Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Structure of the restriction–modification controller protein C.*Esp*1396I

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

Biophysics Laboratories, Institute of Biomedical and Biomolecular Sciences, School of Biological Sciences, University of Portsmouth, Portsmouth PO1 2DY, England

Correspondence e-mail:
geoff.kneale@port.ac.uk,
john.mcgeehan@port.ac.uk

Received 6 May 2009
Accepted 29 May 2009

PDB References: C.*Esp*1396I, 3g5g, r3g5gsf; R35A mutant, 3fya, r3fyasf.

The controller protein of the *Esp*1396I restriction–modification (R–M) system binds differentially to three distinct operator sequences upstream of the methyltransferase (M) and endonuclease (R) genes to regulate the timing of gene expression. The crystal structure of a complex of the protein with two adjacent operator DNA sequences has been reported; however, the structure of the free protein has not yet been determined. Here, the crystal structure of the free protein is reported, with seven dimers in the asymmetric unit. Two of the 14 monomers show an alternative conformation to the major conformer in which the side chains of residues 43–46 in the loop region flanking the DNA-recognition helix are displaced by up to 10 Å. It is proposed that the adoption of these two conformational states may play a role in DNA-sequence promiscuity. The two alternative conformations are also found in the R35A mutant structure, which is otherwise identical to the native protein. Comparison of the free and bound protein structures shows a 1.4 Å displacement of the recognition helices when the dimer is bound to its DNA target.

1. Introduction

Bacterial restriction–modification (R–M) systems act as a primitive ‘immune system’ that serves to protect the cell from invasion by foreign DNA. R–M systems encode a restriction endonuclease and a DNA methyltransferase. The DNA-sequence-specific methyltransferase (M) protects the host DNA from cleavage by the associated restriction enzyme (R); the specific methylation pattern of the host R–M system allows the discrimination of ‘self’ from ‘nonself’ DNA (Wilson & Murray, 1991). The expression of M and R genes must be subject to temporal control such that restriction activity is delayed with respect to methylation, so that the bacterial genome can be methylated and thus protected from the possibility of subsequent endonuclease activity. This delay is often accomplished by means of a regulator (or controller) protein, the C protein, which is required for effective transcription of its own gene and for transcription of the endonuclease (R) gene found on the same operon (Tao *et al.*, 1991; Ives *et al.*, 1992; Rimseliene *et al.*, 1995; Vijesurier *et al.*, 2000; Cesnaviciene *et al.*, 2003; Knowle *et al.*, 2005).

Measurements of C-dependent transcriptional activity *in vitro* have shown the time-dependence of the activity of this switch (Bogdanova *et al.*, 2008) and *in vivo* experiments have directly demonstrated a time lag in the expression of the endonuclease with respect to the methyltransferase (Mruk & Blumenthal, 2008). Previous biochemical and biophysical studies have revealed the general features of this genetic switch in the *Ahd*I R–M system (Streeter *et al.*, 2004; McGeehan *et al.*, 2006; Papapanagiotou *et al.*, 2007). In the

doi:10.1107/S0907444909020514

Table 1
X-ray crystal data, refinement and model statistics.

Values in parentheses are for the highest resolution shells.

	Native	R35A
Data collection		
Space group	$P6_5$	$P6_5$
Unit-cell parameters (Å, °)	$a = b = 128.72$, $c = 137.51$, $\alpha = \beta = 90$, $\gamma = 120$	$a = b = 48.44$, $c = 135.78$, $\alpha = \beta = 90$, $\gamma = 120$
Resolution limits (Å)	50–2.8 (2.9–2.8)	50–3.0 (3.1–3.0)
$R_{\rm merge}$† (%)	13.8 (39.0)	27.7 (45.5)
$I/\sigma(I)$	16.8 (7.7)	1.4 (1.9)
Completeness (%)	96.3 (93.7)	100 (100)
Refinement model statistics		
No. of reflections	30225	3617
$R_{\rm cryst}/R_{\rm free}$‡ (%)	23.7/26.9	24.9/26.7
No. of atoms		
Protein	8696	1197
Water	4	0
B factors (Å^2)		
Protein	33.09	56.74
Water	13.58	n/a
R.m.s. deviations from ideal		
Bond lengths (Å)	0.015	0.014
Angles (°)	1.579	1.690

† $R_{\rm merge} = \sum_{hkl}\sum_i |I_i(hkl) - \langle I(hkl)\rangle| / \sum_{hkl}\sum_i I_i(hkl)$, where $\langle I(hkl)\rangle$ is the mean intensity of reflection $I(hkl)$ and $I_i(hkl)$ is the intensity of an individual measurement of reflection $I(hkl)$. ‡ $R_{\rm cryst} = \sum_{hkl} ||F_{\rm obs}| - |F_{\rm calc}|| / \sum_{hkl} |F_{\rm obs}|$, where $F_{\rm obs}$ is the observed structure-factor amplitude and $F_{\rm calc}$ is the calculated structure-factor amplitude. $R_{\rm free}$ is the same as $R_{\rm cryst}$ but for 5% of structure-factor amplitudes which were set aside during refinement.

R–M system *Esp*1396I, it has recently been shown that the C protein has an additional function of binding to a high-affinity site upstream of the M gene to repress expression of the methyltransferase, thus forming part of an intricate control network in the regulation of R–M activity (Bogdanova *et al.*, 2009).

The first controller-protein structure to be reported was that of C.*Ahd*I (McGeehan *et al.*, 2004, 2005), which revealed a dimeric α-helical protein with a helix–turn–helix motif. A similar structure was subsequently reported for C.*Bcl*I (Sawaya *et al.*, 2005). More recently, we reported the first structure of a controller protein (C.*Esp*1396I) bound to its DNA operator site (McGeehan *et al.*, 2008), in which two dimers were bound adjacently on the DNA to form a tetrameric complex. To enable comparison of the free and bound forms of the protein and to identify possible conformational changes when bound to DNA, we have crystallized and solved the structure of the native C.*Esp*1396I protein dimer. The results show that there is conformational heterogeneity of a small loop region adjacent to the DNA-sequence recognition helix. In addition, we report that the structure of a mutant of C.*Esp*1396I in which a key residue (Arg35) has been mutated shows a similar heterogeneity.

2. Materials and methods

2.1. Crystallization

The expression and purification of native C.*Esp*1396I was performed as described previously (McGeehan *et al.*, 2008), with the exception that the six-histidine tag was retained. Briefly, the protein was overexpressed from plasmid pET-28b/*esp1396IC* in *Escherichia coli* BL21 (DE3). The cells were then harvested and disrupted by sonication. Cell lysates were applied onto a His-Trap HP column in high-salt buffer [500 m*M* NaCl, 20 m*M* imidazole, 40 m*M* Tris–HCl pH 8, 5%(*w*/*v*) glycerol] and eluted by increasing the imidazole concentration to 500 m*M* in a step gradient. Fractions were pooled and placed in Spectrapor dialysis membrane (3000 Da cutoff) and dialysed against 5 l low-salt buffer [150 m*M* NaCl, 20 m*M* imidazole, 40 m*M* Tris–HCl pH 8.0, 5%(*w*/*v*) glycerol, 2.5 m*M* $CaCl_2$] at 277 K. A precipitate formed within the first 2 h of dialysis and was stored in a minimal amount of dialysate at 277 K prior to cryocooling. Site-directed mutagenesis was performed as described previously (McGeehan *et al.*, 2005) and the R35A mutant protein was expressed, purified and crystallized identically.

2.2. X-ray diffraction data collection and structure determination

The crystals obtained following dialysis were collected and transferred to cryoprotectant solution (30% glycerol) prior to cryocooling in liquid nitrogen. For the native protein crystal, 180 images with an oscillation width of 1.0° were collected at a wavelength of 0.933 Å, while 90 1.0° images were collected for the mutant protein crystal owing to high anisotropy. Data extending to 2.8 Å were collected at 100 K on beamline ID14-2 (ESRF, Grenoble) using an ADSC Q4 CCD detector and processing was performed with either *XDS* and *XSCALE* (Kabsch, 1993) or *MOSFLM* (Leslie, 1992) and *SCALA* (Collaborative Computational Project, Number 4, 1994). Molecular replacement was carried out using *Phaser* (McCoy *et al.*, 2005) and the structures were refined with reiterative rounds of model building using *Coot* (Emsley & Cowtan,

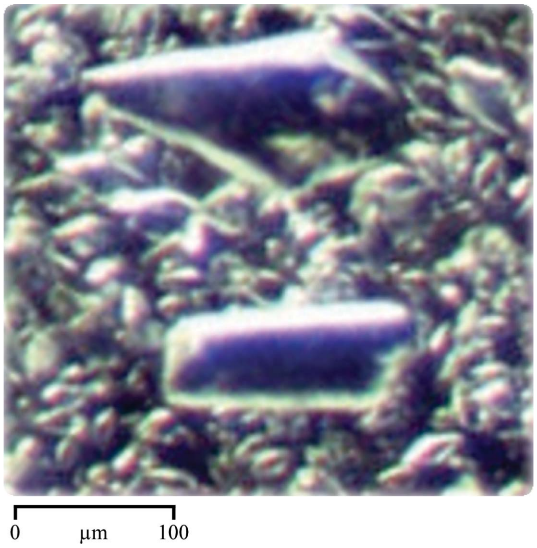

Figure 1
Light-microscope image showing typical precipitate crystals formed during dialysis. Single crystals with an approximate largest dimension of 100 μm can be seen among many small and fragmented crystals.

2004) and TLS-based refinement using *REFMAC5* (Murshudov *et al.*, 1997). NCS restraints were used throughout, except in the loop region consisting of residues 43–46. Analysis of the structural data was performed with *SFCHECK* and *PROCHECK* (Collaborative Computational Project, Number 4, 1994) and the figures were produced with *PyMOL* (DeLano, 2002).

The native and mutant protein structures were deposited in the PDB with codes 3g5g and 3fya, respectively.

3. Results and discussion

3.1. Crystallization and data collection

Traditional sitting-drop and hanging-drop vapour-diffusion methods of crystallizing C.*Esp*1396I resulted in poorly diffracting and often split crystals. Through serendipity, whilst purifying the protein it was noticed that a precipitate formed during routine dialysis of the His-tagged protein and this was identified as crystalline in nature by polarizing light microscopy (Leica MZ12-5). Single crystals were observed with largest dimensions of approximately 200 × 75 × 75 µm. Harvesting of these crystals (Fig. 1) followed by X-ray screening in-house (Xcalibur Nova, Oxford Diffraction) provided diffraction data that were sufficient for indexing.

Both the native and R35A mutant proteins crystallized in space group $P6_5$, although their crystals had different unit-cell parameters (Table 1). There was a strong tendency for splitting of these crystals and, in addition to the presence of secondary lattices, reflections were often smeared, resulting in poor data quality. Following extensive screening using the ESRF/EMBL SC3 sample changer, complete data were collected from both native and mutant C.*Esp*1396I crystals, one from each of the representative unit cells. The native data suffered from poor spot profiles and proved difficult to process with *MOSFLM*. *XDS* produced much improved integration statistics, although despite strong intensities [overall $I/\sigma(I)$ of 17] $R_{\rm merge}$ remained high at 14%. The mutant data were much weaker and also suffered from poor spot profiles. Processing with *XDS* resulted in rejection of a high number of reflections, compromising the overall completeness. *MOSFLM* was able to process these mutant data with high completeness; however, this resulted in high $R_{\rm merge}$ values. Following multiple processing runs for both the native and the mutant data sets, the best compromise between data quality and completeness was achieved using *XDS*/*XSCALE* for the native data and *MOSFLM*/*SCALA* for the mutant data. The data-collection and processing statistics for both proteins are shown in Table 1.

3.2. Structure of native C.*Esp*1396I protein

The native structure was solved by molecular replacement using a single monomer (chain *A*) from the previously solved nucleoprotein structure (PDB code 3clc; McGeehan *et al.*, 2008) and yielded seven independent dimers in the asymmetric unit. Despite the difficulties in processing these data, the electron-density maps were of good quality (Fig. 2) and refinement proceeded smoothly. Over 99% of the residues lie in the preferred regions of the Ramachandran plot, with no outliers, and the $R_{\rm cryst}/R_{\rm free}$ values and bond geometries are reasonable for the resolution cutoff of 2.8 Å (Table 1).

As expected, the overall structure resembles that of the protein in complex with DNA: a compact fold comprising five α-helices per monomer, each with a characteristic helix–turn–helix motif. Resolution limitations prevent a comprehensive comparison of side-chain conformations between the free and bound proteins. However, the large-scale rearrangements of the structure are clear. Superposition of the *A* chains of the bound and the free proteins reveals a global movement of the recognition helix 3 of approximately 1.4 Å (Fig. 3). This hinge action is centred on the dimer interface mediated by helix 5 and is consistent with an opening of the dimer upon DNA binding. The distance between the recognition helices of the dimer is increased by approximately 1–1.5 Å when bound to

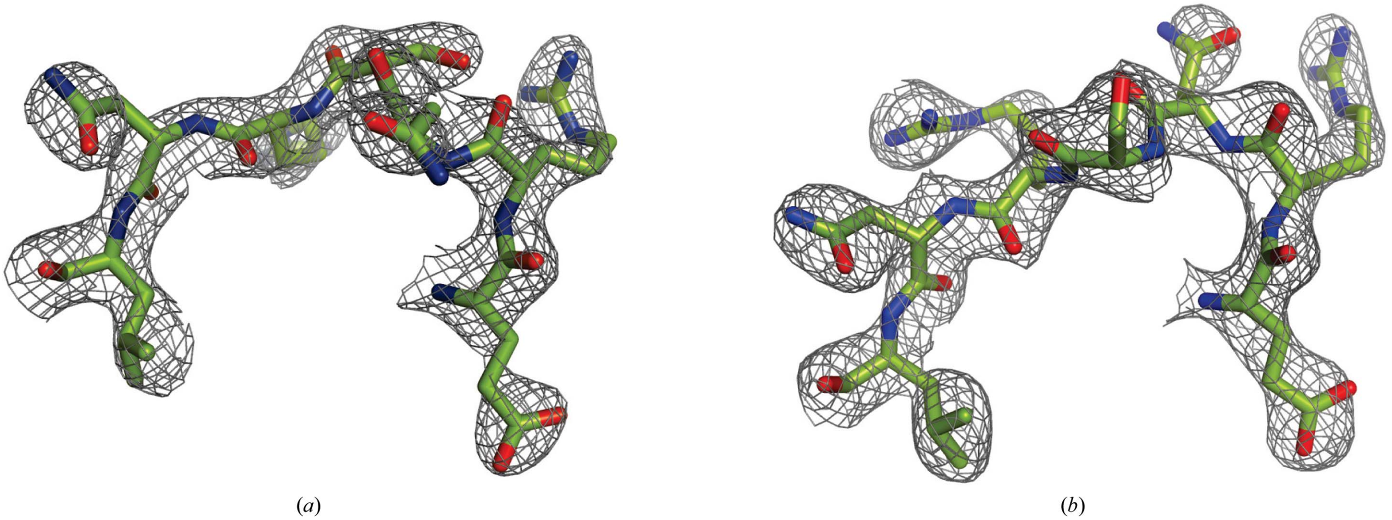

Figure 2
Representative density of the two alternate loop regions in native C.*Esp*1396I. Residues 42–48 are shown for (*a*) chain *A* and (*b*) chain *N*. The $2F_{\rm o} - F_{\rm c}$ electron-density maps are contoured at 1.5σ.

DNA, which may contribute to the significant 54° bend of the operator DNA (McGeehan *et al.*, 2008).

3.3. Flexibility within the loop region may contribute to DNA-sequence recognition

Within the 14 copies of the monomer found in the asymmetric unit of the native protein, there are two that exhibit variation in a loop structure close to the DNA-binding interface (residues 43–46; Fig. 4*a*). Release of NCS restraints in this region allowed the building and refinement of an alternative loop conformation into clearly interpretable electron density (Fig. 2*b*). Compared with the other 12 monomers, which all show remarkably consistent density in this region, the minor conformation present in chains *L* and *N* adopts a markedly different backbone path. Residues from each of the two conformations occupy favoured regions of the Ramachandran plot and a schematic analysis of temperature factors (Fig. 5) demonstrates *B*-factor values that are similar to those of other loop regions in the structure. In fact, although the major loop conformation has *B* factors that are around 14% higher (38 Å^2) than the average for the whole structure (33 Å^2), the *B* factors for residues 43–46 in chains *L* and *N* are 2–3% lower, suggesting increased rigidity afforded by packing interactions. Although higher resolution data are required to gain an accurate refinement of the temperature factors, it appears from these data that there is clear static disorder between two alternative loop conformations in these crystals.

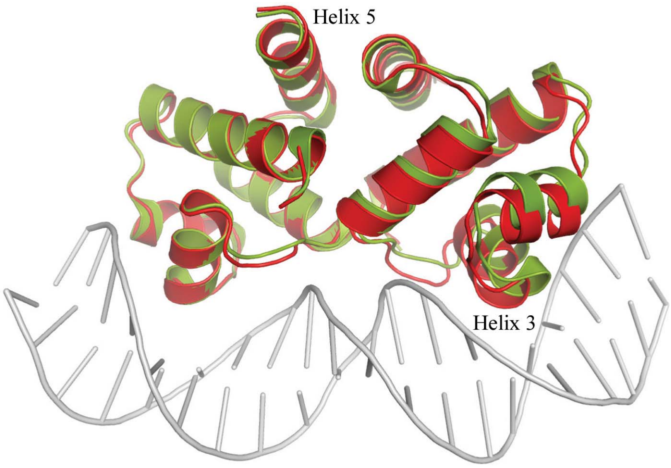

Figure 3
Comparison of free and bound C.*Esp*1396I structures. Chains *A* and *B* of the free protein (red) and chains *A* and *B* of the DNA-bound protein (green; PDB code 3clc) were overlaid, giving an r.m.s.d. of 0.39 Å (201 main-chain atoms). The recognition helix moves 1.4 Å upon DNA binding as calculated by taking the average distances between corresponding C$^\alpha$ atoms within the recognition helix.

In the alternative loop conformation, individual C$^\alpha$ atoms are displaced by up to ~5 Å with significant side-chain rearrangement (Fig. 4*b*). This is evident for the side chain of Arg46, where the NH2 group is displaced by 2.7 Å between the two conformations, and more dramatically for residue Asn44, where a rotation of approximately 180° about the C$^\alpha$ atom results in a maximum displacement of the terminal O atoms of around 10 Å. This alternative conformation appears only twice in 14 chains in the native protein and on both of these occasions residue Arg43 is involved in intermolecular hydrogen bonding to an adjacent monomer (residue Asp64). Presumably, the free-energy difference between the two conformations is small and reflects the local environment afforded by specific packing interactions. It is possible that this

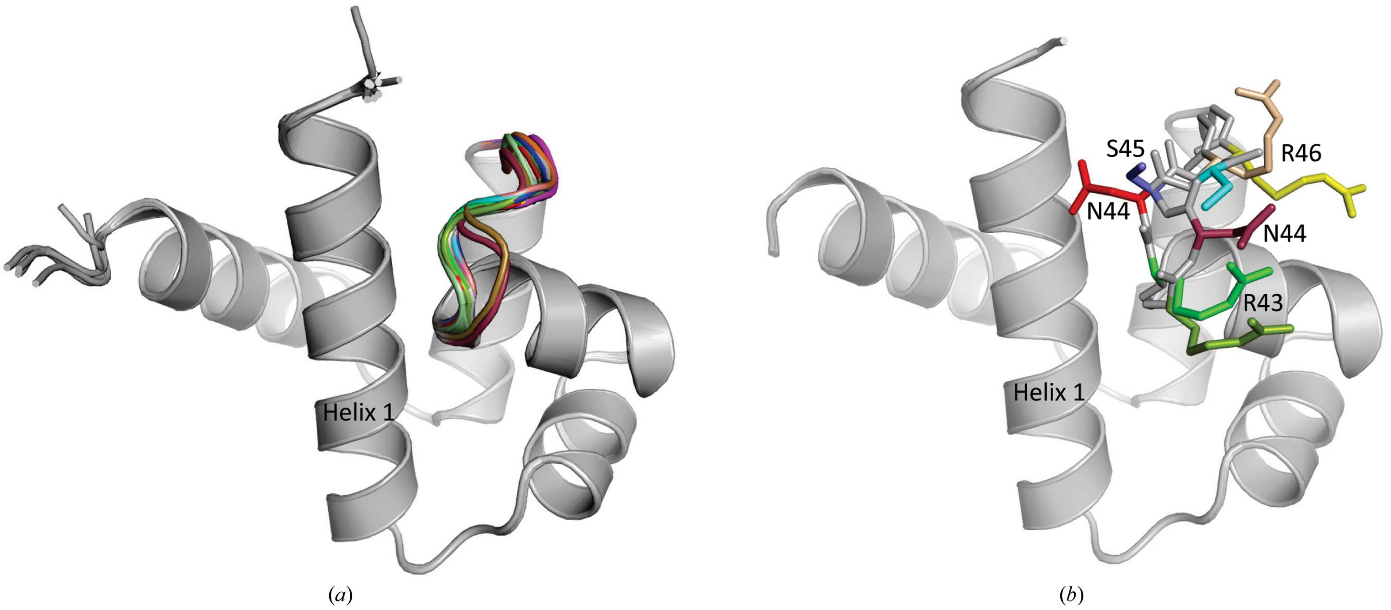

Figure 4
The native protein has two distinct loop conformations. (*a*) All 14 monomers from the asymmetric unit have been overlaid (r.m.s.d. < 1 Å for 180 main-chain atoms excluding residues proximal to the loop). (*b*) Chain *A* and chain *N* superimposed, highlighting the degree of side-chain movement between the alternative loop regions (r.m.s.d. = 0.095 Å for 180 main-chain atoms excluding residues proximal to the loop). Chain *A* is shown in light colours and chain *N* in dark colours; the side chains of Arg46, Ser45, Asn44 and Arg43 are shown in yellow, blue, red and green, respectively.

region contributes to DNA binding and sequence-specificity: this loop is located immediately adjacent to the recognition helix 3, which inserts into the major groove of DNA. It is likely that a direct readout mechanism operates to discriminate between subtly different operator sequences. A degree of flexibility in an extended recognition motif could help to explain some of the features of this control system at the biological level. C.*Esp*1396I has three cognate operator sites, binding each site with a high degree of specificity and a range of affinities over several orders of magnitude (Bogdanova *et al.*, 2009). In order to achieve this biologically important discrimination, it is likely that the DNA-binding interface incorporates a degree of structural flexibility. Whether this loop region imparts the necessary degree of plasticity to explain the binding mechanism of multiple recognition sites will hopefully be revealed by high-resolution nucleoprotein studies.

3.4. Solution and comparison of R35A mutant structure

Previous EMSA studies on this C protein indicated that the single amino-acid change R35A completely abolished binding to a cognate 35 base-pair operator site (McGeehan *et al.*, 2008). The C.*Esp*1396I protein–DNA structure revealed that Arg35 is a strong hydrogen-bonding partner to a conserved guanine located in the DNA-recognition sequence and also stacks with the adjacent thymine base. However, full interpretation of the EMSA binding data hinges on the structural consequences of this mutation, since it could simply destabilize the structure of the protein. The structural integrity of this mutant protein was therefore explored by crystallography.

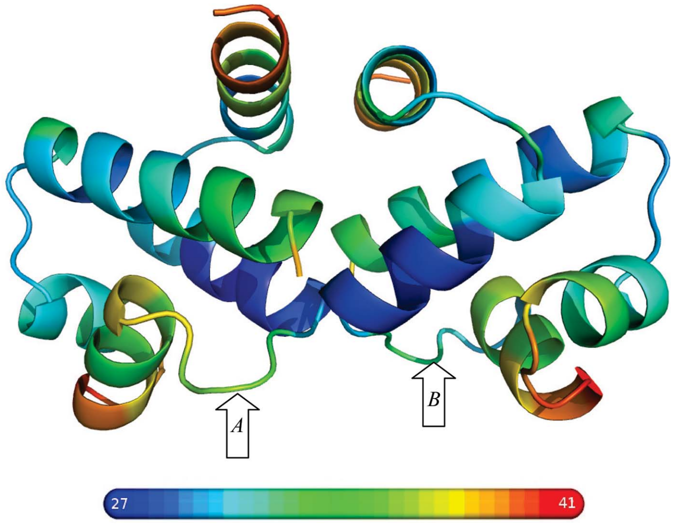

Figure 5
Temperature-factor analysis of a native dimer. Chains *K* and *L* of the native free protein are shown as cartoons, where blue represents low and orange high *B* factors (C$^{\alpha}$ atoms). A colour scale bar illustrates the *B*-factor range from 27 to 41 Å^2. The C-terminal residues have the highest *B* factors (around 20% higher than the mean) in addition to the N-terminal regions of the recognition helices. The core of the protein dimer, including the N-terminal half of helix 4 (shown in dark blue), has lower than average *B* factors as expected. This dimer has both alternative loop conformations, which are highlighted with arrows. The major conformation (*A*) has slightly higher *B* factors compared with the minor conformation (*B*), although both are close to the average for all residues.

The R35A mutant protein crystals contained only one dimer in the asymmetric unit. However, analysis of symmetry-related chains revealed a packing arrangement resembling that of the native protein: a small rotation and translation of adjacent dimers relative to the native crystal packing resulted in a significantly smaller unit cell. It is possible that a degree of heterogeneity between these two crystal forms compounded our search for a single discrete lattice in earlier experiments. Extensive screening of these crystals was productive for the collection of complete data, although the overall intensities were weak [$I/\sigma(I)$ of 1.4; Table 1]. Despite this, following molecular replacement using the C.*Esp*1396I monomer as a search model clear alternative backbone paths for the loop region (residues 43–46) from each monomer were observed in $2F_o - F_c$ and $F_o - F_c$ electron-density maps. In addition, following refinement clear negative density surrounding Arg35 from the model confirmed the presence of the R35A mutation. Although these data were not sufficient to uniquely define the side-chain conformations, they confirmed that the structure of the R35A mutant was essentially the same as that of the wild-type protein. We conclude from these data that the significantly reduced binding affinity of the R35A mutant is a direct consequence of the absence of the Arg35 side chain and does not arise from a conformational change in the protein. This validates our previous hypothesis that the side chain of Arg35, which contacts a conserved TG dinucleotide lying outside the classic inverted-repeat recognition sequence, is central to the overall binding affinity (McGeehan *et al.*, 2008).

Interestingly, the flexible loop region identified in the native structure is also apparent in the mutant structure. Despite only possessing one dimer in the asymmetric unit, each monomer presents an alternative conformation, each corresponding to one of the two states seen in the native protein. Fig. 6 shows the refined structures of the mutant and native proteins superimposed and highlights the position of the Arg35 resi-

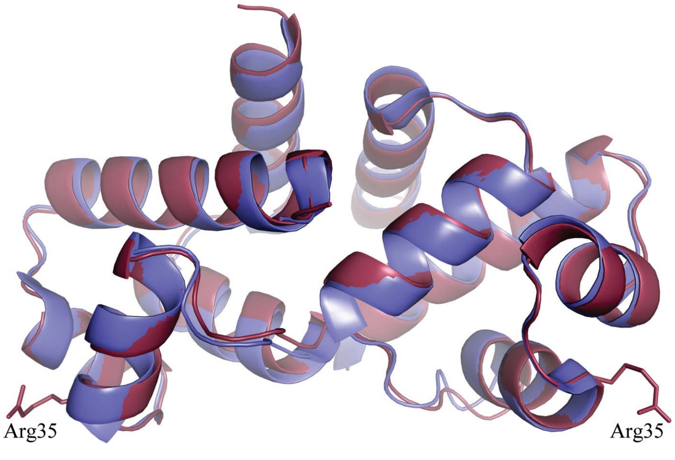

Figure 6
C.*Esp*1396I R35A mutant overlaid with chains *M* and *N* of the native protein. The main-chain atoms of the R35A protein (blue) were overlaid with the native protein (red), giving an r.m.s.d. of 0.48 Å for 408 atoms. The position of the Arg35 side chain in the native protein is shown.

dues that flank the DNA-binding interface. As expected, the overall structure of the two proteins is very similar, with the greatest degree of variation again arising from the flexible loop region (residues 43–46) between helix 3 and helix 4. Clear electron density for each of the conformational states gives further weight to the interpretation that two local energy minima exist in the free protein rather than a highly mobile loop.

High-resolution structures would assist greatly in determining the role of the conformational changes exhibited by C.*Esp*1396I upon DNA binding. To this end, further X-ray crystallography data sets of bound protein have been obtained with various lengths and sequences of DNA and structural analysis is in progress. Understanding the mechanisms that underlie the ability of C.*Esp*1396I to bind multiple DNA-recognition sequences and thus finely tune the expression of the restriction–modification system are fundamental to a complete understanding of this intricate control system.

This work was supported by a BBSRC grant (BB/E000878/1) awarded to GGK. JEM was funded by the RCUK and NB by an IBBS studentship from the University of Portsmouth, England. The authors would like to acknowledge the helpful support of the beamline scientists of the ESRF.

References

Bogdanova, E., Djordjevic, M., Papapanagiotou, I., Heyduk, T., Kneale, G. & Severinov, K. (2008). *Nucleic Acids Res.* **36**, 1429–1442.

Bogdanova, E., Zakharova, M., Streeter, S., Taylor, J., Heyduk, T., Kneale, G. & Severinov, K. (2009). *Nucleic Acids Res.* **37**, 3354–3366.

Cesnaviciene, E., Mitkaite, G., Stankevicius, K., Janulaitis, A. & Lubys, A. (2003). *Nucleic Acids Res.* **31**, 743–749.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

DeLano, W. L. (2002). *The PyMOL Molecular Graphics System.* DeLano Scientific LLC, San Carlos, California, USA.

Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.

Ives, C. L., Nathan, P. D. & Brooks, J. E. (1992). *J. Bacteriol.* **174**, 7194–7201.

Kabsch, W. (1993). *J. Appl. Cryst.* **26**, 795–800.

Knowle, D., Lintner, R. E., Touma, Y. M. & Blumenthal, R. M. (2005). *J. Bacteriol.* **187**, 488–497.

Leslie, A. G. W. (1992). *Jnt CCP4/ESF–EAMCB Newsl. Protein Crystallogr.* **26**.

McCoy, A. J., Grosse-Kunstleve, R. W., Storoni, L. C. & Read, R. J. (2005). *Acta Cryst.* D**61**, 458–464.

McGeehan, J. E., Papapanagiotou, I., Streeter, S. D. & Kneale, G. G. (2006). *J. Mol. Biol.* **358**, 523–531.

McGeehan, J. E., Streeter, S., Cooper, J. B., Mohammed, F., Fox, G. C. & Kneale, G. G. (2004). *Acta Cryst.* D**60**, 323–325.

McGeehan, J. E., Streeter, S., Papapanagiotou, I., Fox, G. C. & Kneale, G. G. (2005). *J. Mol. Biol.* **346**, 689–701.

McGeehan, J. E., Streeter, S. D., Thresh, S.-J., Ball, N., Ravelli, R. B. & Kneale, G. G. (2008). *Nucleic Acids Res.* **36**, 4778–4787.

Mruk, I. & Blumenthal, R. M. (2008). *Nucleic Acids Res.* **36**, 2581–2593.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.

Papapanagiotou, I., Streeter, S. D., Cary, P. D. & Kneale, G. G. (2007). *Nucleic Acids Res.* **35**, 2643–2650.

Rimseliene, R., Vaisvila, R. & Janulaitis, A. (1995). *Gene*, **157**, 217–219.

Sawaya, M. R., Zhu, Z., Mersha, F., Chan, S.-H., Dabur, R., Xu, S.-Y. & Balendiran, G. K. (2005). *Structure*, **13**, 1837–1847.

Streeter, S. D., Papapanagiotou, I., McGeehan, J. E. & Kneale, G. G. (2004). *Nucleic Acids Res.* **32**, 6445–6453.

Tao, T., Bourne, J. C. & Blumenthal, R. M. (1991). *J. Bacteriol.* **173**, 1367–1375.

Vijesurier, R. M., Carlock, L., Blumenthal, R. M. & Dunbar, J. C. (2000). *J. Bacteriol.* **182**, 477–487.

Wilson, G. G. & Murray, N. E. (1991). *Annu. Rev. Genet.* **25**, 585–627.