4158–4167 *Nucleic Acids Research, 2012, Vol. 40, No. 9* *Published online 30 December 2011*
doi:10.1093/nar/gkr1250

Recognition of dual symmetry by the controller protein C.Esp1396I based on the structure of the transcriptional activation complex

J. E. McGeehan, N. J. Ball, S. D. Streeter, S.-J. Thresh and G. G. Kneale*

Biomolecular Structure Group, Institute of Biomedical and Biomolecular Sciences, School of Biological Sciences, University of Portsmouth, Portsmouth, Hampshire PO1 2DY, UK

Received October 16, 2011; Revised November 24, 2011; Accepted November 30, 2011

ABSTRACT

The controller protein C.Esp1396I regulates the timing of gene expression of the restriction–modification (RM) genes of the RM system Esp1396I. The molecular recognition of promoter sequences by such transcriptional regulators is poorly understood, in part because the DNA sequence motifs do not conform to a well-defined symmetry. We report here the crystal structure of the controller protein bound to a DNA operator site. The structure reveals how two different symmetries within the operator are simultaneously recognized by the homo-dimeric protein, underpinned by a conformational change in one of the protein subunits. The recognition of two different DNA symmetries through movement of a flexible loop in one of the protein subunits may represent a general mechanism for the recognition of pseudo-symmetric DNA sequences.

INTRODUCTION

Restriction–modification (RM) systems play a central role in modulating the horizontal transfer of genes in bacterial populations and thus in the transmission of antibiotic resistance between bacterial species (1). An understanding of the molecular mechanisms of gene regulation in RM systems, and their impact on the flow of genetic information in bacterial populations, is thus of great interest.

RM systems encode a restriction endonuclease (ENase) and a DNA methyltransferase (MTase). The sequence-specific DNA methyltransferase protects the host DNA from cleavage by the associated restriction enzyme, and the specific methylation pattern of the host RM system allows the discrimination of 'self' from 'non-self' DNA (2). Premature expression of the endonuclease prior to protection of the host DNA by the methyltransferase would be lethal. Thus, there are a variety of control mechanisms that ensure the correct temporal expression of RM genes. In many systems, this is accomplished by means of a 'controller' (C) protein encoded by a gene downstream of its own promoter, and the C-gene is co-transcribed with the endonuclease (R) gene as a single transcriptional unit (3–7). The C-protein binds at various sites within the C/R promoter to regulate transcription of its own gene and the associated endonuclease gene (8).

Measurements of C-dependent transcriptional activity *in vitro*, together with mathematical modelling of the gene control circuits, have shown the time dependence of the activity of this switch (9). *In vivo* experiments have directly demonstrated a time lag in the expression of the ENase with respect to the MTase when the C-protein is expressed in a new host (10).

In most C-protein systems so far investigated, the operator sequence at the C/R promoter has binding sites (denoted O_L and O_R) that can accommodate two C-protein dimers (11,12). O_L is distal to the gene and has the highest affinity for a C-protein dimer. O_R is proximal to the gene and the intrinsic affinity for this site is weak; however, when a C-protein dimer is bound to O_L then the affinity for O_R increases around 1000-fold (12,13).

Earlier biochemical and biophysical analysis in our laboratory suggested the basis of the genetic switch in AhdI (11–15). Low-level expression of the C-protein from a weak promoter leads to a delay in transcription until sufficient protein accumulates to form a functional dimer. The C-protein dimer activates transcription of the C/R operon, forming a positive feedback loop, which leads to a rapid increase in C-protein expression; at higher concentrations, a second dimer is recruited to the promoter, displacing the σ subunit of RNA polymerase and thereby repressing transcription of its own gene (and hence expression of the R gene) in a negative feedback loop (Figure 1). A similar, but more complex, mechanism has been proposed for the R–M system

*To whom correspondence should be addressed. Tel: +02392 842 678; Fax: +02392 842 053; Email: geoff.kneale@port.ac.uk

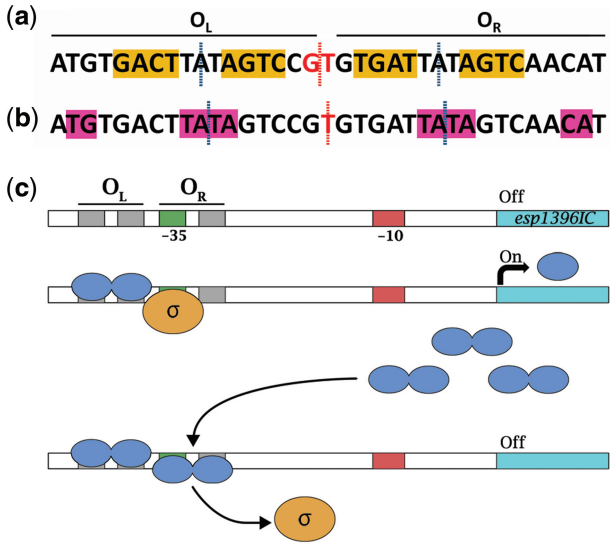

Figure 1. The conserved DNA sequences obey different symmetries. (**a**) The C-boxes are symmetrical if the pseudo-dyad axis is placed between the central GT. (**b**) If the pseudo-dyad axis is placed at the central T, the C-boxes are no longer symmetrical but the other conserved element are. The pseudo dyad axes within operators (blue) and between operators (red) are shown as dotted lines. Figure reproduced from McGeehan *et al.* (2008). (**c**) C-proteins act as a genetic switch regulating the timing and expression of R–M genes. The −35 (green) and −10 (red) regions are indicated upstream of the C-gene (light blue) and the R-gene (data not shown). C-protein dimers are shown in blue and the sigma subunit of RNA polymerase in orange. C-protein is expressed at low levels from a weak C-independent promoter (data not shown). A C-protein dimer first occupies the high-affinity O_L site and stimulates transcription of the C-gene through recruitment of RNA polymerase sigma subunit to the −35 site. As the C-protein concentration increases, a dimer occupies the O_R site and occludes the −35 site down-regulating the expression of the C- and R-genes. Adapted from McGeehan *et al.* (11).

Esp1396I. Experiments conducted in collaboration with Severinov and colleagues (16) have shown that in the R–M system Esp1396I, the C-protein, in addition to regulation of the R gene, represses the M gene by binding as a dimer to a high-affinity site that overlaps the transcriptional start site.

Bioinformatic analysis of known and potential C-protein binding sites has identified a repeating symmetrical 'consensus' sequence consisting of four quasi-symmetrical 'C-boxes' GACT—AGTC—-GACT—AGTC upstream of the C/R genes in a wide variety of R-M systems (6,8), and a similar sequence is found within the 35-bp sequence that has been identified by DNA footprinting in AhdI (12). However, the degree of sequence homology between species is moderate and the internal symmetry between 'C-boxes' is far from perfect (Figure 1a). The GT dinucleotide in the centre of the proposed consensus sequence is in fact more highly conserved than the proposed GATC tetranucleotide recognition sequences (8), but clearly lacks dyad symmetry. The proposed 3-bp 'spacers' within the left and right operator sequences are equally well conserved, the consensus sequence being TAT.

We initially solved the structure of the controller protein C.Esp1396I bound as a tetramer (i.e. two dimers) to its 35-bp operator sequence—the first DNA–protein structure for any C-protein complex (11). The structure of the nucleoprotein complex shows the molecular basis of cooperative binding, consisting of protein–protein electrostatic contacts between dimers, together with structural changes in the DNA that facilitate binding of the second dimer. In the crystal structure of the C.Esp1396I—35-bp operator complex (PDB code: 3CLC), the pseudo-dyad axis relating the two operators is shifted by half a base (i.e. centred on T rather than the expected GT). Although the pseudo-dyad between GACT/AGTC sequences is then lost, there are instead perfectly symmetrical TATA sequences at the centre of each operator (Figure 1b).

The structure suggested the mechanism whereby cooperative binding of dimers to the DNA operator governed the switch from activation to repression of the C and R genes (11). The overall structure of the complex comprises two dimers bound to the DNA, each centred on the pseudo-dyad located between the central A and T bases in the TATA sequence that is found at the centre of each operator site. The two dimers are bound to approximately opposite faces of the DNA. Each dimer bends the DNA by ca. 50°, and inserts helix-3 of the classical HTH motif into the major groove of DNA, either side of the central TATA within each operator. In this structure, the two protein dimers are related by a dyad axis that coincides with the pseudo-dyad axis lying within the central T:A base pair of the 35-bp duplex.

Some clear protein–DNA interactions were also identifiable, in particular the interaction of R35 with the conserved G3 on both DNA strands. However, the structure was relatively low resolution and, moreover, since both orientations of the DNA were present in the asymmetric unit, the resulting structure was symmetry-averaged, which precluded a detailed analysis of the protein–DNA contacts.

In order to clearly identify the protein–DNA interactions, and thus determine the molecular basis of DNA sequence recognition (and in particular how deviations from symmetry within the DNA recognition site are accommodated), we have crystallized a C.Esp1396I dimer bound to a single dimer binding site, O_L (the stronger of the two binding sites that are located upstream of the endonuclease gene). The DNA sequences employed were designed with overhanging bases, to facilitate intermolecular packing via end-to-end stacking of unpaired bases. What we found, however, was a completely novel packing interaction, in which the bases formed DNA triplets between adjacent DNA duplexes.

MATERIALS AND METHODS

Purification

Purification of C.Esp1396I was carried out as previously described (11). Briefly, large-scale cultures of *E. coli* BL21(DE3) containing the plasmid pET-28b/*esp1396IC* were grown. Over-expressed C.Esp1396I was harvested by sonication and purified using nickel affinity chromatography. The N-terminal hexa-histidine tag was removed by thrombin digestion but the purified protein retained

a GSH tripeptide. DNA oligonucleotides were purified as previously described and annealed to form a duplex (11).

Crystallization

C.Esp1396I was incubated with an 18-bp duplex DNA at varying ratios prior to crystal screening. Selected crystals were formed in the PACT Premier 67 condition (0.2 M sodium acetate, 0.1 M *Bis–Tris*-propane, 20% PEG3350, at pH 5.0 in 4 µl hanging drops (2 µl protein/DNA and 2 µl mother liquor) at 16°C. The final protein:DNA ratio was 2:1 (monomer:DNA) at ~30 µM final DNA concentration. The crystals were mounted in litholoops (Molecular Dimensions), cryoprotected in 35% v/v ethylene glycol and cryo-cooled in liquid nitrogen.

Structure solution and refinement

Data were collected from cryo-cooled crystals of the 19-mer complexes on ID14-4 at the ESRF (Grenoble) at 100 K using an ADSC 4Q CCD detector. The complex crystallized in the space group $P22_12_1$; reflections extended to ~1.9 Å and 110 images were collected with an oscillation angle of 1°. The data were processed and scaled using XDS/XSCALE (17) with an overall R_{merge} of 11% and an overall completeness of 90.7% at 2.1 Å (Table 1). The scaled data was phased by molecular replacement with C.Esp1396I dimer bound to the left operator within the tetrameric complex (chains A and B, residues 5–75, chain C, residues 6–13 and chain D, residues 23–30) as the search model using Phaser (18). Iterative refinement was carried out using Refmac5 (19) with TLS restraints enabled (Table 1). The missing DNA bases were manually added into interpretable electron density using Coot (20), as were 11 of the 16 missing terminal amino acid residues. Waters were added during refinement with Refmac5 and checked manually. The final structure contained all 38 bases and amino acid residues 2–77 and 3–79 in chains A and B, respectively. The final parameters used during refinement are shown in Table 1. The DNA base pair parameters were calculated using the software package CURVES+ (http://gbio-pbil.ibcp.fr/cgi/Curves_plus). The coordinates of the nucleoprotein complex have been deposited in the Protein Structure Database (PDB code 3S8Q).

RESULTS

Structure solution

The DNA sequence (an 18-bp duplex with an overhanging base at the 5′-end of each strand) was designed to aid the formation of pseudo-continuous DNA in a single orientation and thus overcome the symmetry-averaging problems encountered in the tetramer complex structure (11). The averaging problem was indeed overcome in this structure, but the DNA did not form a pseudo-continuous helix. Instead, the DNA ends are involved in crystal packing interactions between symmetry related molecules and form triple helical interactions (Figure 2). The terminal two bases are paired on both the Hoogstein and Watson–Crick edges to form a base 'triplet' at both ends of the DNA (T-AT and A-GC), which maximizes base stacking. These triple helical interactions help to stabilize the DNA ends, which refined with low B-factors, despite not being involved in protein–DNA interactions.

One complex, consisting of a C.Esp1396I dimer bound to a DNA duplex (Figure 3), is present in the asymmetric unit and the structure refined to 2.1 Å with a final R/R_{free} of 16.8/22.4% (Table 1). Iterative refinement was carried out using Refmac (19) with TLS restraints enabled. All of the DNA bases are clearly resolved, as are all except a few amino acid residues at the N and C termini of each protein subunit. In addition, a total of 314 solvent molecules could be located, including a number of water molecules mediating protein–DNA interactions.

From a superposition of the dimeric O_L complex and the appropriate region of the tetrameric complex (Supplementary Figure S1), the protein and the DNA components of both complexes are for the most part identical, although the side-chains of the protein and the bases of the DNA can be positioned with far greater reliability in the dimeric complex. One major difference, however, is that one region of the protein (residues 43–47) exhibits a different conformation in each subunit in the dimeric complex. This is probably also the case in the tetrameric

Table 1. X-ray crystal data, refinement parameters and model statistics for the 19-mer O_L complex structure

Data collection	
Space group	$P22_12_1$
Unit-cell parameters (Å, °)	$a = 44.3$ $b = 61.5$ $c = 113.7$ $\alpha = \beta = \gamma = 90$
Resolution limits (Å)	50–2.1 (2.2–2.1)
R_{merge}* (%)	11.0 (37.1)
$I/\sigma(I)$	15.3 (4.6)
Completeness (%)	90.7 (86.9)
Refinement parameters	
Scaling	Babinet
TLS	
No. of Groups	4
Description	Individual chains
Refinement model statistics	
No. of reflections	17 096
R_{cryst}/R_{free}‡ (%)	16.8/22.4
No. of atoms	
Protein	1279
DNA	773
Water	314
Average B-factors (Å^2)	
Protein	31.85
DNA	35.49
Water	47.29
RMS deviations from ideal	
Bond lengths (Å)	0.01
Angles (°)	1.3

Values in parentheses are for the highest resolution shell. *$R_{merge} = \Sigma_{hkl}\Sigma_i|I_i(hkl) - «I(hkl)»|/\Sigma_{hkl}\Sigma_i I_i(hkl)$, where «$I$(hkl)» is the mean intensity of reflection I(hkl) and I_i(hkl) is the intensity of an individual measurement of reflection I(hkl). ‡$R_{cryst} = \Sigma_{hkl}|F_{obs}| - |F_{calc}|/\Sigma_{hkl}|F_{obs}|$, where F_{obs} is the observed structure–factor amplitude and F_{calc} is the calculated structure–factor amplitude. R_{free} is the same as R_{cryst} but for the 5% of structure–factor amplitudes that were set aside during refinement.

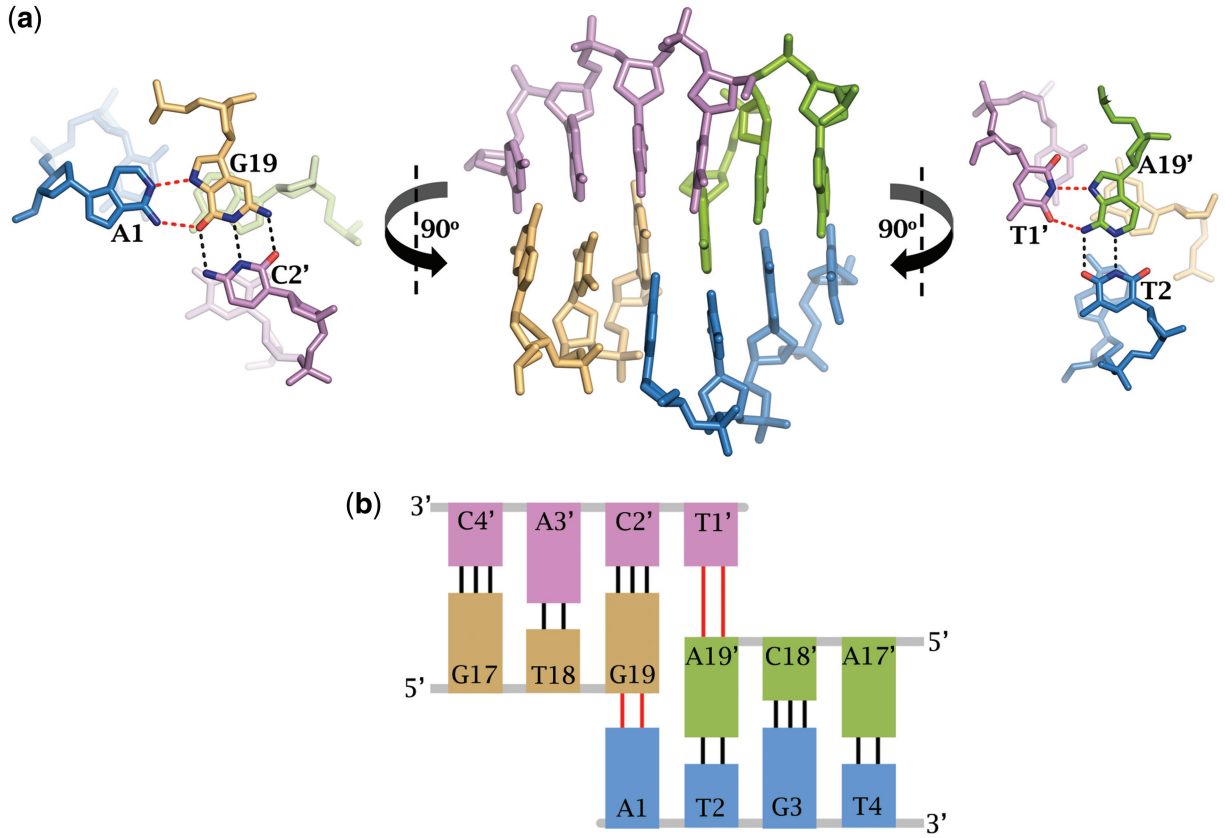

Figure 2. Triple helical DNA interactions between symmetry related 19-mer complexes. (**a**) The triple helical interactions occur between A_1 and G_{19} of chain C (blue and beige, respectively) and T'_1 and A'_{19} of chain D (pink and green, respectively). (**b**) A cartoon representation depicting how the Watson–crick edge of the overhanging 5′ base (either T'_1 or A_1) forms hydrogen bonds with the Hoogstein edge of the terminal base of a symmetry related molecule (A'_{19} or G_{19}, respectively).

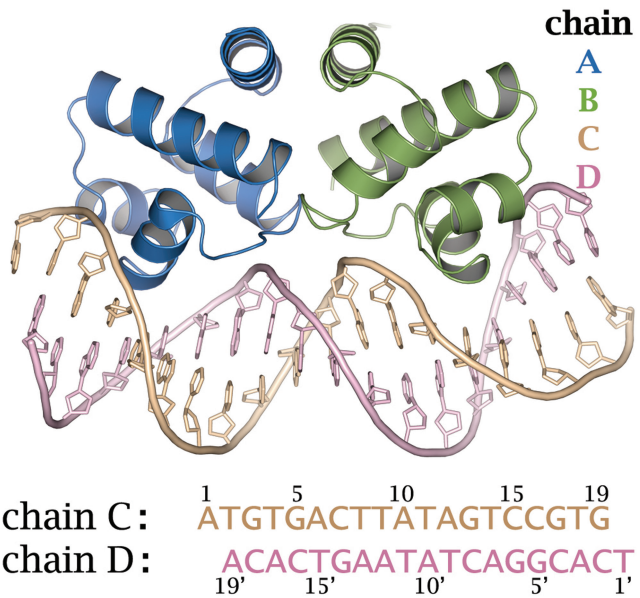

Figure 3. The C.Esp1396I O_L nucleoprotein complex structure. The asymmetric unit of the crystal contains a C.Esp1396I dimer (chains A and B; blue and green, respectively) bound to DNA (chains C and D; beige and pink, respectively). The DNA duplex consists of 18 bp with a 5′ overhang on each strand.

complex, but the combined averaging effects, together with the low resolution of the data, resulted in smeared density in these areas and prevented unambiguous refinement. These two alternative loop conformations were first identified in the free protein structure, where there are seven dimers in the asymmetric unit (21). Two of the 14 subunits in the asymmetric unit adopted a different conformation to the other 12, and the two conformations are likely to be of comparable stability. In the case of the O_L nucleoprotein complex, both conformations can be found in the same dimer; in monomer A, the loop adopts the 'major' conformation and in monomer B, it adopts the 'minor' conformation, as can be clearly seen in the electron density (Supplementary Figure S2). The different loop conformations in the DNA–protein complex may reflect the departure from true dyad symmetry in the O_L operator sequence (Figure 1).

Analysis of protein–DNA interactions

The recognition helix. The recognition helix (residues 35–43) of each subunit inserts into the major groove of the DNA. Two residues in this helix, R35 and T36, are involved in direct readout of the DNA sequence while other residues are involved in non-sequence-specific interactions with the phosphate backbone (Supplementary

Figures S3 and S4). In both subunits of the dimer, the γ-hydroxyl of T36 H-bonds to the N4 group of a cytosine; however, T36 in chain A recognizes C'_{16} of one DNA strand and the T36 in chain B recognizes C_{15} of the other strand.

The amino groups of R35 in each subunit interact with the N7 and O6 of a guanine. R35 in chain A recognizes G_3 on one DNA strand (Figure 4a,b). However, the R35 in chain B cannot make the symmetry equivalent interaction on the other strand, as an adenine (A'_3) rather than a guanine is in the equivalent position in chain D. Instead, the R35 interacts with the N7 and O6 of the G_{17} on chain C (Figure 4c); it is clear that the flexible side-chain of arginine is capable of accommodating the departure from dyad symmetry in the O_L DNA sequence. In addition to hydrogen bonding with G_3, R35 in chain A is involved in indirect readout by stacking of the planar guanidinium group with the exposed face of T_2 as described elsewhere (11) and is thus specific for a TG dinucleotide at this position. There is no equivalent stacking of R35 of chain B, however, since there is no equivalent TG dinucleotide at this site (there is an adjacent T but it is on the 3′ side of the G, which therefore adopts a very different base stacking pattern).

Helix 3 of C.Esp1396I recognizes bases within the 5′-AGTC sequence (the 'bottom' strand), rather than the 5′-GACT on the complementary ('top') strand. Each recognition helix of the dimer makes these two direct sequence-specific interactions, one (from each subunit) to the highly conserved G_3 or G_{17} base outside of the C-box and another to the cytosine complementary to the G in the AGTC sequence. In addition, however, the G and T bases in this sequence are recognized by R46 that lies within the flexible loop region (as discussed below). The two C.Esp1396I monomers are able to make non-symmetrical interactions with the DNA due to the flexible nature of the R35 sidechain, and can thus adapt to the asymmetrical location of the C-boxes in the O_L DNA sequence.

The alternative loop conformations. In the free protein structure, two alternative conformations of the loop region (residues 43–46) between helices 3 and 4 were observed (21). Comparing the two conformations, the side-chains of N44 and S45 are flipped almost 180° about the peptide backbone. In the minor conformation, it was postulated that the polar head groups of the asparagine (N44) sidechain would be in close proximity to the DNA and may be involved in protein–DNA interactions (21).

Figure 5 shows both loop conformations with respect to the DNA, and the atoms involved (see also Supplementary Figure S5). The side-chain of N44 in the major loop (Figure 5a) points towards the core of the protein and the terminal carbonyl and amino groups, are stabilized directly through interactions with the backbone amino group of S7 and the γ-hydroxyl of S10. The δ-amino of N44 and the backbone carbonyl of S7 also coordinate a water. These interactions provide stability for the major loop conformation. In the minor loop conformation (Figure 5b) the N44 rotates ~180° about the backbone and the side-chain points towards the DNA backbone. The δ-amino of N44, the η-amino of R43 and the phosphate oxygen of G5 coordinate a water that stabilizes the N44 sidechain.

Although the side-chain of S45 is rotated ~180° between the two loop conformations, in both instances the γ-hydroxyl is stabilized by interactions with other amino acid side-chains. In the major loop conformation, the γ-hydroxyl interacts with the backbone carbonyl of a glycine (G4) while in the minor loop conformation the

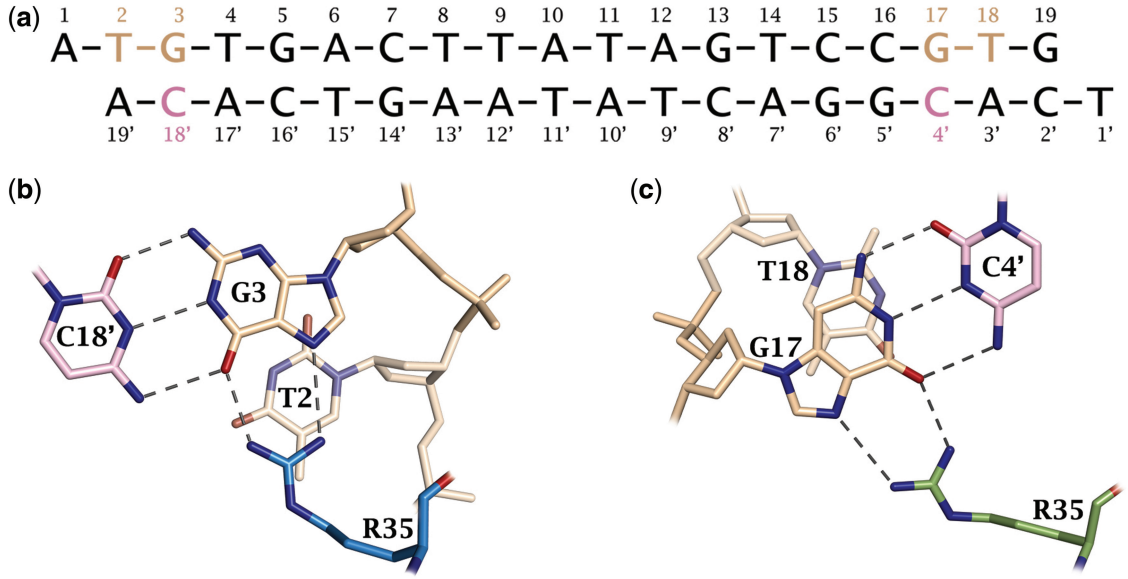

Figure 4. Direct and indirect readout of DNA by R35. (**a**) The R35 recognizes the G3 and G17 in the DNA sequence and the highlighted bases (beige: chain C and pink: chain D) are shown in b and c. (**b**) The R35 from chain A recognizes the conserved TG by interacting with the O6 and N7 of the guanine base. The planar guanidinium group of R35 also stacks with the thymine base. (**c**) The symmetry related interaction cannot be made by chain B, which instead recognizes G17 via the O6 and N7. All hydrogen bond distances are <3.2 Å.

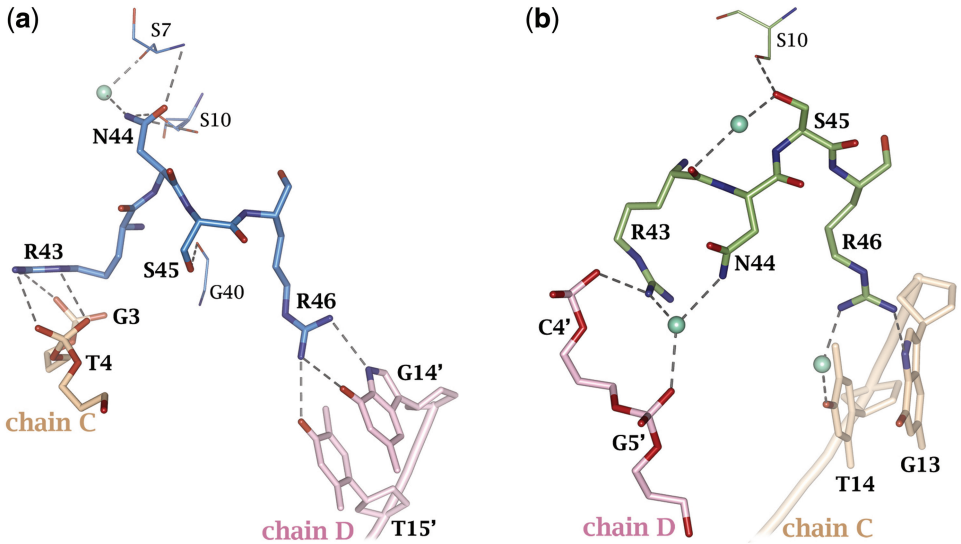

Figure 5. Comparison of the interactions made by the flexible loop region in chains A and B of the 19-mer O_L complex structure. The hydrogen bond interactions made by the flexible loop regions in chains A and B (**a** and **b**, respectively) are shown as black dashes. Residues involved in stabilizing the loop region are represented as thin lines. Water molecules are represented by green spheres.

γ-hydroxyl interacts with the γ-hydroxyl of a serine (S10) and also coordinates a water. The R43 interacts with the phosphate backbone in both loop conformations and the polar groups of N44 are also stabilized in both conformations, albeit by different groups.

In both conformations of the loop region, arginine 46 is involved in direct readout of the DNA sequence. In the major loop conformation (monomer A), R46 interacts with the N7 and O6 of G'_{14} from chain D, as well as H-bonding to the O4 of the adjacent T'_{15} (Figure 5a and Supplementary Figure S5a). Only guanine and thymine have electron donors in the correct position to interact with the amino groups of this arginine; thus the recognition of these two bases confers specificity. In the minor loop conformation (monomer B), R46 is involved in interactions with the G_{13} and T_{14} of chain C (Figure 5b and Supplementary Figure S5b). However, the head group of the arginine is positioned in such a way as to only directly interact with the N7 of G_{15}. This interaction is sufficient to distinguish between purine and pyrimidine but is unable to distinguish between adenine and guanine. The η-amino group of R46 also makes a water mediated contact with the O4 of T_{14}. As water can act as either a donor or an acceptor in hydrogen bonding, this interaction cannot distinguish thymine from cytosine. This water forms part of the network of highly organized waters found in the major groove of the DNA.

Symmetry of the protein subunits and DNA. Figure 6a shows a superposition of subunits A and B by a rotation about the dyad axis that relates them. The overall backbone structure of the two subunits is very similar, with the only notable difference occurring in the flexible loop region. The GTC motif recognized the by amino acid side-chains of T36 and R46 shifts by approximately half a base pair relative to the protein. The flexible loop is able to accommodate this half base pair shift, permitting recognition of the GTC by monomer B. In an alternative view, if the GTC bases that are recognized by each subunit (one in each half-site) are superimposed (Figure 6b), the protein rotates by ~30° around the helix axis of the DNA.

DNA structure and backbone interactions. Analysis of the OL DNA structure in the complex was performed using CURVES (22). The minor groove at the TATA site is compressed (from ~7 Å to ~2 Å), which leads to the DNA being significantly bent about this sequence, as shown in Figure 7. The overall bend of the DNA duplex is ~40°. From circular permutation assays, it is clear that the DNA is not intrinsically bent; rather, the bending is induced when the C-protein binds to its operator, i.e. the sequence of the operator DNA is one that can readily be deformed when the C-protein binds (11,14). DNA bending around the TATA site permits a form of indirect readout. The bend in the DNA at the TATA site is accompanied by deviations in the base pair and step parameters (Figure 8 and Supplementary Figure S6). The parameters for the OL sequence show values that are typical of TATA sequences (23) except for the roll parameter, which is closer to standard B-form DNA. The twist values for the two thymines differ significantly from the standard B-form values, suggesting that the DNA bending can be achieved by partially unstacking the TATA bases.

The tetrameric complex structure suggested a possible role for Y37 and S52 of each subunit in the dimer in compressing the minor groove at the TATA sequences by binding to the phosphate backbone of the DNA on either side (11). These interactions are seen much more clearly in the O_L complex structure, with additional backbone contacts being made by N47 (Figure 7). For each subunit, the hydroxyl groups of Y37 and S52 interact with the same phosphate group of the DNA (5′ of nucleotide 13 in both DNA strands) and the amino group of N47 interacts with the phosphate 5′ of

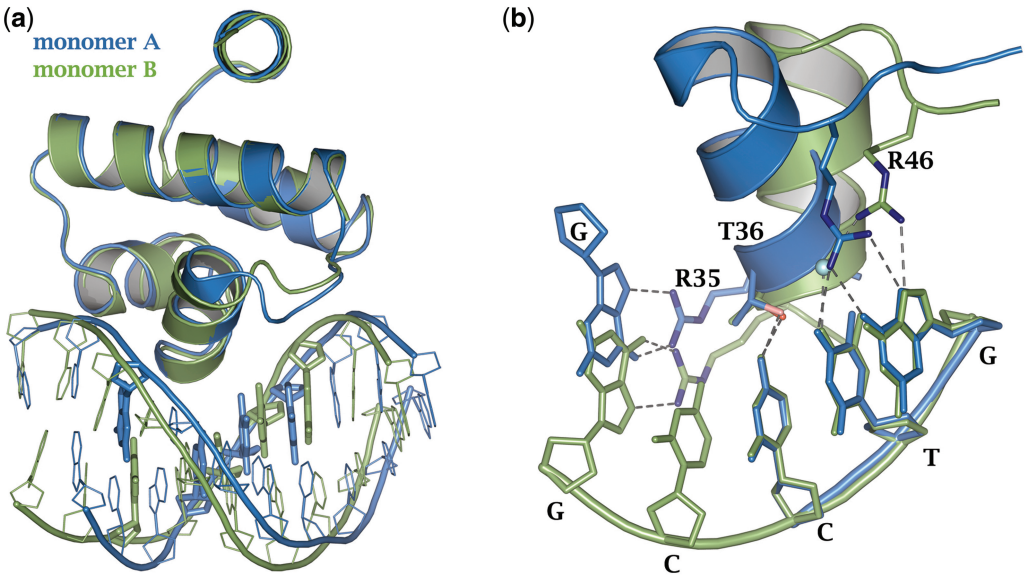

Figure 6. Comparison between the two half sites in the 19-mer O_L complex structure. (**a**) Monomers A and B (blue and green, respectively) were superimposed with RMSD = 0.34 Å (222 main chain atoms). The DNA bases are offset by approximately half a base pair. Bases involved in direct readout are shown as thick lines. The half base pair shift is compensated for through the flexibility of the loop region, permitting recognition of the GTC bases. (**b**) Residues 13–15 of chain C (blue) were superimposed upon residues 14–16 of chain D (green) (RMSD = 0.54 Å over 61 backbone atoms). Hydrogen bonds are represented by dashed lines.

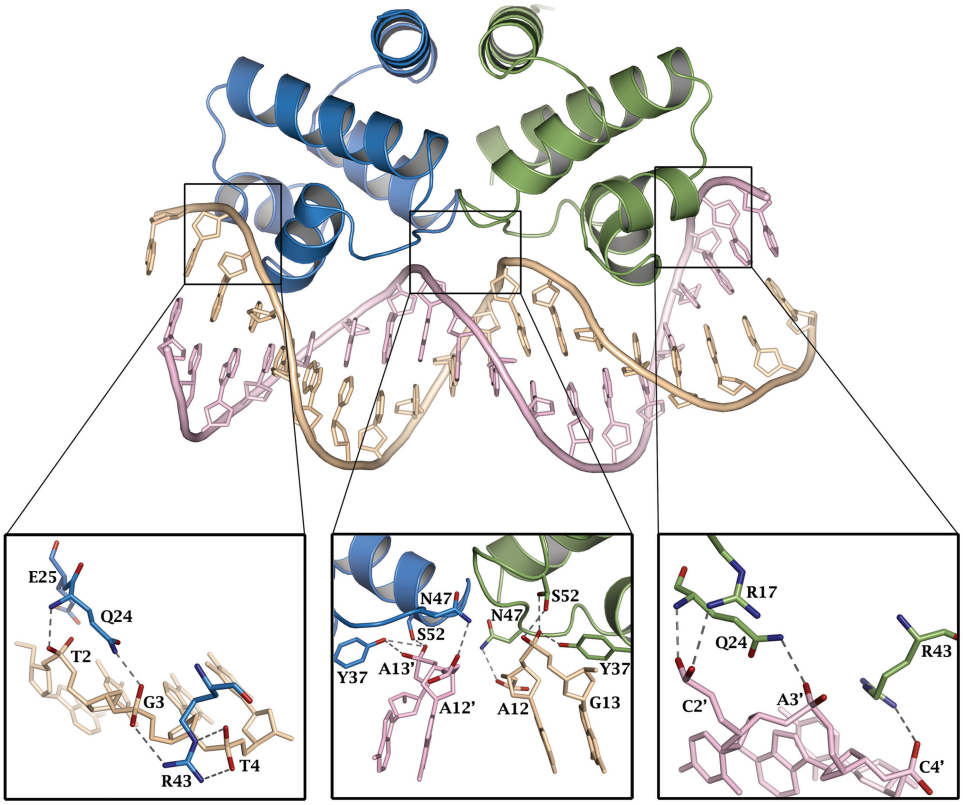

Figure 7. Non-specific DNA contacts stabilize the complex and the compression of the minor groove. Chains A–D are coloured blue, green, beige and pink, respectively. Hydrogen bonds are represented by dashed lines. The nucleoprotein complex is stabilized by non-specific interactions between amino acids (e.g. R43; inset left and right) and the phosphodiester backbone of DNA. The minor groove is compressed at the TATA sequence, which results in the DNA being significantly bent. Y37, N47 and S52 from both monomers are involved in stabilizing the bend (inset centre).

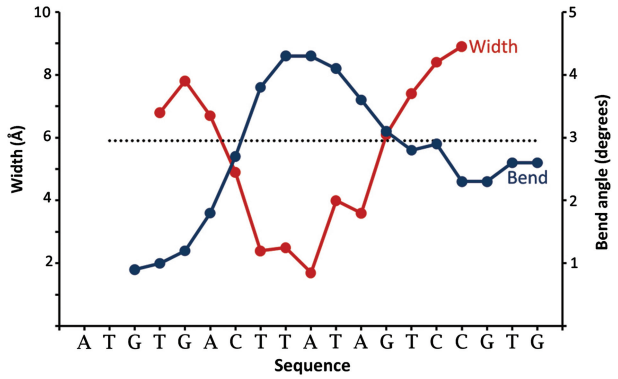

Figure 8. Compression of the minor groove at the TATA sequence results in DNA bending. The overall DNA bend is ~41° and the local bend angle between adjacent base pairs (calculated as the angle formed between the normals of adjacent base pairs) is greatest at the TATA sequence (red line: minor groove width; blue line: local bend angle; dashed line: minor groove width of standard B-form DNA).

nucleotide 12. The serine in chain B has a dual conformation (both conformations were refined with 50% occupancy), but both conformations interact with the same phosphate group. These interactions cause the minor groove to be compressed and the DNA to be bent. In addition, there are interactions of the side-chains of residues R17, N24, S39 and R43 with phosphate groups at either end of the DNA (Figure 9 and Supplementary Figures S7 and S8), which further stabilize the bent DNA conformation. The interactions of the negatively charged phosphate groups with the positively charged guanidinium groups of R17 and R43 will be particularly strong, and should make an important contribution to the overall binding energy.

DISCUSSION

A highly conserved inverted repeat with the consensus GACT...AGTC was originally thought to be the binding motif, and this recognition sequence appeared to

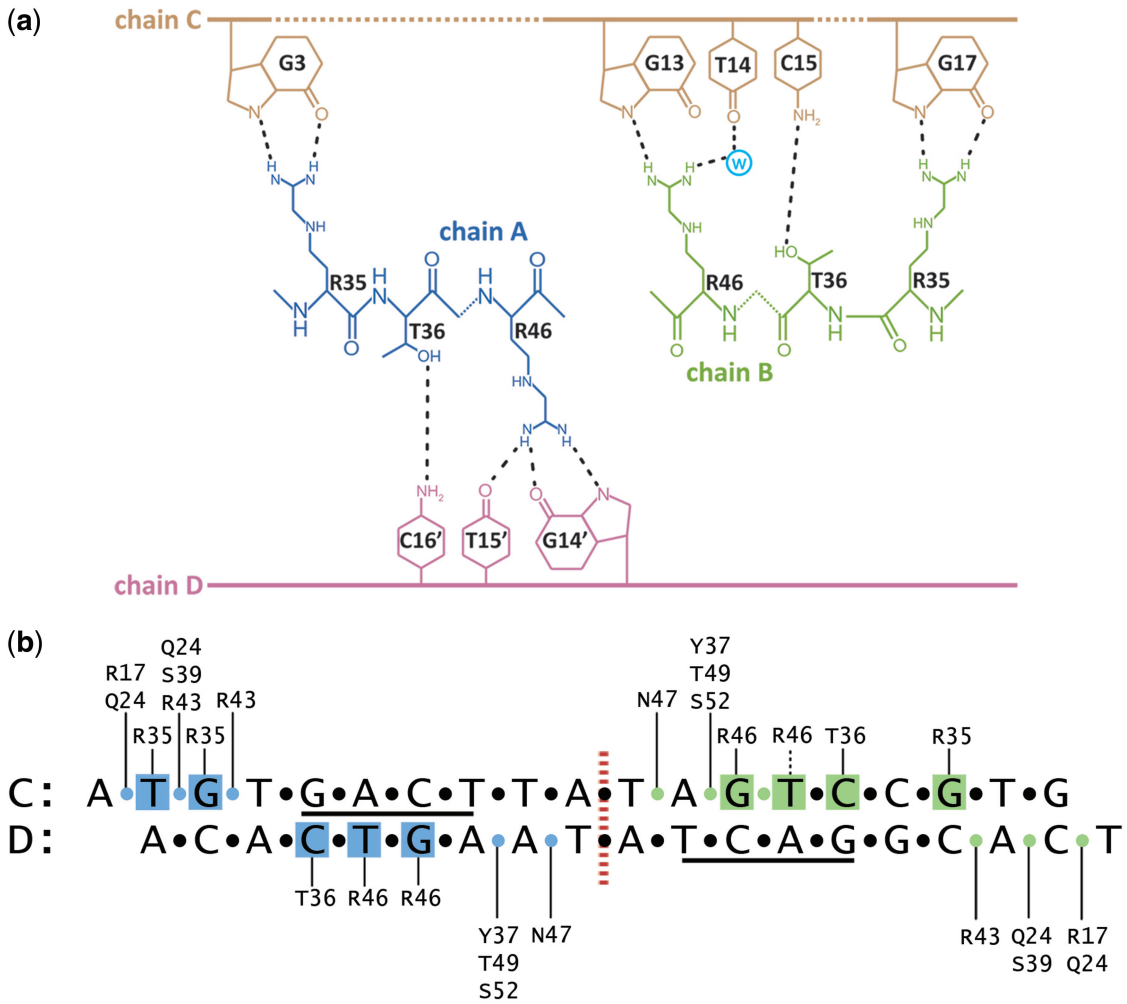

Figure 9. Schematic representation of nucleoprotein interactions. (**a**) Residues Arg35, Thr36 and Arg46 are involved in direct readout of the DNA sequence. (**b**) Overview of protein–DNA interactions. Phosphate groups are represented as circles, and those interacting with the protein are coloured according to the subunit contacted. Interactions between chain A and the DNA are highlighted in blue and interactions between chain B and DNA are highlighted in green (for further details, see Supplementary Figures S7 and S8).

be common to a large family of C-proteins (and in most cases is itself duplicated) in the region upstream of the C/R promoter (6,8,10). In addition to these C-boxes, it was noted that the flanking TG (and the symmetry related CA) is also highly conserved, as are the central GT between dimer binding sites and the TAT sequences between the C-boxes.

However, these motifs follow different symmetries (Figure 1a and b) with the pseudo-dyad axis either between the central GT (C-box symmetry), or through the central T, respectively (11). As each of the above sequence elements are conserved and both have been shown to be important in C-protein binding (16), the protein must be able to accommodate both symmetries. The structure we now report of the C.Esp1396I controller protein dimer bound to the O_L operator shows how this dual recognition is achieved.

The flexible loop that was observed in the free protein plays a fundamental role in breaking the symmetry of the protein dimer and permits base-specific interactions with a variety of DNA sequence motifs that are not completely symmetric. The GTC bases in the C-box are recognized by T36 and R46 of each subunit, and the conserved TG motif (T_2/G_3) and G_{17} are specifically recognized by R35 in the recognition helix of subunits A and B, respectively (Figure 9). In order for the R46 in chain B to interact with the DNA bases in the second half-site, which are displaced from their symmetry equivalent positions, the flexible loop adopts the minor conformation.

From close inspection of Figure 9, it can be seen that the non-specific interactions of the protein with the phosphate groups in the DNA backbone follow the symmetry that has the dyad centred on the TATA sequence, as in Figure 1b. In contrast, the interactions of the GTC bases in the C-box motifs follow the symmetry that is centred on TAT (Figure 1a). The amino acid residue (R46) responsible for the majority of the interactions with the GTC motif moves to accommodate the ~1.7 Å shift (and ~18° rotation) of these bases. This is enabled by the change in the loop conformation in subunit B, as well as by the inherent flexibility of the arginine side-chain, resulting in the relative displacement between the two subunits that can be seen in Figure 6.

Pseudo-symmetric DNA sequences are common in gene control regions of DNA, and the asymmetry plays an important biological role in determining the differential binding affinity for different promoters (24). In the structural analysis of such systems, symmetrized operator sequences have often been employed (25–27). In others where the natural operators have been used, electron density around the non-symmetric bases is unclear [as was indeed the case for the C.Esp1396I tetramer complex (11)]. However, in the lambda cI repressor, non-symmetrical interactions can be seen that depend upon the movement of the flexible N-terminal tail of the protein (28,29).

In C-protein recognition sites, there are additional and more pronounced deviations from symmetry, i.e. a translation between the dyad axis that defines the major recognition motif GTC/GAC (Figure 1a), and the dyad axis defining the protein–DNA backbone interactions (Figure 1b). As discussed above, this displacement (1.7 Å) together with the resulting ~18° rotation between the axes defining the two symmetries is accommodated by a conformational change in one of the subunits, thus breaking the symmetry of the C-protein homo-dimer in order to match that of the DNA recognition sequence. Many other known and putative C-protein recognition sites are similar to those found in the Esp1396I restriction-modification system (30) and it would not be surprising if a similar mechanism of recognition applied in these cases. Indeed, it is possible that the recognition of two different DNA symmetries through movement of a flexible loop in one of the protein subunits may represent a general mechanism for the recognition of such DNA binding sites.

ACCESSION NUMBER

PDB code: 3S8Q.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figures 1–8.

ACKNOWLEDGEMENTS

We are grateful to the ESRF (France) and Diamond Light Source (UK) and associated beam-line staff for provision of synchrotron radiation facilities.

FUNDING

Biotechnology and Biological Sciences Research Council (BBSRC) (research grants BB/E000878/1 to G.G.K. and BB/H00680X/1 to G.G.K. and J.E.M.); Research Councils UK Academic Fellowship (award to J.E.M.); University of Portsmouth studentship (to N.B.). Funding for open access charge: BBSRC.

Conflict of interest statement. None declared.

REFERENCES

1. Akiba,T., Koyama,K., Ishiki,Y., Kimura,S. and Fukushima,T. (1960) On the mechanism of the development of multiple-drug-resistant clones of Shigella. *Jpn. J. Microbiol*, **4**, 219–227.
2. Wilson,G.G. and Murray,N.E. (1991) Restriction and modification systems. *Annu. Rev. Genetics*, **25**, 585–627.
3. Tao,T., Bourne,J.C. and Blumenthal,R.M. (1991) A family of regulatory genes associated with type II restriction-modification systems. *J. Bacteriol*, **173**, 1367–1375.
4. Ives,C.L., Nathan,P.D. and Brooks,J.E. (1992) Regulation of the BamHI restriction-modification system by a small intergenic open reading frame, bamHIC, in both *Escherichia coli* and *Bacillus subtilis*. *J. Bacteriol*., **174**, 7194–7201.
5. Rimseliene,R., Vaisvila,R. and Janulaitis,A. (1995) The eco72IC gene specifies a trans-acting factor which influences expression of both DNA methyltransferase and endonuclease from the Eco72I restriction-modification system. *Gene*, **157**, 217–219.
6. Vijesurier,R.M., Carlock,L., Blumenthal,R.M. and Dunbar,J.C. (2000) Role and mechanism of action of C•PvuII, a regulatory

protein conserved among restriction-modification systems. *J. Bacteriol*, **182**, 477–487.
7. Cesnaviciene,E., Mitkaite,G., Stankevicius,K., Janulaitis,A. and Lubys,A. (2003) Esp1396I restriction-modification system: structural organization and mode of regulation. *Nucleic Acids Res*, **31**, 743–749.
8. Knowle,D., Lintner,R.E., Touma,Y.M. and Blumenthal,R.M. (2005) Nature of the promoter activated by C.PvuII, an unusual regulatory protein conserved among restriction-modification systems. *J. Bacteriol*, **187**, 488–497.
9. Bogdanova,E., Djordjevic,M., Papapanagiotou,I., Heyduk,T., Kneale,G. and Severinov,K. (2008) Transcription regulation of the type II restriction-modification system AhdI. *Nucleic Acids Res*, **36**, 1429–1442.
10. Mruk,I. and Blumenthal,R.M. (2008) Real-time kinetics of restriction–modification gene expression after entry into a new host cell. *Nucleic Acids Res*, **36**, 2581–2593.
11. McGeehan,J.E., Streeter,S.D., Thresh,S.J., Ball,N., Ravelli,R.B. and Kneale,G.G. (2008) Structural analysis of the genetic switch that regulates the expression of restriction-modification genes. *Nucleic Acids Res*, **36**, 4778–4787.
12. Streeter,S.D., Papapanagiotou,I., McGeehan,J.E. and Kneale,G.G. (2004) DNA footprinting and biophysical characterisation of the controller protein C.AhdI suggests the basis of a genetic switch. *Nucleic Acids Res*, **32**, 6445–6453.
13. McGeehan,J.E., Papapanagiotou,I., Streeter,S.D. and Kneale,G.G. (2006) Cooperative binding of the C.AhdI controller protein to the C/R promoter and its role in endonuclease gene expression. *J. Mol. Biol*, **358**, 523–531.
14. Papapanagiotou,I., Streeter,S.D., Cary,P.D. and Kneale,G.G. (2007) DNA structural deformations in the interaction of the controller protein C.AhdI with its operator sequence. *Nucleic Acids Res*, **35**, 2643–2650.
15. McGeehan,J.E., Streeter,S., Papapanagiotou,I., Fox,G.C. and Kneale,G.G. (2005) High-resolution crystal structure of the restriction-modification controller protein C.AhdI from *Aeromonas hydrophila*. *J. Mol. Biol*, **346**, 689–701.
16. Bogdanova,E., Zakharova,M., Streeter,S., Taylor,J., Heyduk,T., Kneale,G. and Severinov,K. (2009) Transcription regulation of restriction-modification system Esp1396I. *Nucleic Acids Res*, **37**, 3354–3366.
17. Kabsch,W. (1993) Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J. Appl. Crystallogr.*, **26**, 795–800.
18. McCoy,A.J., Grosse-Kunstleve,R.W., Storoni,L.C. and Read,R.J. (2005) Likelihood-enhanced fast translation functions. *Acta Crystallogr.*, **61**, 458–464.
19. Murshudov,G.N., Vagin,A.A. and Dodson,E.J. (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr.*, **53**, 240–255.
20. Emsley,P. and Cowtan,K. (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr.*, **60**, 2126–2132.
21. Ball,N., Streeter,S., Kneale,G.G. and McGeehan,J. (2009) Structure of the restriction-modification controller protein C.Esp1396I. *Acta Crystallogr.*, **D65**, 900–905.
22. Lavery,R., Moakher,M., Maddocks,J.H., Petkeviciute,D. and Zakrzewska,K. (2009) Conformational analysis of nucleic acids revisited: curves+. *Nucleic Acids Res.*, **37**, 5917–5929.
23. Olson,W.K., Bansal,M., Burley,S.K., Dickerson,R.E., Gerstein,M., Harvey,S.C., Heinemann,U., Lu,X.-J., Neidle,S., Shakked,Z. *et al.* (2001) A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol.*, **313**, 229–237.
24. Ptashne,M. (2004) *A Genetic Switch: Phage Lambda Revisited*, 3rd edn. Cold Spring Harbor, New York.
25. Brennan,R.G., Roderick,S.L., Takeda,Y. and Matthews,B.W. (1990) Protein-DNA conformational changes in the crystal structure of a lambda Cro-operator complex. *Proc. Natl Acad. Sci. USA*, **87**, 8165–8169.
26. Albright,R.A. and Matthews,B.W. (1998) Crystal structure of lambda-Cro bound to a consensus operator at 3.0 A resolution. *J. Mol. Biol*, **280**, 137–151.
27. Schumacher,M.A., Miller,M.C., Grkovic,S., Brown,M.H., Skurray,R.A. and Brennan,R.G. (2002) Structural basis for cooperative DNA binding by two dimers of the multidrug-binding protein QacR. *EMBO J*, **21**, 1210–1218.
28. Beamer,L.J. and Pabo,C.O. (1992) Refined 1.8 A crystal structure of the lambda repressor-operator complex. *J. Mol. Biol.*, **227**, 177–196.
29. Jain,D., Nickels,B.E., Sun,L., Hochschild,A. and Darst,S.A. (2004) Structure of a ternary transcription activation complex. *Mol. Cell*, **13**, 45–53.
30. Sorokin,V., Severinov,K. and Gelfand,M.S. (2009) Systematic prediction of control proteins and their DNA binding sites. *Nucleic Acids Res.*, **37**, 441–451.