Nucleic Acids Research Advance Access published August 31, 2012

*Nucleic Acids Research, 2012, 1–11*
*doi:10.1093/nar/gks718*

# The structural basis of differential DNA sequence recognition by restriction–modification controller proteins

**N. J. Ball, J. E. McGeehan, 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 PO1 2DY, UK

Received June 6, 2012; Revised July 1, 2012; Accepted July 3, 2012



## ABSTRACT

**Controller (C) proteins regulate the expression of restriction–modification (RM) genes in a wide variety of RM systems. However, the RM system Esp1396I is of particular interest as the C protein regulates both the restriction endonuclease (R) gene and the methyltransferase (M) gene. The mechanism of this finely tuned genetic switch depends on differential binding affinities for the promoters controlling the R and M genes, which in turn depends on differential DNA sequence recognition and the ability to recognize dual symmetries. We report here the crystal structure of the C protein bound to the M promoter, and compare the binding affinities for each operator sequence by surface plasmon resonance. Comparison of the structure of the transcriptional repression complex at the M promoter with that of the transcriptional activation complex at the R promoter shows how subtle changes in protein–DNA interactions, underpinned by small conformational changes in the protein, can explain the molecular basis of differential regulation of gene expression.**

## INTRODUCTION

Restriction–modification (RM) systems protect bacteria from invasion by bacteriophage and may play a role in restricting the flow of genetic information in bacterial populations (1, 2). RM systems encode a restriction endonuclease (ENase) and a DNA methyltransferase (MTase) that recognize the same DNA sequence. The DNA MTase protects the host DNA from cleavage by the associated restriction enzyme, while digesting (restricting) foreign DNA (2). There are a variety of control mechanisms that ensure the correct temporal expression of RM genes, to ensure that the host DNA is methylated prior to exposure to the ENase.

The best known of these mechanisms employs a 'controller' (C) protein encoded by a gene downstream of its own promoter, and co-transcribed with the restriction 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). The time-dependence of the activity of this switch has been demonstrated *in vitro*, and ENase expression was shown to be delayed with respect to the MTase when the C protein is expressed in a new host *in vivo* (9,10).

In typical C-protein systems, the operator sequence at the C/R promoter has two operator sites (denoted $O_L$ and $O_R$) (11,12). $O_L$ is distal to the gene and has a high affinity for a C-protein dimer. When bound at this site, the σ subunit of RNA polymerase is recruited and both the C and R genes are switched on. $O_R$ is a much weaker binding site proximal to the gene; however, when a C-protein dimer is bound to $O_L$ then the affinity for $O_R$ is greatly increased and at high protein concentrations, this site is occupied and the gene is down-regulated (12–14). In the RM system Esp1396I, the C protein also represses the constitutively expressed methyltransferase (M) gene by binding as a dimer to the promoter that overlaps the transcriptional start site of this gene (15). The C/R genes and the M gene in this system are transcribed convergently from different promoters (See Figure 1).

Analysis of C-protein binding sites in a wide variety of RM systems suggested a repeating quasi-symmetrical consensus sequence consisting of two sets of inverted repeats or 'C-boxes' [$GACT(N_3)AGTC(N_4)GACT(N_3)AGTC$] upstream of the C/R genes (6,8,12). However, the degree of sequence homology between species is moderate and the internal symmetry within and between 'C-boxes' is also weak in most C/R promoters (16).

*To whom correspondence should be addressed. Tel: +44 23 9284 2678; Fax: +44 23 9284 2053; Email: geoff.kneale@port.ac.uk

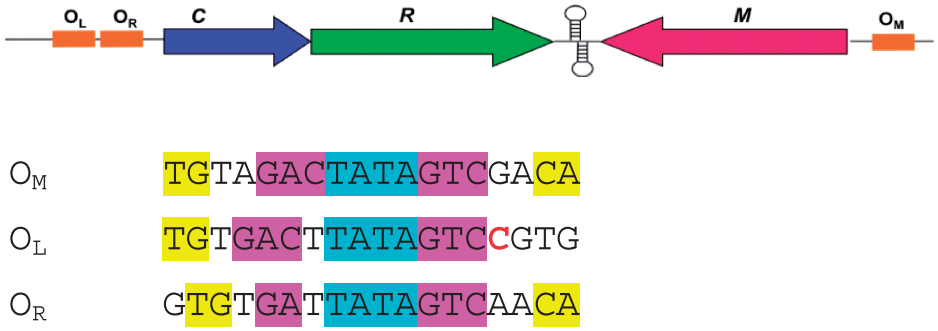


**Figure 1.** Regulation of restriction (R) and modification (M) genes by C.Esp1396I. The upper figure shows convergent gene organization and the location of the three operator sites: $O_M$, $O_L$ and $O_R$. The sequences of these sites are shown below, with the specific recognition motifs shown in magenta and yellow, and the central TATA in cyan. The C implicated in a possible interaction with D34 is indicated in red. Adapted from Bogdanova *et al.* (15).

Moreover, the proposed 3-bp 'spacers' within the left and right operator sequences are also largely conserved between species, the consensus sequence being TAT. However, subsequent structural studies of C-protein–DNA complexes suggest that the binding site may be better described as a 4-bp alternating pyrimidine–purine spacer (e.g. TATA) separating two tri-nucleotide recognition sites, rather than a 3-bp spacer separating two 4-bp recognition sequences (11,17).

The first published structure of a C protein bound to DNA was that of C.Esp1396I bound as a tetramer, with two dimers bound adjacently on the 35-bp operator sequence ($O_L + O_R$) of the C/R promoter (11). The structure revealed the mechanism whereby cooperative binding of dimers to the DNA operator control the switch from activation to repression of the C and R genes. In the crystal structure of the complex (PDB code: 3CLC), two dimers are bound to the DNA, each centred on the pseudo-dyad located between the central A and T bases in the TATA sequence within each operator site, and interacting across the major groove at the centre of the DNA.

Subsequent high resolution crystallographic studies of the complex with the $O_L$ operator (17) showed more clearly the nature of the sequence specific contacts to the bases within the recognition site ('direct readout'), as well as the non-specific interactions with the severely bent phosphodiester backbone ('indirect readout'). We now report the structure of a dimer of C.Esp1396I bound to $O_M$ and investigate the affinities of the protein for its three natural promoters, $O_M$, $O_R$ and $O_L$, in order to understand the structural and mechanistic basis of differential DNA sequence recognition that underpins this elegant genetic switch.

## MATERIALS AND METHODS

### Purification

Large-scale cultures of *Escherichia coli* BL21(DE3) containing the plasmid pET–28b/*esp1396IC* were grown. Over-expressed C.Esp1396I containing an N-terminal hexa-histidine tag (C.Esp1396I-6His) was harvested by sonication and separated from the cell lysate using nickel affinity chromatography. The His-tag was removed by thrombin digestion but the purified protein retained a GSH tripeptide (C.Esp1396I-GSH). Size exclusion chromatography was performed on a 26/60 Sephacryl S-200 HR size exclusion column in order to separate C.Esp1396I-GSH from cleaved His-tag, uncleaved protein and thrombin. For structural studies and for biophysical analysis, the protein was concentrated using heparin affinity chromatography. The DNA oligonucleotides were purified as previously described and annealed to form a duplex, prior to complex formation (11).

### Analytical ultracentrifugation

Sedimentation equilibrium experiments were performed at 20°C with a range of protein concentrations using an Optima XL-A analytical ultracentrifuge (Beckman–Coulter, Palo Alto, CA, USA). Preliminary studies were done at 28 000 r.p.m. covering the range 1–30 μM protein. Subsequent runs were carried out at rotor speeds of 15 000, 21 000 and 28 000 r.p.m. Scans were done at wavelengths of 225 and 280 nm with a radial step size of 0.01 mm after 21 h equilibration. The scans for 1, 5 and 10 μM protein were globally fitted to a self-association model using SEDPHAT to determine the dissociation constant for the dimer ($K_{dim}$). The values for partial specific volume and buffer density were calculated using SEDNTERP and the errors were estimated using F-statistics. The $K_{dim}$ was used to calculate the dimer concentration [D] in a sample of known total protein concentration, $P_T$, using the following relationship:

$$[D] = 0.125 \times \{4.P_T + K_{dim} \pm \sqrt{[K_{dim}(K_{dim} + 8.P_T)]}\} \quad (1)$$

### Surface plasmon resonance

5' biotinylated synthetic oligonucleotides containing either the $O_M$, $O_L$, $O_R$ or both the $O_L$ and $O_R$ sequences ($O_{L+R}$) were immobilized on the surface of a SA sensor chip on a Biacore T-100. C.Esp1396I-GSH was dialyzed against the running buffer (10 mM HEPES pH 7.4, 100 mM NaCl, 5 mM $MgCl_2$, 5 mM $CaCl_2$, 0.05% v/v Tween-20) before a range of concentrations were injected over the chip for 30 s at a flow rate of 30 μl/min. Kinetic analysis was performed using the 1:1 binding model (with mass-transfer correction enabled) provided in the BiaEval software

(version 2.0.2). For the kinetic analysis, the protein concentration was adjusted to the actual dimer concentration using Equation (1). Equilibrium analysis was performed by fitting the data to either a one-site model:

$$R = R_{max.}[D]/(K_D+[D]) \quad (2)$$

or (for the $O_{L+R}$ data) a two-site model:

$$R = R_{max.}(1+2[D]/K_{D2}).([D]/K_{D1})/ 2\{1+(1+[D]/K_{D2}).([D]/K_{D1})\} \quad (3)$$

using GraFit version 5.0.11 (Erithacus Software Ltd.). R is the response generated on reaching equilibrium, $R_{max}$ is the maximum response that can be generated by saturating the binding sites on the immobilized ligand, $K_D$ is the dissociation constant for the interaction (with $K_{D1}$ and $K_{D2}$ denoting the dissociation constants for the interaction between $O_L$ and $O_R$, respectively) and the dimer concentration, [D], is given by Equation (1).

### Crystallization of complexes

The DNA containing $O_M$ was designed to promote the crystallization of the complex in a single orientation in a similar manner to the $O_L$ complex. The DNA consisted of an 18-bp duplex with 5' overhangs of A on one strand and T on the other. C.Esp1396I was incubated with the DNA at varying ratios (1:1, 1.5:1, 2:1, 2.5:1, 3:1 and 4:1 protein monomer:DNA) prior to crystal screening. The protein–DNA complex was subjected to sparse matrix screening using the Honeybee robot (Digilabs) to set up sitting drops. Subsequent crystallizations of protein–DNA complex were done at a ratio of 2:1 (protein monomers:DNA) with a final DNA concentration of ~20 μM. Sitting drops were set up using 2 μl complex and 2 μl of the well solution. The initial conditions were optimized by varying the pH from 7 to 8.5 (in 0.5 unit increments), while simultaneously varying the PEG 1500 concentration from 5 to 30% w/v (in 5% increments). The trays were incubated at 16°C and checked at regular intervals using polarizing light microscopy. Suitable crystals were cryoprotected in 30% v/v glycerol, cryocooled in liquid nitrogen and stored, prior to exposure to synchrotron radiation. The crystals that gave rise to the final $O_M$ structure formed in 0.1 M SPG (succinate/phosphate/glycine) buffer pH 8, 25% w/v PEG 1500 with spermidine at a final concentration of 10 μM in the drop.

### Structure solution and refinement

Cryocooled crystals of the $O_M$ complex were exposed to synchrotron radiation on ID14-4 at the ESRF (Grenoble). A selection of crystals was screened using the automated sample changer and data sets were collected at 100 K using an ADSC 4Q CCD detector. The $O_M$ complex crystallized in space group $P2_1$ and 180 images were collected with an oscillation angle of 1°. The data were processed and scaled using MOSFLM/SCALA (18) as this provided better integration statistics than processing the data using XDS/XSCALE (19). The collection and refinement statistics are shown in Table 1.

**Table 1.** Crystallographic parameters

| | |
|---|---|
| Data collection | |
| Space group | $P2_1$ |
| Unit-cell parameters (Å, °) | $a$ = 47.5 |
| | $b$ = 147.1 |
| | $c$ = 47.8 |
| | $\alpha = \gamma = 90$ |
| | $\beta$ = 93.7 |
| Resolution limits (Å) | 45.36–2.7 (2.85–2.7) |
| $R_{merge}$[a] (%) | 6.6 (20.1) |
| $I/\sigma(I)$ | 7.4 (3.8) |
| Completeness (%) | 98.9 (99.7) |
| Refinement parameters | |
| NCS | |
| Groups | 1 |
| Chains in group | A, B, E and F |
| Residue range | 5–75 |
| Restraint level | Tight |
| TLS | |
| Groups | 10 |
| Chains (residues) | A, C, D, F, G and H (1–79) |
| | B and E (1–41,48–79) |
| | B and E (42–47) |
| Refinement model statistics | |
| No. of reflections | 61 350 |
| $R_{cryst}/R_{free}$[b] (%) | 19.6/23.7 |
| No. of atoms | |
| Protein | 2496 |
| DNA | 1546 |
| Water | 12 |
| Average B factors (Å$^2$) | |
| Protein | 31.8 |
| DNA | 35.6 |
| Water | 34.8 |
| RMS deviations from ideal | |
| Bond lengths (Å) | 0.015 |
| Angles (°) | 2.1 |

X-ray crystal data, refinement and model statistics for the $O_M$ complex structure. Values in parentheses are for the highest resolution shell.
[a]$R_{merge} = \Sigma_{hkl}\Sigma_i|I_i(hkl)-\langle\langle I(hkl)\rangle\rangle|/\Sigma_{hkl}\Sigma_i I_i(hkl)$, where $\langle\langle I(hkl)\rangle\rangle$ is the mean intensity of reflection $I(hkl)$ and $I_i(hkl)$ is the intensity of an individual measurement of reflection $I(hkl)$.
[b]$R_{cryst} = \Sigma_{hkl}||F_{obs}|-|F_{calc}||/\Sigma_{hkl}|F_{obs}|$, where $F_{obs}$ is the observed structure factor amplitude and $Fcalc$ is the calculated structure factor amplitude. $R$free is the same as $R$cryst but for 5% of structure factor amplitudes that were set aside during refinement.

The scaled data were phased by molecular replacement using Phaser (20). Chains A and B along with 10 bp from the $O_L$ structure (chains C and D) were used as separate ensembles to search for a replacement solution. The $O_M$ structure contained two complexes (i.e. two dimers, each bound to a DNA duplex) in the asymmetric unit. The structure was refined to 2.7 Å using iterative cycles of REFMAC5 (21) and real-space refinement in COOT (22). Non-crystallographic symmetry (NCS) and TLS restraints were used in REFMAC5 and the missing bases were manually added into interpretable electron density using COOT. The restraints used in refinement are shown in Table 1. Solvent atoms were added manually in COOT. 5% of structure-factor amplitudes were set aside during refinement for $R_{free}$ calculations. The final structure refined with $R/R_{free}$ = 19.6/23.7% and contained all 76 DNA bases (38 per duplex) and the following amino acid residues: 2–77 (chain A), 2–78 (chain B), 1–77 (chain E) and 4–78 (chain D); 99.7% of amino acid

residues were in the preferred region of the Ramachandran plot. The coordinates of the DNA–protein complex have been deposited in the protein databank (PDB code: 3UFD). Molecular structures were visualized with Pymol (23). Amino acid residue numbers refer to the native sequence; the tripeptide sequence remaining after removal of the affinity tag was not observed in the electron density map as, presumably, it is disordered.

## RESULTS

The C protein C.Esp1396I was expressed and purified as described previously (24; see also 'Materials and Methods' section). Structural analysis of the interaction of the protein with the M operator was then undertaken by means of X-ray crystallography and the interaction was further characterized in solution by analytical ultracentrifugation (AUC). Surface plasmon resonance (SPR) was then performed to compare binding affinities of the protein for each of the three natural operator sites.

### Crystallographic analysis of the DNA–protein complex

DNA–protein complexes were formed with an 18-bp DNA duplex consisting of two 19-bases oligonucleotides (thus forming 5' A/T overhangs). This sequence contains the MTase gene operator sequence ($O_M$) and 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). Optimum crystallization conditions for the complex were determined from trials based on the PACT screen (Molecular Dimensions). X-ray diffraction data from suitable crystals were collected at 100 K at the ESRF (Grenoble). The space group was determined as $P2_1$, with two independent protein–DNA complexes in the asymmetric unit. The structure was solved by molecular replacement and refined by iterative cycles of reciprocal space refinement (REFMAC5) and real space refinement (COOT) to 2.7 Å resolution (see Table 1). Chains A–D comprise one DNA–protein complex, where A and B refer to the two subunits of a protein dimer, and C and D to the two strands of the DNA duplex (Figure 2). Chains E–H comprise the second DNA–protein complex in the asymmetric unit (E and F corresponding to the protein dimer, G and H to the DNA duplex).

The non-symmetric bases were identifiable during the building of the DNA duplex and thus the orientation of the DNA was defined. In particular, the purine/pyrimidine (A5/T16) and the pyrimidine/purine (C5′/G16′) base pairs could be distinguished (Figure 2). The terminal A and T bases could also be distinguished in the map, thus confirming the orientation of the DNA duplex. In contrast to the structure of the $O_L$ complex, where Hoogsteen base pairs are involved in the interaction between adjacent duplexes (17), the two terminal bases in the $O_M$ complex form Watson–Crick base pairs between duplexes, resulting in end-to-end packing of the DNA (Figure 2). In addition, R43 and K17 side-chains from adjacent complexes are involved in packing interactions that are mediated by an anion, most probably chloride (Supplementary Figure S1). Representative electron density in the map is illustrated in the vicinity of the dimer interface and around a region of the DNA (Supplementary Figure S2).

During the initial stages of refinement, the flexible loop regions (residues 43–46) were not subject to NCS restraints, since there are two stable conformations available for this loop in the free protein (24). However, after the initial refinement, the electron density maps were sufficiently clear to see that all four subunits had the flexible loop in the same conformation. Thus, subsequent rounds of refinement were carried out with tight NCS restraints also applied to the flexible loop region. Subsequent models therefore refer to the structure of a single complex (chains A–D). Although NCS restraints were not applied to the two DNA duplexes in the asymmetric unit, subsequent analysis shows that the two DNA helices have almost identical structures (see below).

### DNA structure in the complex

The DNA conformation in the nucleoprotein complex was analysed using the online CURVES server (25). The local DNA bend angle and the compression of the minor groove in the two complexes in the asymmetric unit is illustrated in Figure 3. The DNA helices in both complexes exhibit an overall bend angle of 56°. Additionally, both complexes show a very similar degree of local bending and minor groove compression at equivalent base pairs, despite the fact that no NCS restraints were applied to the DNA during refinement. The minor groove width varies from ~10 Å to ~2 Å, being most severely compressed at the TATA sequence. The compression of the minor groove is accompanied by an increased local bend angle.

The DNA in the $O_M$ complex has a higher overall bend angle (56°) than the DNA in the $O_L$ complex (41°), possibly reflecting the decreased spacing of the conserved elements in the sequence in $O_M$. In the $O_L$ complex, the GAC/GTC sequences are separated by 5 bp and are positioned non-symmetrically relative to the TATA sequence (Figure 1). In the $O_M$ complex, the C-boxes are separated by 4 bp and are positioned symmetrically around the TATA sequence.

The compression of the minor groove is achieved through interactions between the phosphate backbone of the DNA and the amino acid residues of C.Esp1396I (Supplementary Figure S3). Residues D34, Y37, T49, S52 and N47 play a critical role in the compression of the phosphodiester backbone around the TATA sequence. Equivalent residues from each monomer interact with the backbone of a DNA strand either side of the TATA site and the distances between these residues in the two monomers determine the angle by which the DNA is bent. There are additional protein–DNA backbone contacts on the opposite strand that stabilize the complex, notably from amino acid residues R17, Q24, S39, R43 and N44 to the DNA phosphate backbone around the conserved TG nucleotides.

### DNA sequence recognition

Direct readout of the $O_M$ operator DNA sequence is accomplished via the sidechains of the amino acid residues

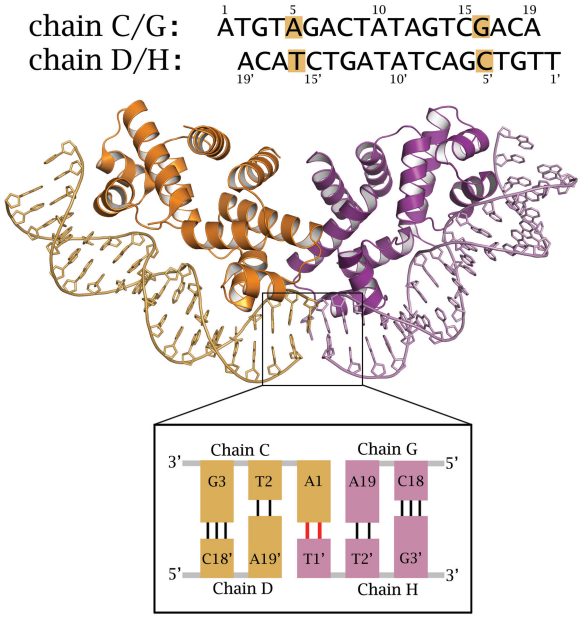


**Figure 2.** Structure of the two nucleoprotein complexes in the asymmetric unit of the C.Esp1396I/$O_M$ complex. Top: The sequence of the two DNA chains highlights the non-symmetric base pairs (AT and CG). Bottom: The two DNA duplexes in each complex are held together by an AT base pair formed from the 5' overhanging bases.

R35, T36 and R46 (Figure 4), which interact with the GAC/GTC motifs. In fact, all the contacts to this motif are made to the GTC bases on one strand. The γ-hydroxyl of the T36 sidechain interacts with the N4 amino group of cytosine $C_{15}$. The R46 sidechain interacts with the N7 of $G_{13}$. The interaction of the second NH of the guanidinium group with the carbonyl oxygen (O4) of $G_{13}$ appears to be mediated through a water molecule. Likewise, there is a water molecule in a position to mediate the interaction of the R46 guanidinium group with the carbonyl oxygen of the thymine base, $T_{14}$.

The R35 sidechain is involved in both direct and indirect readout of the DNA sequence in the $O_M$ complex, as was found in both the $O_L$ complex and the tetrameric repression complex structures. The planar guanidinium head group of the arginine forms hydrogen bonds with the N7 and O6 of $G_3$ but it is also involved in a π-stacking interaction with the adjacent base, $T_2$ (Supplementary Figure S4). However, in contrast to the $O_L$ structure, the interaction of R35 with the TG motif is equivalent for both subunits, reflecting the symmetry of the DNA sequence at the M operator. It should be noted that the R35 from a given subunit (e.g. subunit A) interacts with different DNA strands when recognizing the GTC motif (on strand D) and the TG motif (on strand C). These interactions will further strengthen the integrity of the dimer in the nucleoprotein complex.

From comparison of C protein sequences and their cognate DNA binding sequences, it has been shown that there is a correlation between the identity of an amino acid residue in the recognition helix and the base sequence of the operator that it binds (26); specifically, it has been proposed that an aspartate at position 34 (or its equivalent in other C proteins) correlates with a cytosine base being present at the 3' side of one of the GTC motifs, whereas a histidine at this site is most often found when there is a thymine at this site in the DNA sequence. C.Esp1396I belongs to the former category (i.e. possessing a DRTY rather than an HRTY motif in the recognition helix). We see no interaction of D34 with this base in the complex with the $O_M$ or the $O_L$ operator; instead, the D34 sidechain contacts the phosphodiester backbone of the DNA (see Supplementary Figure S3). There may conceivably be 'indirect' interactions to the base via a solvent molecule (although none is visible in the crystal structure).

Are there any other clues to a possible structural/biological role for D34? Somewhat surprisingly, the correlation observed only applies to the second of the four 'C-boxes' in the promoters studied [box 1B in the nomenclature of Mruk *et al.* (26)] and not to box 1A, where the symmetry related subunit of the dimer binds at the $O_L$ site (and nor does it apply to either site in $O_R$). We also note that of the three DNA sequences that C.Esp1396I binds, each has a different base (G, C, A) at the site that has been proposed to interact with D34 (Figure 1). Indeed, the strongest binding site ($O_M$) has a G at this site, a clear exception to the observations of Mruk *et al.* (26). Thus a direct role for D34 in binding to an isolated DNA operator site is unlikely.

However, we have previously shown that the C-protein subunit bound to box 1B of $O_L$ is involved in cooperative binding to the adjacent subunit bound to box 2A, at the interface between the two dimers in the tetrameric repression complex (11). Since the DRTY correlation with a cytosine base is only found at the second of the four repeating elements in the C/R promoter, we are tempted to speculate that D34 may play a role in repression at the promoter. The adjacent residue, R35, of this subunit interacts with E25 of the adjacent subunit via an ion pair mechanism in the tetrameric (repression) complex, and is a major contributor to the observed cooperativity between the two sites (11). The R35 of the adjacent subunit, however, binds to the G of the highly conserved central GT motif between $O_L$ and $O_R$. It is possible that D34 may play an as yet unidentified role in that complex network of interactions, perhaps also involving the cytosine on the 3' side of the GTC motif. If so, then presumably a histidine in the HRTY motif could make an equivalent interaction with a thymine at that site, to explain the observations of Mruk *et al.*(26).

## Hydrodynamic analysis

The dissociation constant ($K_{dim}$) for the monomer–dimer equilibrium is an important parameter in the operation of the genetic switch, especially at low levels of expression of the C protein. Moreover, an accurate value of $K_{dim}$ needs to be determined experimentally in order to obtain the relevant DNA binding constants. Thus, in order to obtain the DNA binding affinities of C.Esp1396I for its various operators, we first analysed the monomer–dimer equilibrium of the protein by sedimentation equilibrium in the AUC.

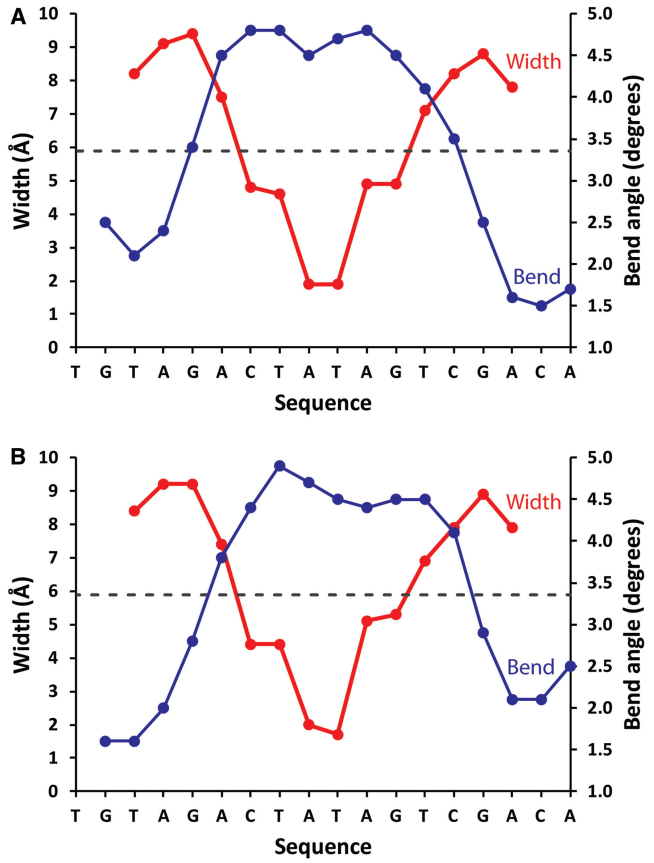


**Figure 3.** Analysis of DNA structure in the two DNA–protein complexes in the asymmetric unit, showing the local bend angle and groove width at each base pair.

Since the protein has only one tyrosine, its extinction coefficient is too low to allow accurate determination of the $K_{dim,}$ when low concentrations of protein are required. We therefore mutated the tyrosine residue Y29 into a tryptophan by site-directed mutagenesis. The mutation was confirmed by DNA sequencing of the gene, and the presence of a tryptophan could also be deduced from the fluorescence emission spectrum of the purified protein. Y29 is located far from the dimerization interface, and does not participate in DNA binding since it lies at the C-terminal end of helix 2. From dynamic light scattering, the hydrodynamic radius of the Y29W mutant protein (2.4 nm) was indistinguishable from that of the native protein, and its DNA binding properties were also found to be unchanged (data not shown).

The absorbance scans of the Y29W mutant of C.Esp1396I in the concentration range 1–30 μM were analysed using a single species model in SEDPHAT in order to determine the weight average molecular weight (Supplementary Figure S5). At low concentrations, the molecular weight was determined to be 8.8 ± 1.2 kDa, in agreement with the theoretical mass of a C.Esp1396I monomer (9.5 kDa). At higher concentrations (>10 μM), the molecular weight was found to be 19.4 ± 0.5 kDa, corresponding to the expected mass of a C.Esp1396I dimer. Thus, the $K_{dim}$ for the monomer–dimer equilibrium is within the range of 1–10 μM.

A more accurate equilibrium constant was then determined by globally fitting the absorption scans measured at three different concentrations (1, 5 and 10 μM) and three different rotor speeds (15, 21 and 28 k.r.p.m.) to a self-associating species model (27) using SEDPHAT (see Supplementary Figure S6). This yielded a value for the $K_{dim}$ of 1.6 μM corresponding to a free energy of dimerization (at 20°C) of −32.5 kJ/mol. The dimerization constant is of the same order of magnitude as that for C.AhdI ($K_{dim}$ = 2.5 μM), consistent with the surface areas of their respective dimer interfaces (~1900 Å$^2$ versus ~1400 Å$^2$) and the similar H-bonding interactions between monomers in each case (14,24).

## DNA binding analysis

(SPR experiments were conducted to investigate the DNA binding affinities of C.Esp1396I for the relevant promoter sites. Four different biotinylated duplexes containing either $O_M$, $O_L$, $O_R$ or the double site ($O_{L+R}$) were each immobilized on separate streptavadin chips. For each experiment, the C.Esp1396I protein was injected using a range of concentrations, and the response measured as a function of time. The range of protein concentrations required to elicit a significant response for each DNA sequence varied greatly (up to 50-fold), reflecting the variation in DNA binding affinity at each site.

Following injection of the protein, the sensorgrams quickly reached their maximum response, and then remained constant throughout the 30 s injection (Figure 5). At this point the rates of binding and dissociation are equal and equilibrium has been attained. It is possible to obtain $K_D$ for the interaction by plotting the equilibrium response against protein concentration and fitting to the relevant binding equation. However, since C.Esp1396I binds as a dimer, the concentration of the active dimer must first be determined, since the total protein concentration is, in some cases, below the $K_{dim}$. This can be estimated using the dimer dissociation constant of 1.6 μM determined by AUC (see 'Materials and Methods' section). The analysis assumes that the monomer–dimer equilibrium is not affected by the small amount of protein dimer that binds to DNA during the experiment; for the relatively low loadings of DNA immobilized on the surface, this is likely to be a valid approximation.

For the individual operator sites, the standard single-site binding equation was used to determine the dissociation constant of the dimer–DNA interaction, $K_D$ (Figure 6). For the duplex containing $O_{L+R}$, a 2-site model was used with dissociation constants, $K_{D1}$ and $K_{D2}$, where $K_{D1}$ describes binding to $O_L$ and $K_{D2}$ describes binding to $O_R$. By determining the affinity of C.Esp1396I for $O_L$ in isolation, $K_{D1}$ can be fixed, which permits an accurate determination of the affinity for the second site ($O_R$) when $O_L$ is already occupied.

From the result of the equilibrium analysis, C.Esp1396I has the highest affinity for $O_M$ ($K_D$ = 0.61 nM), intermediate affinity for $O_L$ ($K_D$ = 5.6 nM) and the lowest affinity for $O_R$ ($K_D$ = 120 nM). Once the $O_L$ site has been occupied by a C.Esp1396I dimer, however, the affinity between C.Esp1396I and $O_R$ increases ~130-fold

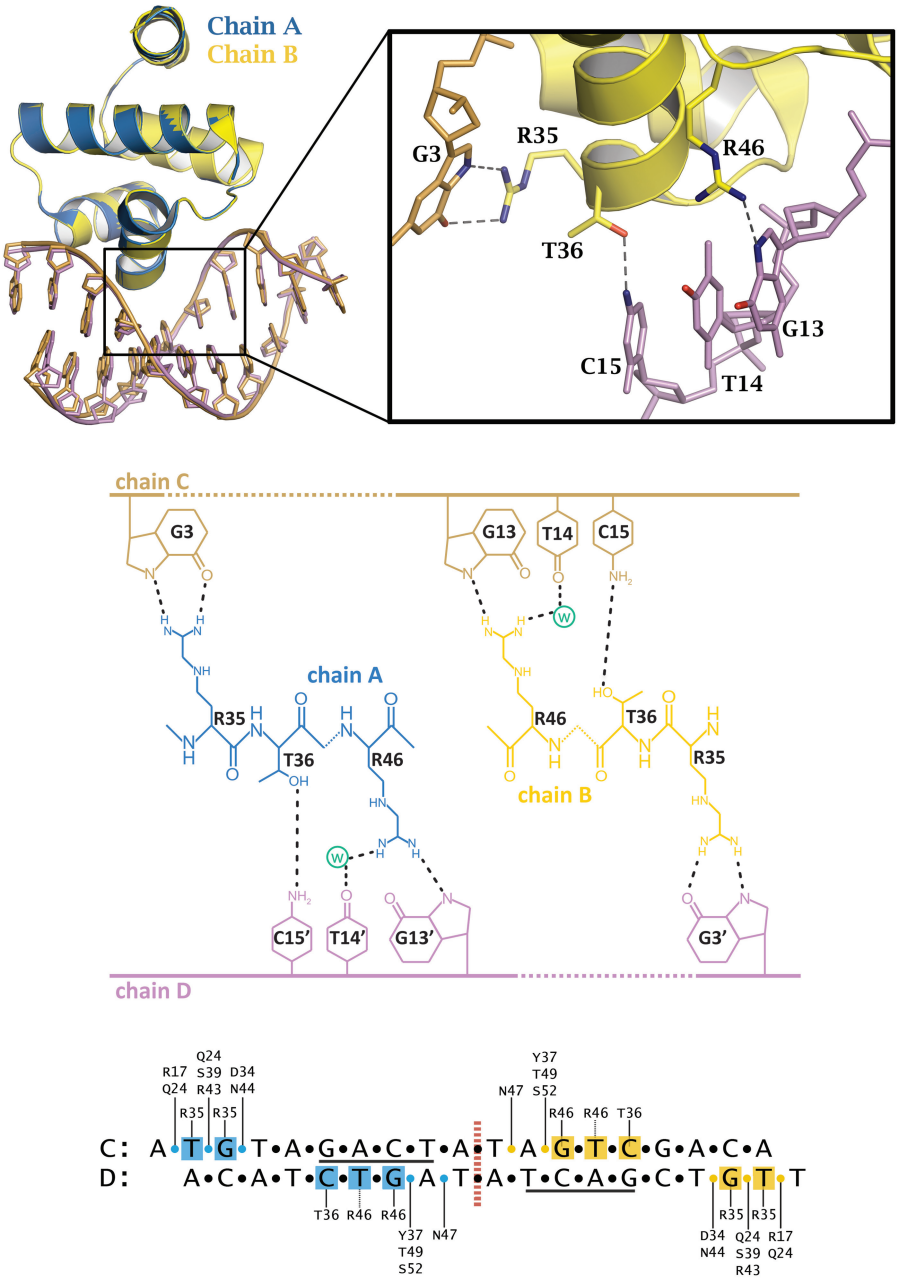


**Figure 4.** DNA–protein contacts. Top: Rotation and superposition of the two subunits of the complex show symmetrical interactions to the DNA (inset: interactions of amino acids R35, T36 and R46 with bases $G_3$ on one strand and $G_{13}$, $T_{14}$ and $C_{15}$ on the other; the water atom is omitted for clarity). Middle: Schematic representation of the hydrogen bonding contacts. Bottom: Overview of specific base contacts and contacts to the DNA phosphates (yellow and blue circles).

($K_{D2}$ = 0.94 nM), indicating that there is a very high cooperativity of binding between the two operator sites.

The on- and off-rates of the interaction can also be measured by kinetic analysis of the sensorgrams (Figure 5), except for the case of two-site binding to $O_{L+R}$, which cannot be described by any of the available binding models. From the ratio of the on- and off-rates, the binding constants for the three single operator sites can be obtained (See Table 2). The $K_D$ obtained by equilibrium measurements were generally higher than those obtained from kinetic analysis, but they were of the same magnitude, and were in approximately the same ratio (200:8:1 for $O_R$:$O_L$:$O_M$). Thus, the SPR experiments show that the affinity for the $O_M$ site is around 8-fold higher than that for $O_L$ which, in turn, is 25-fold higher than that for $O_R$.

## DISCUSSION

Overall, the structure of the C.Esp1396I/$O_M$ complex resembles that determined for the $O_L$ complex at the C/R promoter (17). Crucially, however, there are key structural

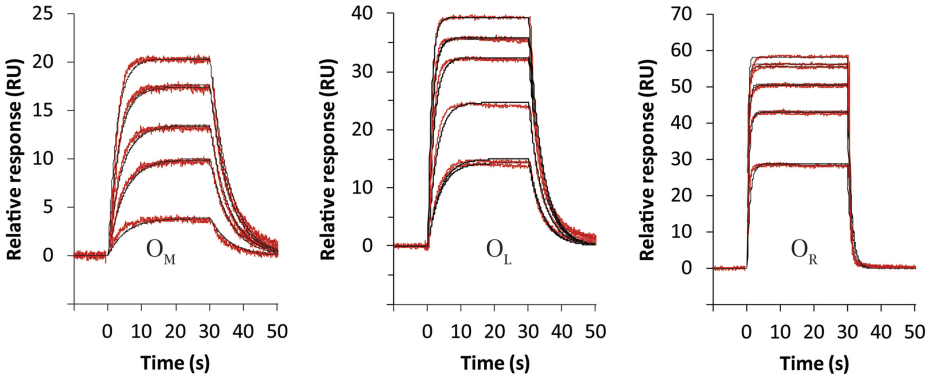


**Figure 5.** SPR kinetic analysis. For each operator site, the C protein was injected into the sample channel at five different protein concentrations (in duplicate), and the responses recorded after subtraction from the reference channel. Data were fitted to obtain the on- and off-rates for the interaction (see Table 1).

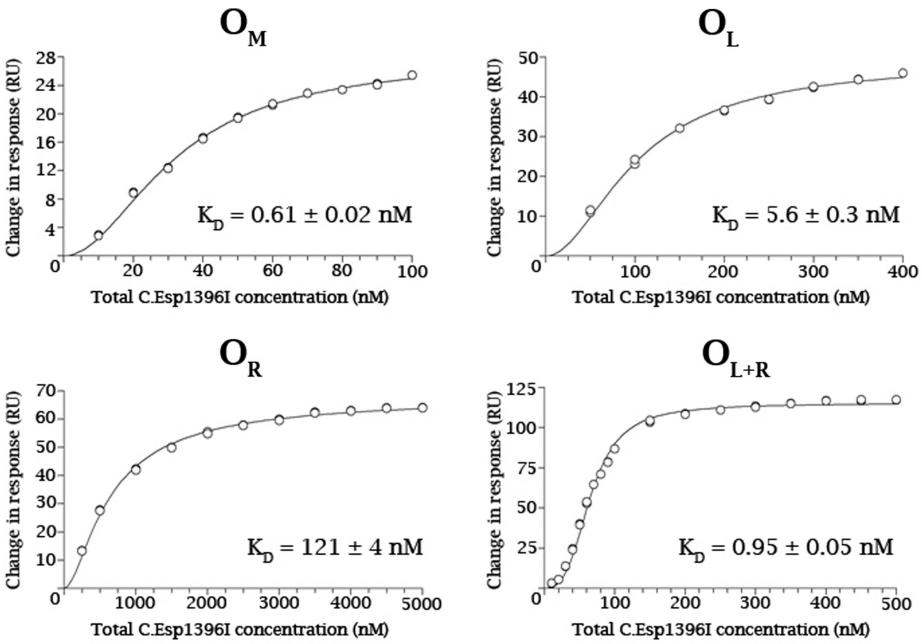


**Figure 6.** Equilibrium binding analysis. Equilibrium binding at saturation was plotted against total protein concentration (expressed as monomer) from the SPR data shown in Figure 5. For $O_M$, $O_L$ and $O_R$, the curves were fitted to a single-site binding model to obtain the relevant dissociation constants, $K_D$. For $O_L + O_R$, a two site binding model was employed; the $K_d$ for binding to $O_L$ was fixed at the value obtained experimentally for the isolated operator site, thus permitting the determination of the affinity for the second site ($O_R$) when $O_L$ is already occupied.

**Table 2.** Rate constants from kinetic analysis of the SPR data for C.Esp1396I binding to the three operator sites, $O_M$, $O_L$ and $O_R$

| | $k_a$ ($M^{-1}.s^{-1}$) | $k_d$ ($s^{-1}$) | $K_D$ (nM) |
|---|---|---|---|
| $O_M$ | $1.61 \pm 0.01 \times 10^8$ | 0.177 ± 0.001 | 1.10 ± 0.01 |
| $O_L$ | $2.99 \pm 0.02 \times 10^7$ | 0.254 ± 0.001 | 8.5 ± 0.1 |
| $O_R$ | $3.88 \pm 0.05 \times 10^6$ | 0.887 ± 0.004 | 229 ± 4 |

The equilibrium dissociation constants, $K_D$, were in each case determined from the ratio of the off-rate ($k_a$) to the on-rate ($k_d$).

differences that determine the differential DNA binding affinity for the endonuclease and M promoters. Figure 7 shows the superposition of the $O_M$ and $O_L$ complexes (RMSD of 0.36 Å). The majority of backbone and sidechain positions are essentially identical (Figure 7a), the exception being the conformation of the flexible loop (residues 43–46) of one of the two subunits of the dimer, which differs significantly between the two complexes (Figure 7b).

The DNA bend in both complexes is centred on the alternating pyrimidine–purine sequence, TATA. Either side of this, in both complexes, the GAC motif (or more specifically, the complementary sequence GTC) is recognized by hydrogen bonding interactions with amino acid residues T36 and R46. In the $O_M$ complex, this motif is symmetrically disposed, 2 bp either side of the dyad axis within the central TATA, so that the centre of the GAC (=GTC) motifs are separated by 7 bp. However, in the $O_L$ complex, these motifs are asymmetrically arranged, 2 and

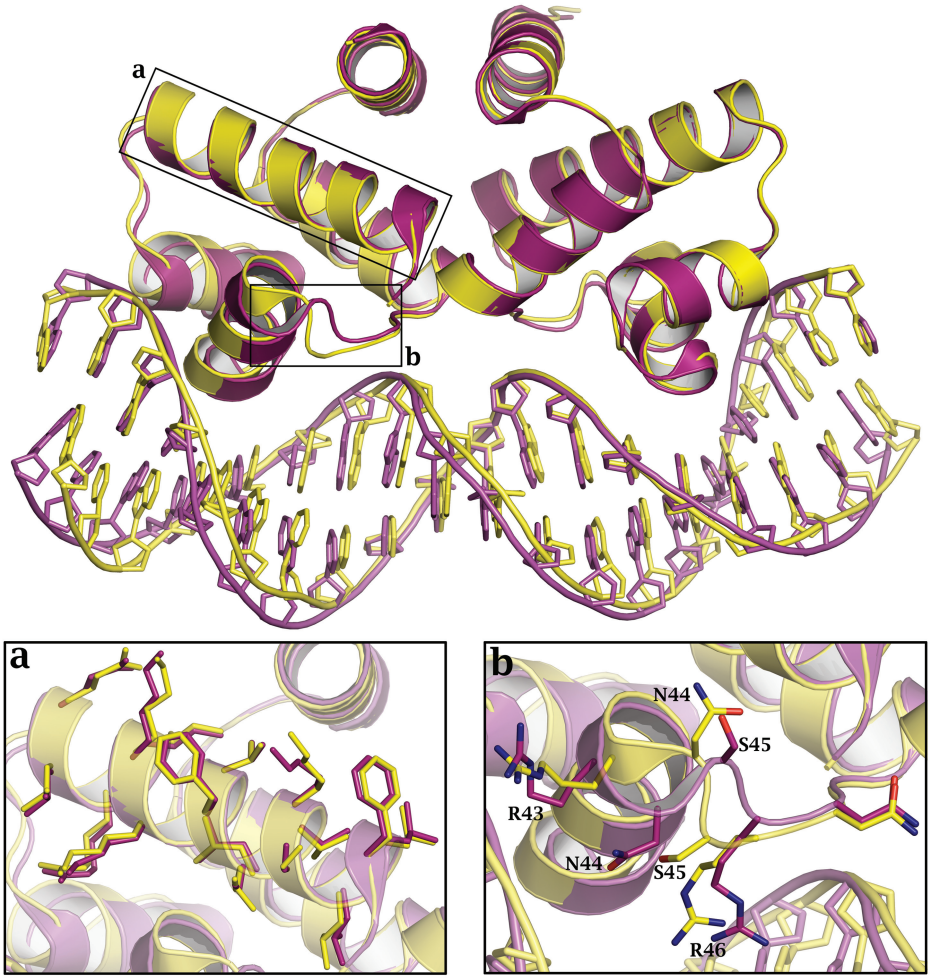


**Figure 7.** Comparison of the structures of complexes of C.Esp1396I bound to the operators $O_L$ and $O_M$ (yellow and magenta, respectively) showing the displacement of the DNA bases. Although the sidechains of the alpha-helices in the two complexes are superimposed (**a**) the loop regions (**b**) are in quite different conformations, resulting in large displacements of amino acid side chains of N44 and S45, together with a smaller movement of R46.

3 bp respectively from the pseudo-dyad axis, leading to an 8-bp separation between their centres (see Figure 1).

This additional separation of ca. 3.4 Å between these sites forces a conformational change in one of the subunits of the $O_L$ complex, in order to accommodate the displacement of the GTC motifs. It is notable that the overall position of the alpha-helices remains the same when compared with the $O_M$ complex; however, a localized conformational change in the flexible loop region (amino acid residues 43–46), leads to a significant rotation of the R46 sidechain that contacts the GTC motif.

The $O_M$ sequence is almost perfectly symmetrical, differing by only 1 bp (the A:T at position 5 is a C:G at the equivalent position on the other strand—see Figure 2). However, there are no contacts from the protein to the DNA at this position. The TG/CA and the GAC/GTC sequences are symmetrically arranged, and make specific hydrogen bonds to each subunit of the protein (including one via a water molecule). The TATA sequence does not make sequence-specific H-bonds, but instead makes numerous interactions with the protein via the phosphate groups of the DNA backbone to stabilize the highly deformed DNA helix at this point—a form of 'indirect' sequence read-out.

In comparison, the $O_L$ sequence lacks one of the TG motifs (see Figure 1), and thus loses a strong interaction with R35 (including two charged H-bonds and a base stacking interaction). Although $O_L$ has the GAC/GTC motif that is recognized by the protein, the extra base 'insertion' requires a conformational change in one of the protein subunits. Together, these changes in DNA sequence reduce the binding affinity by a factor of ~8. We have previously shown that mutation of R35 to alanine abrogates binding to the $O_{L+R}$ operator, since in this case interactions with two TG motifs are lost (11).

There is no structure available for the $O_R$ complex, but such a complex would most likely lose the interaction with one of the TG motifs (Figure 1), unless the change in spacing can be accommodated by a conformational change, which itself would add an energy penalty. In addition, one of the GAC motifs becomes a GAT, thus losing the interaction with R46 on one subunit. Compared

to $O_M$, these two alterations to the DNA sequence, taken together, reduce the binding affinity by a factor of ~200.

The order in which C.Esp1396I binds to its operator sequences is vital for the temporal regulation of the RM system. This is determined initially by the relative affinities of the C protein for the $O_L$ and $O_M$ binding sites, and subsequently by the cooperativity between $O_L$ and $O_R$ binding sites at the C/R promoter. Initially, C.Esp1396I is expressed at a low level from a weak C-independent promoter. The M gene (*esp1396IM*) is expressed constitutively, allowing the host genome to be methylated and thus protected from the action of the restriction enzyme. As the C.Esp1396I concentration slowly increases, protein dimers are formed. Initially, C-protein dimers bind to the highest affinity site $O_M$ and down-regulate the expression of *esp1396IM*, where the binding site overlaps the start of transcription (15). Subsequently, C-protein dimers bind to $O_L$, up-regulating transcription from the C/R operon (*esp1396IC/R*) through cooperative recruitment of RNA polymerase, leading to a positive feedback loop. Thus the concentration of C.Esp1396I dimers will increase exponentially. At these higher concentrations, a further C-protein dimer binds cooperatively to the $O_R$ site to displace RNA polymerase, resulting in a negative feedback loop as the expression of *esp1396IC/R* is down-regulated. Ultimately, when both C/R and M promoters are repressed, the levels of C protein will fall, leading to de-repression of the M gene and thus enabling transcription of the M gene. Further regulation at the level of translation may also be involved, adding an additional level of fine tuning of the genetic switch.

The transcriptional regulation of the RM genes is ultimately dependent upon a localized conformational change in the C protein that is confined to a few amino acids residues in the loop region between helices 3 and 4. This conformational change is sufficient to allow variations in the spacing between specific DNA sequences of the 'C-box' motifs (specifically the trinucleotide sequence GAC/GTC) in relation to the TATA sequence that defines the centre of the bend in the DNA. There is, however, a free energy penalty to pay, as is evident from the 200-fold variation in DNA binding affinities between the three operator sites. In the $O_M$ promoter complex, there is almost perfect dyad symmetry within the C-protein dimer, matching a similar symmetry in the DNA sequence. In contrast, the shift in the pseudo-dyad axis relating the C-boxes in $O_L$ forces a conformational change in the loop of one subunit of the protein dimer, thus breaking the symmetry, and contributing to an almost 10-fold decrease in binding affinity, compared to the symmetrical binding site, $O_M$.

This subtle change in the conformation of the protein underpins the differential affinity for the respective operator sites and controls the order in which the RM genes are switched on and off. The correct balance between methylation and restriction is thereby maintained, thus ensuring that the integrity of the bacterial genome is not compromised by premature expression of the endonuclease, while at the same time ensuring that DNA methyltransferase activity is kept in check.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figures 1–6.

## 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) [BB/E000878/1 to G.G.K. and BB/H00680X/1 to G.G.K. and J.E.M.]; Research Councils UK Academic Fellowship (to J.E.M.); University of Portsmouth IBBS PhD studentship (to N.J.B.). Funding for open access charge: BBSRC.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Jeltsch,A. (2003) Maintenance of species identity and controlling speciation of bacteria: a new function for restriction/ modification systems? *Gene*, **317**, 13–16.
2. Wilson,G.G. and Murray,N.E. (1991) Restriction and modification systems. *Ann. Rev. Genet.*, **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. 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.
15. Bogdanova,E., Zakharova,M., Streeter,S., Taylor,J., Heyduk,T., Kneale,G.G. and Severinov,K. (2009) Transcription regulation of restriction-modification system Esp1396I. *Nucleic Acids Res.*, **37**, 3354–3366.
16. 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.
17. McGeehan,J., Ball,N.J., Streeter,S.D., Thresh,S.-J. and Kneale,G.G. (2012) Recognition of dual symmetry by the controller protein C.Esp1396I based on the structure of the transcriptional activation complex. *Nucleic. Acids Res.*, **40**, 4158–4167.
18. Leslie,A.G.W. (1992) Recent changes to the MOSFLM package for processing film and image plate data. *Joint CCP4 + ESF-EAMCB Newsletter on Protein Crystallography*, No. 26.
19. Kabsch,W. (1993) Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J. Appl. Crystallogr.*, **26**, 795–800.
20. McCoy,A.J., Grosse-Kunstleve,R.W., Storoni,L.C. and Read,R.J. (2005) Likelihood-enhanced fast translation functions. *Acta Crystallogr. D: Biol. Crystallogr.*, **61**, 458–464.
21. Murshudov,G.N., Vagin,A.A. and Dodson,E.J. (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D: Biol. Crystallogr.*, **53**, 240–255.
22. Emsley,P. and Cowtan,K. (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr. D: Biol. Crystallogr.*, **60**, 2126–2132.
23. Delano,W.L. (2002) *The PyMOL Molecular Graphics System*. DeLano Scientific, San Carlos, CA, USA.
24. Ball,N., Streeter,S., Kneale,G.G. and McGeehan,J. (2009) Structure of the restriction-modification controller protein C.Esp1396I. *Acta Crystallogr. D Biol. Crystallogr.*, **D65**, 900–905.
25. 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.
26. Mruk,I., Rajesh,P. and Blumenthal,R.M. (2007) Regulatory circuit based on autogenous activation-repression: roles of C-boxes and spacer sequences in control of the PvuII restriction-modification system. *Nucleic Acids Res.*, **35**, 6935–6952.
27. Vistica,J., Dam,J., Balbo,A., Yikilmaz,E., Mariuzza,R.A., Rouault,T.A. and Schuck,P. (2004) Sedimentation equilibrium analysis of protein interactions with global implicit mass conservation constraints and systematic noise decomposition. *Anal. Biochem.*, **326**, 234–256.