that was not used for calculation of R_{work}.

in any structure with generously allowed or unfavorable backbone dihedral angles.

Protein Data Bank accession codes

The I-*Mso*I/DNA (RCSB accession code 1M5X), *Cre*L (IN3E) and *Cre*R (1N3F) structural models and data have been deposited in the Protein Data Bank.

Acknowledgements

The work described here was funded by grants from the NIH to B.S. (GM49857) and to R.M. (CA88942), and by a grant from the Natural Sciences and Engineering Research Council of Canada (GP0002830) to C.L. and M.T. B.C. was supported by an Interdisciplinary Training Grant in Cancer Research pre-doctoral fellowship (T32 CA80416). We thank members of the FHCRC structural biology program, particularly Adrian Ferre D'Amare, for helpful criticisms and advice.

References

1. Dujon, B. (1989). Group I introns as mobile genetic elements: facts and mechanistic speculations—a review. *Gene*, **82**, 91–114.
2. Lambowitz, A. M. & Belfort, M. (1993). Introns as mobile genetic elements. *Annu. Rev. Biochem.* **62**, 587–622.
3. Belfort, M. & Perlman, P. S. (1995). Mechanisms of intron mobility. *J. Biol. Chem.* **270**, 30237–30240.
4. Belfort, M. & Roberts, R. J. (1997). Homing endonucleases—keeping the house in order. *Nucl. Acids Res.* **25**, 3379–3388.
5. Chevalier, B., Monnatt, R. J. & Stoddard, B. L. (2001). The LAGLIDADG homing endonuclease I-CreI shares three divalent cations between two active sites. *Nature Struct. Biol.* **8**, 312–316.
6. Jacquier, A. & Dujon, B. (1985). An intron-encoded protein is active in a gene conversion process that spreads an intron into a mitochondrial gene. *Cell*, **41**, 383–394.
7. Turmel, M., Cote, V., Otis, C., Mercier, J. P., Gray, M. W., Lonergan, K. M. & Lemieux, C. (1995). Evolutionary transfer of ORF-containing group I introns between different subcellular compartments (chloroplast and mitochondrion). *Mol. Biol. Evol.* **12**, 533–545.
8. Cho, Y., Qiu, Y.-L., Kuhlman, P. & Palmer, J. D.

(1998). Explosive invasion of plant mitochondria by a group I intron. *Proc. Natl Acad. Sci. USA*, **95**, 14244–14249.
9. Dalgaard, J. Z., Klar, A. J., Moser, M. J., Holley, W. R., Chatterjee, A. & Mian, I. S. (1997). Statistical modeling and analysis of the LAGLIDADG family of site-specific endonucleases and identification of an intein that encodes a site-specific endonuclease of the HNH family. *Nucl. Acids Res.* **25**, 4626–4638.
10. Heath, P. J., Stephens, K. M., Monnat, R. J. & Stoddard, B. L. (1997). The structure of *I-CreI*, a group I intron-encoded homing endonuclease. *Nature Struct. Biol.* **4**, 468–476.
11. Duan, X., Gimble, F. S. & Quiocho, F. A. (1997). Crystal structure of PI-SceI, a homing endonuclease with protein splicing activity. *Cell*, **89**, 555–564.
12. Chevalier, B. S. & Stoddard, B. L. (2001). Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility. *Nucl. Acids Res.* **29**, 3757–3774.
13. Colleaux, L., D'Auriol, L., Galibert, F. & Dujon, B. (1988). Recognition and cleavage site of the intron-encoded omega transposase. *Proc. Natl Acad. Sci. USA*, **85**, 6022–6026.
14. Thompson, A. J., Yuan, X., Kudlicki, W. & Herrin, D. L. (1992). Cleavage and recognition pattern of a double-strand-specific endonuclease (I-*Cre* I) encoded by the chloroplast 23S rRNA intron of *Chlamydomonas reinhardtii*. *Gene*, **119**, 247–251.
15. Marshall, P. & Lemieux, C. (1991). Cleavage pattern of the homing endonuclease encoded by the fifth intron in the chloroplast large subunit rRNA-encoding gene of *Chlamydomonas eugametos*. *Gene*, **104**, 241–245.
16. Dalgaard, J. Z., Garrett, R. A. & Belfort, M. (1993). A site-specific endonuclease encoded by a typical archaeal intron. *Proc. Natl Acad. Sci. USA*, **90**, 5414–5417.
17. Durrenberger, F. & Rochaix, J.-D. (1993). Characterization of the cleavage site and the recognition sequence of the I-*Cre* I DNA endonuclease encoded by the chloroplast ribosomal intron of *Chlamydomonas reinhardtii*. *Mol. Gen. Genet.* **236**, 409–414.
18. Argast, G. M. (1998). I-*Ppo* I and I-*Cre* I homing site sequence degeneracy determined by random mutagenesis and sequential *in vitro* enrichment. *J. Mol. Biol.* **280**, 345–353.
19. Jurica, M. S., Monnat, R. J. & Stoddard, B. L. (1998). DNA recognition and cleavage by the LAGLIDADG homing endonuclease I-*Cre* I. *Mol. Cell*, **2**, 469–476.
20. Koufopanou, V., Goddard, M. R. & Burt, A. (2002). Adaptation for horizontal transfer in a homing endonuclease. *Mol. Biol. Evol.* **19**, 239–246.
21. Goddard, M. R. & Burt, A. (1999). Recurrent invasion and extinction of a selfish gene. *Proc. Natl Acad Sci. USA*, **96**, 13880–13885.
22. Lucas, P., Otis, C., Mercier, J. P., Turmel, M. & Lemieux, C. (2001). Rapid evolution of the DNA-binding site in LAGLIDADG homing endonucleases. *Nucl. Acids Res.* **29**, 960–969.
23. Ichiyanagi, K., Ishino, Y., Ariyoshi, M., Komori, K. & Morikawa, K. (2000). Crystal structure of an archaeal intein-encoded homing endonuclease PI-PfuI. *J. Mol. Biol.* **300**, 889–901.
24. Silva, G. H., Dalgaard, J. Z., Belfort, M. & Van Roey, P. (2003). Crystal structure of the thermostable archaeal intron-encoded endonuclease I-*Dmo* I. *J. Mol. Biol.* **286**, 1123–1136.
25. Moure, C. M., Gimble, F. S. & Quiocho, F. A. (2002). Crystal structure of the intein homing endonuclease PI-SceI bound to its recognition sequence. *Nature Struct. Biol.* **9**, 764–770.
26. Pingoud, A. & Jeltsch, A. (2001). Structure and function of type II restriction endonucleases. *Nucl. Acids Res.*, 3705–3727.
27. Jurica, M. S., Monnat, R. J., Jr & Stoddard, B. L. (1998). DNA recognition and cleavage by the LAGLIDADG homing endonuclease I-CreI. *Mol. Cell*, **2**, 469–476.
28. Schneider, T. D., Stormo, G. D., Gold, L. & Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* **188**, 415–431.
29. 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.
30. Joachimiak, A., Haran, T. E. & Sigler, P. B. (1994). Mutagenesis supports water mediated recognition in the trp repressor–operator system. *EMBO J.* **13**, 367–372.
31. Ladbury, J. E., Wright, J. G., Sturtevant, J. M. & Sigler, P. B. (1994). A thermodynamic study of the trp repressor–operator interaction. *J. Mol. Biol.* **238**, 669–681.
32. Wilson, D. S., Guenther, B., Desplan, C. & Kuriyan, J. (1995). High resolution crystal structure of a paired (Pax) class cooperative homeodomain dimer on DNA. *Cell*, **82**, 709–719.
33. Wilson, D. S., Sheng, G., Jun, S. & Desplan, C. (1996). Conservation and diversification in homeodomain–DNA interactions: a comparative genetic analysis. *Proc. Natl Acad. Sci. USA*, **93**, 6886–6891.
34. Deibert, M., Grazulis, S., Janulaitis, A., Siksnys, V. & Huber, R. (1999). Crystal structure of MunI restriction endonuclease in complex with cognate DNA at 1.7 Å resolution. *EMBO J.* **18**, 5805–5816.
35. Seligman, L., Chisholm, K. M., Chevalier, B. S., Chadsey, M. S., Edwards, S. T., Savage, J. H. & Veillet, A. L. (2002). Mutations altering the cleavage specificity of a homing endonuclease. *Nucl. Acids Res.* **30**, 3870–3879.
36. Lukacs, C. M., Kucera, R., Schildkraut, I. & Aggarwal, A. K. (2000). Understanding the immutability of restriction enzymes: crystal structure of BglII and its DNA substrate at 1.5 Å resolution. *Nature Struct. Biol.* **7**, 134–140.
37. Grazulis, S., Deibert, M., Rimseliene, R., Skirgaila, R., Sasnauskas, G., Lagunavicius, A. *et al.* (2002). Crystal structure of the Bse634I restriction endonuclease: comparison of two enzymes recognizing the same DNA sequence. *Nucl. Acids Res.* **30**, 876–885.
38. Yang, Z. & Bielawaki, J. P. (2000). Statistical methods for detecting molecular adaptation. *TREE*, **15**, 496–503.
39. Hughes, A. L. (1999). *Adaptive Evolution of Genes and Genomes*, Oxford University Press, Oxford.
40. Chevalier, B. S., Kortemme, T., Chadsey, M. S., Baker, D. R. J., Monnat, J. & Stoddard, B. L. (2002). Design, activity and structure of a highly specific artificial endonuclease. *Molec. Cell*, **10**, 895–905.
41. Wang, J., Kim, H.-H., Yuan, X. & Herrin, D. L. (1997). Purification, biochemical characterization and protein–DNA interactions of the I-*Cre* I endonuclease produced in *Escherichia coli*. *Nucl. Acids Res.* **25**, 3767–3776.
42. McRee, D. E. (1999). A versatile program for manipu-

lating atomic coordinates and electron density. *J. Struct. Biol.* **125**, 156–165.
43. Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallog. sect. A*, **47**, 110–119.
44. Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W. *et al.* (1998). Crystallography and NMR system: a new software suite for macromolecular structure determination. *Acta Crystallog. sect. D*, **54**, 905–921.
45. Brunger, A. (1993). Assessment of phase accuracy by cross validation: the free R value. *Meth. Appl. Acta Crystallog. sect. D*, **49**, 24–36.
46. Laskowski, R. J., Macarthur, M. W., Moss, D. S. & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallog.* **26**, 283–291.
47. Lu, X.-J., Zippora, S. & Olson, W. K. (2000). A DNA conformational motifs in ligand bound double helices. *J. Mol. Biol.* **300**, 819–840.

Edited by K. Morikawa

(Received 11 December 2002; received in revised form 27 March 2003; accepted 27 March 2003)