The EMBO Journal vol.13 no.17 pp.3927–3935, 1994

Structure of *Pvu*II endonuclease with cognate DNA

Xiaodong Cheng[1], K.Balendiran[2], Ira Schildkraut[3] and John E.Anderson

W.M.Keck Structural Biology Laboratory, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724 and [3]New England Biolabs, 32 Tozer Road, Beverly, MA 01915, USA

[2]Present address: Temple University School of Medicine, 3307 North Broad Street, Philadelphia, PA 19140, USA

[1]Corresponding author

Communicated by J.D.Watson

We have determined the structure of *Pvu*II endonuclease complexed with cognate DNA by X-ray crystallography. The DNA substrate is bound with a single homodimeric protein, each subunit of which reveals three structural regions. The catalytic region strongly resembles structures of other restriction endonucleases, even though these regions have dissimilar primary sequences. Comparison of the active site with those of *Eco*RV and *Eco*RI endonucleases reveals a conserved triplet sequence close to the reactive phosphodiester group and a conserved acidic pair that may represent the ligands for the catalytic cofactor Mg^{2+}. The DNA duplex is not significantly bent and maintains a B-DNA-like conformation. The subunit interface region of the homodimeric protein consists of a pseudo-three-helix bundle. Direct contacts between the protein and the base pairs of the *Pvu*II recognition site occur exclusively in the major groove through two antiparallel β strands from the sequence recognition region of the protein. Water-mediated contacts are made in the minor grooves to central bases of the site. If restriction enzymes do share a common ancestor, as has been proposed, their catalytic regions have been very strongly conserved, while their subunit interfaces and DNA sequence recognition regions have undergone remarkable structural variation.

Key words: common evolutionary origin/DNA recognition/restriction endonuclease/structural similarity/X-ray crystallography

Introduction

*Pvu*II endonuclease (R.*Pvu*II), part of the restriction–modification system from the bacterium *Proteus vulgaris*, cleaves duplex DNA with the sequence 5′-CAGCTG-3′ between the central guanine and cytosine bases in both strands to generate blunt-ended products (Gingeras *et al.*, 1981). *Pvu*II methyltransferase (M.*Pvu*II) modifies the central cytosine (Blumenthal *et al.*, 1985) to generate *N*4-methyl-cytosine (Butkus *et al.*, 1987). The *Pvu*II restriction–modification system has been cloned and sequenced (Blumenthal *et al.*, 1985; Tao *et al.*, 1989; Athanasiadis *et al.*, 1990; Tao and Blumenthal, 1992). The R.*Pvu*II coding sequence predicts a polypeptide of 157 amino acids with a molecular weight of 18 345 Da, making R.*Pvu*II the smallest restriction endonuclease found so far (Wilson and Murray, 1991).

Crystal structures are known for three other endonucleases, including R.*Eco*RI in complex with cognate DNA (Kim *et al.*, 1990), R.*Eco*RV in the absence and presence of both cognate and non-cognate DNA (Winkler *et al.*, 1993) and the apo-form of R.*Bam*HI (Newman *et al.*, 1994a,b). R.*Pvu*II has been overproduced and crystallized in apo-form (I.Schildkraut and J.Bonventre, unpublished observation; Athanasiadis and Kokkinidis, 1991) and as a protein–DNA complex (Balendiran *et al.*, 1994). We report here the crystal structure of the complex.

Results

The protein

Two views of the structure and a diagram of the locations of structural elements in the amino acid sequence are shown in Figure 1. The cleft bisecting the U-shaped R.*Pvu*II homodimer provides the binding site for DNA. Each subunit contains five α helices and eight β strands, designated A–E for helices and 1–8 for strands from the amino to the carboxyl terminus (Figure 1D). The loops between helices αA and αB, and between strands β1 and β2 are designated L-AB and L-12, respectively. The subunit is folded into three discernible and functionally distinct regions, which we term the subunit interface region, the catalytic region and the DNA recognition region. At the level of primary sequence, the interface region is contiguous, while the other two regions are mutually interspersed.

Subunit interface region. The first 46 N-terminal amino acids form the subunit interface through two helices connected by a loop (αA, loop L-AB and αB). The 21 residue helix αA is bent due to proline P14. The two αA helices of the dimer cross over one another and, as a result, the N-terminus of one subunit is close to the C-terminus of the other subunit (Figure 1B). Two loops (L-AB) of the dimer have their main chain atoms hydrogen-bonded to each other like two β strands (Figure 1C). The 10 residue αB helices are aligned head-to-head in the dimer, with the two loops L-AB in between, as if there were a single helix with a kink in the middle, like αA. Thus, the two helix–loop–helix structures are arranged as a pseudo-three-helix bundle, with the hydrophobic side chains from all four helices packing into a hydrophobic core in the center of the bundle. The only other contact of the two subunits is a hydrogen bond between the side chains of H85 (from loop L-4C), which both lie in the major groove of the DNA (Figure 1A and C). We think

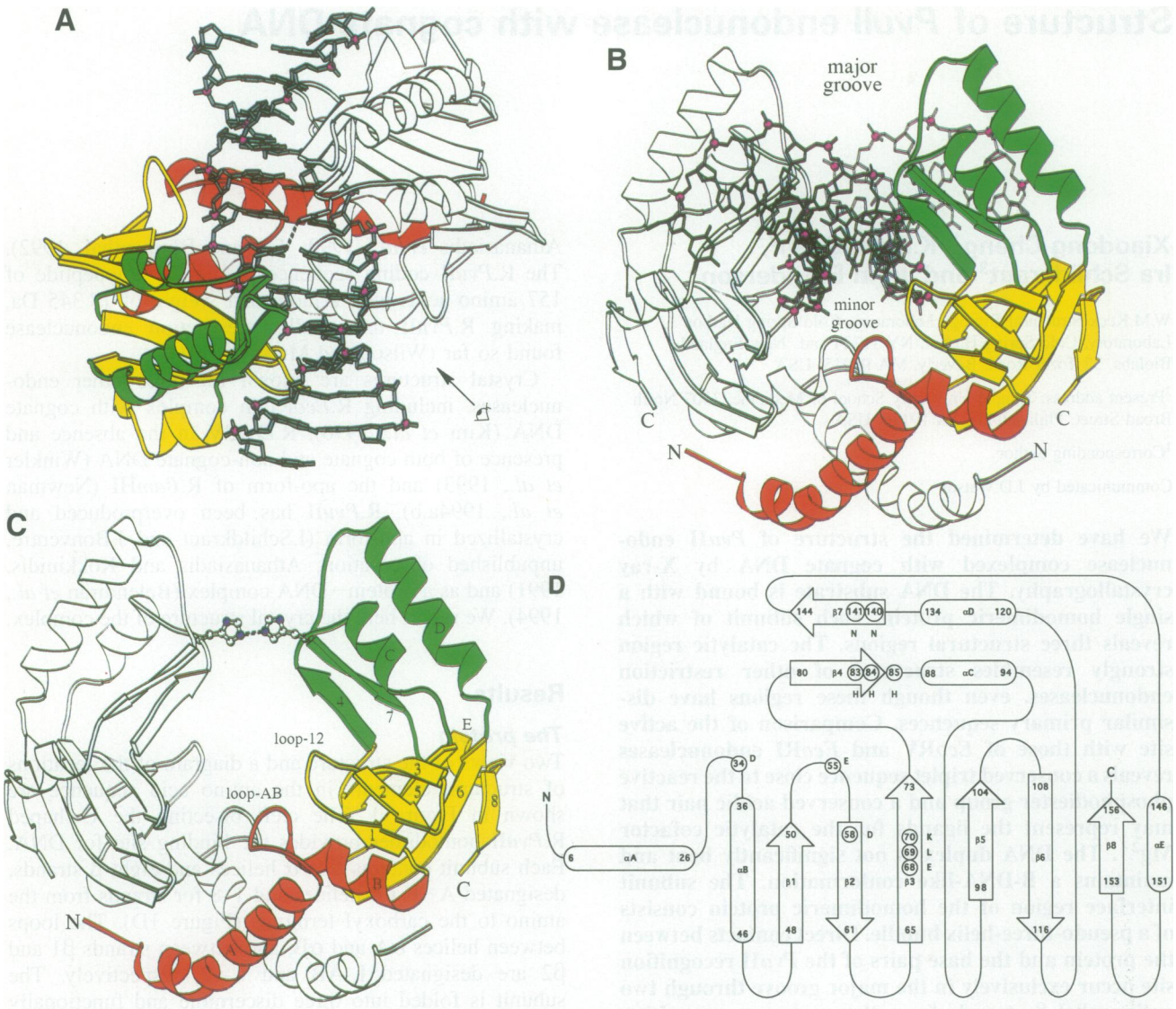

Fig. 1. Structure of R.*Pvu*II. Two subunits are shown in gray or in color, respectively, with a ball and stick model of the bound cognate DNA segment. Three regions are colored in red, green and blue for dimer interaction, catalysis and DNA recognition, respectively. (**A**) Front view of the protein–DNA complex. (**B**) Side view of the protein–DNA complex from an angle as indicated in (A). (**C**) Same view as in (B) of dimer structure without DNA. The interaction between two H85 side chains closes off the DNA binding cleft. (**D**) Locations of secondary structural elements in the amino acid sequence.

that this contact may only occur when DNA is bound (see below).

Catalytic region. The catalytic region consists of a mixed antiparallel/parallel sheet formed by three short strands (β1, β2 and β8) and three long strands (β3, β5 and β6) with one helix (αB or αE) on either side of the sheet. The sheet is twisted around strand β3, whose N-terminal half forms hydrogen bonds with strand β2 and whose C-terminal half pairs with strand β5. As a result, the hairpin of β5 and β6 is almost perpendicular to the hairpin of β1 and β2. Potential catalytic residues lie in loop L-12, in the N-terminal end of strand β2 and in the middle of the twisted strand β3, all in the vicinity of the reactive phosphodiester group (see below).

DNA recognition region. The recognition region comprises two substructures from different locations in the primary sequence, each containing one strand and one helix. Strand β4 and helix αC (residues 80–94) run antiparallel with helix αD and strand β7 (residues 120–144) (Figure 1D). The resulting two antiparallel β strands belonging to each monomer make specific interactions with DNA bases in the major groove. A similar mode of base specific recognition was seen in the structures of *Escherichia coli* MetJ repressor and bacteriophage P22 Arc repressor, in which a β strand of one monomer and its two-fold symmetry-related counterpart of the other monomer form an antiparallel β ribbon contacting bases in the major groove (Somers and Phillips, 1992; Raumann *et al.*, 1994). The β ribbon as a DNA recognition motif may be as 'popular' as the helix–turn–helix and zinc finger motifs (Kim, 1992).

The recognition region also contains a histidine triplet (H83–H84–H85) and an asparagine doublet (N140–N141). The side chain of H83 contacts the phosphate of one central guanine (−1), while the side chain atom $N_{\delta 1}$ of H84 forms a hydrogen bond to the O6 atom of the other central guanine (+1) (Figure 2B). The closest protein atom to the N4 atom of Cyt (+1) is the C_β of H84, which

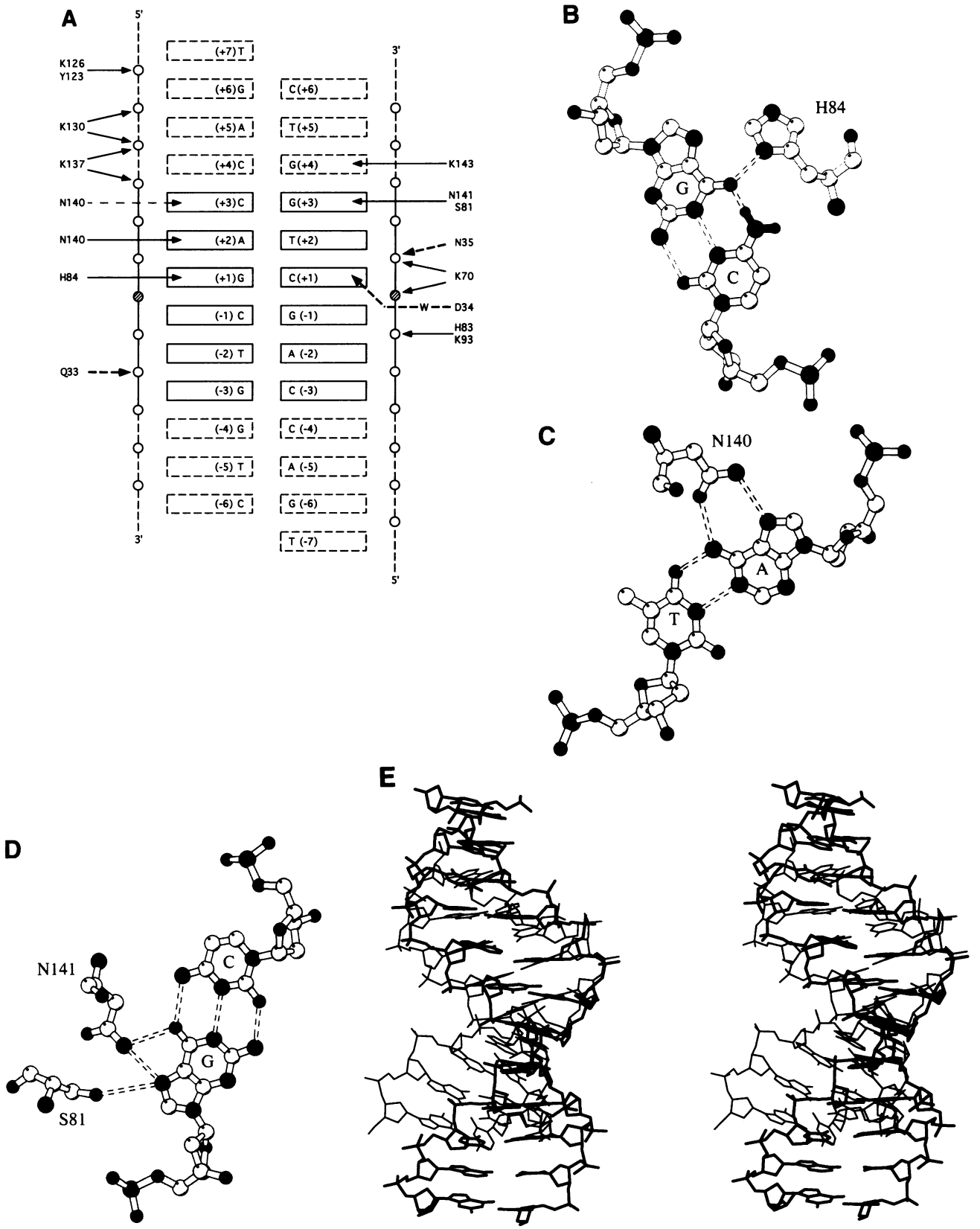

Fig. 2. Protein–DNA interactions. (**A**) Schematic representation showing the base and phosphate contacts between DNA and one subunit of R.*Pvu*II. The recognized nucleotides are in solid lines and the target phosphate is shaded. Contacts to the minor groove are shown with thick dashed lines and contact by the main chain atom with a thin dashed line. The symbol W indicates a water-mediated contact. (**B**) The O6 atom of Gua (+1) forms a hydrogen bond with $N_{\delta 1}$ of H84. Oxygen and nitrogen atoms and phosphate molecules are shaded, carbon is in white and two N4 protons of Cyt (+1) are also shaded. The effect of *N*4-methylation of Cyt (+1), which prevents cleavage by R.*Pvu*II, may be due to a collision between the methyl group and H84. (**C**) The N6 and N7 atoms of Ade (+2) form two hydrogen bonds with $O_{\delta 1}$ and $N_{\delta 2}$ of N140, respectively. (**D**) The O6 and N7 atoms of Gua (+3) form hydrogen bonds with $N_{\delta 2}$ of N141. The N7 atom also interacts with hydroxyl group of S81. (**E**) Stereo view showing the superimposition of the DNA conformations in the complexes of R.*Pvu*II–DNA (thick line) and R.*Eco*RV–DNA (thin line).

is 4.3 Å away. The replacement of the proton at N4 by a methyl group may force the side chain of H84 farther away and break the hydrogen bond between $N_{\delta 1}$ of H84 and the O6 atom of guanine (+1). This probably explains how *N*4-methylation of Cyt (+1) prevents cleavage by R.*Pvu*II. As mentioned above, the side chain atom $N_{\varepsilon 2}$ of H85 from one subunit is hydrogen-bonded to H85 of the other subunit by sharing one proton. This dimeric interaction wraps around the central G–C base pairs and closes off the DNA binding cleft and therefore must occur after the DNA is bound (Figure 1A and C). It is not yet clear whether this H85–H85 interaction forms only with a specific complex or if it forms even on binding non-specific DNA, perhaps to enhance sequence scanning by linear diffusion (Jack *et al.*, 1982; Terry *et al.*, 1985).

It is interesting that the two purines, Ade (+2) and Gua (+3), both interact with asparagine: N140 with the adenine and N141 with the guanine. Mutations at either of these two positions eliminate R.*Pvu*II activity (see below). The side chain oxygen atom and amide group of N140 make a hydrogen bond to the N6 and N7 atoms of adenine, respectively, in which N6 is the hydrogen donor and N7 is the hydrogen acceptor (Figure 2C). In the case of N141, the amide group makes two hydrogen bonds to the O6 and N7 atoms of guanine, in which both atoms are hydrogen acceptors (Figure 2D). Thus an asparagine can recognize either purine base by forming distinct pairs of hydrogen bonds. The side chains of Q33–D34–N35 in the loop L-AB, which forms the bottom of the DNA binding cleft and spans the minor groove (Figure 1C), make direct contacts to backbone phosphates and water-mediated contacts to the central G–C base pair (Figure 2A). One base interaction outside the recognized nucleotides is observed in the major groove between the side chain of K143 and Gua (+4). This may explain the observation that the DNA adjacent to the recognition site has an effect on the cleavage efficiency (Chen *et al.*, 1991).

A series of mutants (D34G in loop L-12, S81L and T82A in strand β4 and N140D, N141K and I144K in strand β7) have been isolated that have essentially lost R.*Pvu*II activity (P.Riggs and P.Evans, personal communication) and all map to the DNA interface. Each of the mutations in D34, N140 and N141 disrupts the interactions between the protein and DNA, as discussed above. The side chain of S81 is located near N141 and makes a hydrogen bond with the N7 atom of Gua (+3) (Figure 2D); its replacement by Leu abolishes this sequence-specific interaction. The hydroxyl group of T82 in strand β4 forms a hydrogen bond with the main chain oxygen atom of amino acid 142 in strand β7. Replacing T82 with Ala would destabilize the two antiparallel β strands (β4 and β7) of the DNA recognition region. The side chain of I144 points away from DNA and is buried in the hydrophobic interior and its substitution with Lys would affect the stability of the folded protein.

The active site and comparison with other endonucleases

The two proposed catalytic mechanisms for R.*Eco*RV and R.*Eco*RI both involve an attack on the scissile phosphate by a Mg^{2+}-activated water molecule, but they differ in whether the water is also activated by one or more acidic

Fig. 3. The catalytic domain of type II endonucleases. The residue numbers are indicated above the sequence. Possible catalytic side chains are shaded and the β strands are indicated by the numbers below the sequence, according to the designation of R.*Pvu*II.

side chains (Selent *et al.*, 1992) or by an adjacent phosphate group from the DNA substrate (Jeltsch *et al.*, 1992, 1993). A common structural motif (P73–D74 . . . D90–I91–K92 in R.*Eco*RV and P90–D91 . . . E111–A112–K113 in R.*Eco*RI) was found in the vicinity of the target scissile phosphodiester bond in both enzymes (Jeltsch *et al.*, 1992, 1993). A third acidic side chain, E45 in R.*Eco*RV and D59 in R.*Eco*RI, was also identified in the vicinity of the active site. The acidic side chains probably form a high-affinity metal ion binding site and the basic lysine and Mg^{2+} stabilize the pentavalent doubly-negative phosphate in the transition state. However, only nine out of the 36 type II endonucleases in the Swissprot protein sequence database contain this exact motif (Anderson, 1993). No exact match to this motif was identified in R.*Pvu*II.

A triplet in R.*Pvu*II, E68–L69–K70, is close to the active site, but no PD dipeptide can be found nearby. The side chain amino group of K70 bridges the phosphates of cytosine (+1) and thymine (+2). This triplet does bear a steric resemblance to E111–A112–K113 of R.*Eco*RI and D90–I91–K92 of R.*Eco*RV. All three triplets are situated in strand β3 in the active site β sheet (Figure 3). Superposition based on the alignment of the triplets reveals that D58 of R.*Pvu*II occupies a similar main chain position to that of D74 of R.*Eco*RV and D91 of R.*Eco*RI, but there is no proline immediately before D58. There is another acidic side chain in R.*Pvu*II, E55, whose C_α atom is also <7 Å from the scissile phosphate. We think that both E55 and D58 could potentially participate in the formation of the Mg^{2+} binding site or in the activation of the water molecule which attacks the scissile phosphate, analogously to the acidic pairs E45 and D74 of R.*Eco*RV and D59 and D91 of R.*Eco*RI. The following observations argue that E55 is more likely to be the counterpart for the aspartate in the PD dipeptide of R.*Eco*RV and R.*Eco*RI. We note that E55 of R.*Pvu*II is located in a loop region, L-12, and so are D74 of R.*Eco*RV and D91 of R.*Eco*RI (Figure 3), while D58 is part of a β strand. Unlike the acidic residues in R.*Eco*RV and R.*Eco*RI, E55 of R.*Pvu*II is not preceded by a proline. However, E55 is followed by a glycine and so is D91 of R.*Eco*RI. These prolines and glycines may be required to properly position the Mg^{2+} binding site. The electron density for the side chain of D58 is well localized, while the side chain of E55 is more flexible, as indicated by a relatively higher thermal factor. If the side chain of E55 normally interacts with

the cofactor, then this flexibility may be due to the absence of Mg^{2+} in the crystal.

Spatial alignment of the triplets reveals that the β sheets in the catalytic region of all four characterized enzymes superimpose quite well, except for the difference in size of the strands (Figure 4). However, the strand β6 in R.*Pvu*II and R.*Eco*RV is oppositely oriented to the strand β6 in R.*Eco*RI and R.*Bam*HI (according to the designation of R.*Pvu*II in Figure 3), despite almost perfect spatial overlap. Topologically, the antiparallel strands β5 and β6 in R.*Pvu*II and R.*Eco*RV are formed by a hairpin (Figure 1D), while the parallel strands β5 and β6 in R.*Eco*RI and R.*Bam*HI are formed by a β−α−β structure (Newman *et al.*, 1994a,b). No obvious similarities were found for the subunit interface region, and the mode of DNA recognition is also different. When the apo-structure of R.*Bam*HI is superimposed onto the R.*Eco*RI structure, many places, including all of the active site, DNA binding cleft and dimer interface, fit with an overall r.m.s. error of < 2 Å, despite the lack of detectable sequence similarity (Newman *et al.*, 1994a,b).

The DNA

The DNA in our structure is a 12 base duplex containing the recognition sequence (CAGCTG) and single base (T) 5′-overhangs (Figure 2A). The length of the duplex is the same as the crystallographic *c*-axis (48.5 Å). The DNA duplexes are stacked end to end, with the helical axis parallel to the *c*-axis of the crystal. The only other crystallographic contact occurs between side chains from the C-terminal strand (β8) and strands β1 and β2. The DNA is located in the cleft between the two subunits. Contacts between the protein and backbone phosphates span all 12 base pairs of the duplex. The minor groove of the DNA faces the bottom of the cleft and the major groove faces the open end (Figure 1B).

The bound duplex has the characteristic shape of B-form DNA. Average helical and conformational values (± SD) are: twist, 29 ± 6°; rise, 3.7 ± 0.3 Å; roll, 0 ± 8°; tilt, 0 ± 4°; slide, −2.7 ± 0.8 Å. Calculations were done with the program NEWHEL93 (R.E.Dickerson) obtained through the Brookhaven Protein Data Bank. This conformation is quite different from that of the decanucleotide bound to R.*Eco*RV, in which a kink was observed at the central base pair resulting in a compression of the major groove (Figure 2E). It was suggested that the kinked DNA conformation characterizes the specific DNA binding for R.*Eco*RV, since a complex containing non-specific DNA does not have a kink (Winkler *et al.*, 1993). The structure of the DNA in the R.*Eco*RI−DNA complex is also deformed from its normal B-DNA conformation, so as to suggest that this may be necessary for specific complex formation (Kim *et al.*, 1990). However, our structure for R.*Pvu*II with cognate DNA indicates that DNA deformation is not necessary for specific binding by R.*Pvu*II. We think that it is probably the orientation of and distance between the catalytic side chain(s) and the target bond that are critical for a specific interaction that leads to cleavage, rather than a distortion in DNA conformation, since the distance between the negatively charged side chain(s) in the loop region and the scissile phosphate remains the same for both R.*Pvu*II and R.*Eco*RV.

Discussion

The loop L-12 in R.*Pvu*II is the shortest one among the four structurally characterized endonucleases (Figure 3) and the small size of this loop may allow non-distorting interaction with B-form DNA. The longer and bulkier equivalent loop in R.*Eco*RV, however, may require a kink in the DNA to ensure similar distances between the catalytic residue(s) and the target, while at the same time avoiding steric collision between the protein and DNA. DNA bending is associated with a transition enthalpy of 1.6 kcal/(mol base pair) (Park and Breslauer, 1991) and an endonuclease paying this penalty would presumably have to compensate with additional specific interactions with the DNA. A loop swap experiment between the two enzymes will be helpful in addressing these questions.

Comparison of the R.*Bam*HI structure with that of R.*Eco*RI shows that many regions of the protein can be superimposed and that the similarity is strong enough to propose a common evolutionary origin (Newman *et al.*, 1994a,b). A common feature of R.*Bam*HI and R.*Eco*RI is that they recognize similar hexanucleotide sequences (GGATCC and GAATTC) which differ only in two positions and both cleave after the first base. R.*Pvu*II and R.*Eco*RV recognize more dissimilar hexanucleotide sequences (CAGCTG and GATATC) which differ in four positions, but again cleave their sequences at identical positions, in the middle, to leave a blunt end. The structure of R.*Pvu*II reported here partially resembles that of R.*Eco*RV. For this pair of enzymes the structural similarity is less profound (perhaps partly because of the greater divergence of their recognition sequences), but might still reflect the trace of a common origin.

There is little detectable sequence similarity among the more than 60 Type II restriction endonuclease genes found so far. The structural similarity in the catalytic region shared by all four proteins presumably reflects the common reaction catalyzed by type II restriction enzymes, which cleave both DNA strands to yield 5′ phosphate and 3′ hydroxy groups. This similarity may also reflect a common ancestry for all restriction enzymes.

In the structure of R.*Eco*RI−DNA, the DNA binding cleft of the dimeric R.*Eco*RI molecule faces the major groove of the DNA (Kim *et al.*, 1990), while in the R.*Eco*RV−DNA complex, the DNA binding cleft of R.*Eco*RV faces the minor groove side (Winkler *et al.*, 1993). The bonds cleaved by R.*Eco*RI to leave four-base, single-stranded 5′ extensions are on the major groove side of the DNA, while those cleaved by R.*Eco*RV to leave blunt ends are on the minor groove side. Thus the difference in the directions in which R.*Eco*RI and R.*Eco*RV approach the DNA may be related to the positions of their scissile bonds (Anderson, 1993). The orientation of the catalytic region may need to be stabilized by the dimerization region for efficient cleavage, requiring that the DNA binding cleft face the side of the DNA the scissile bonds are on. This hypothesis is supported by the structure of R.*Pvu*II−DNA, in which the enzyme binds to DNA from the minor groove side. We predict that enzymes that leave single-stranded 3′ extensions approach DNA from the minor groove side as well. To date, there are no structures available for enzymes that leave 3′ overhangs to test this prediction.

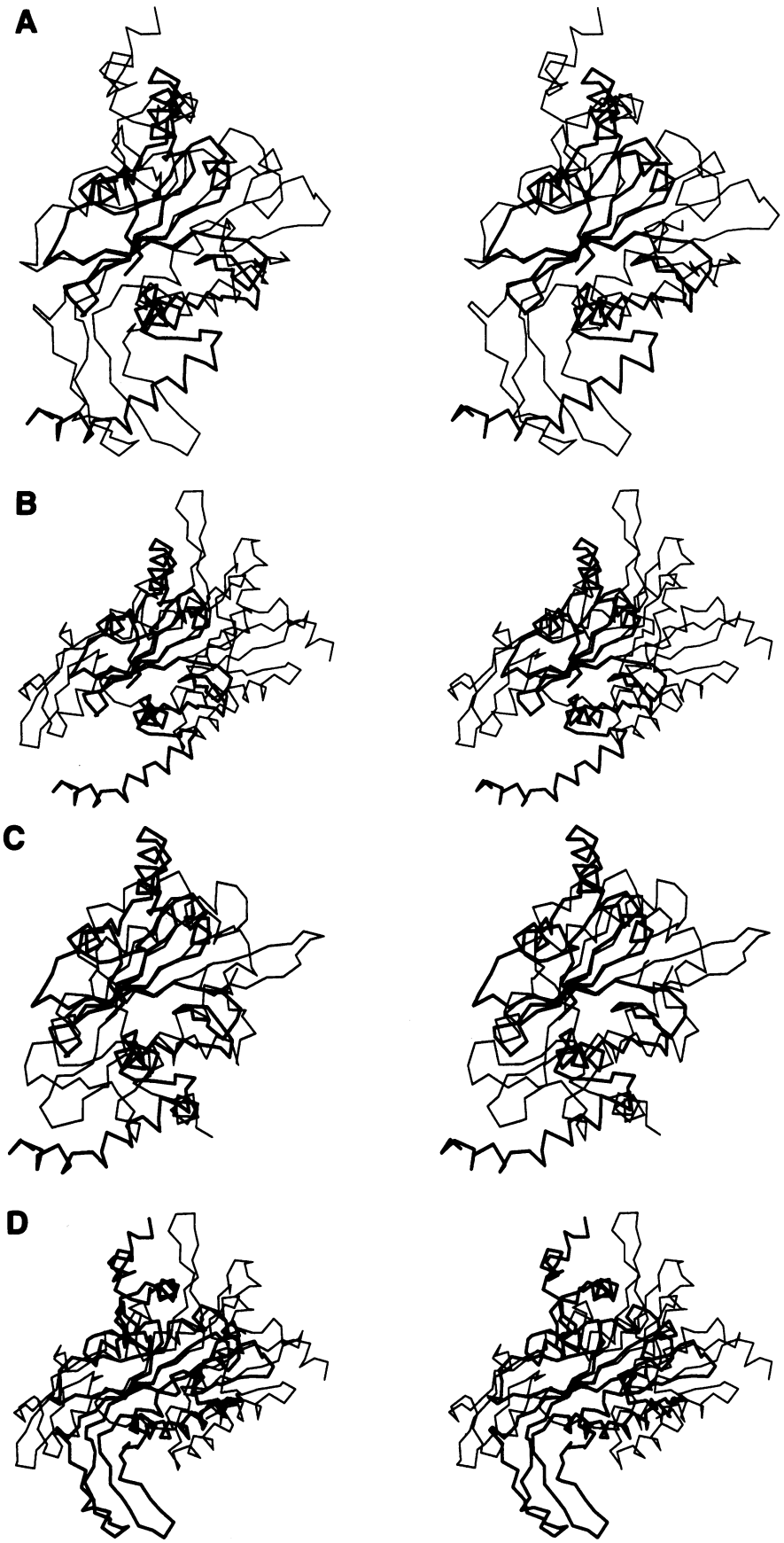

Fig. 4. Stereo views showing the superimposition of the protein C_α chains of (**A**) R.*Pvu*II–DNA (thick line) and R.*Eco*RV–DNA, (**B**) R.*Pvu*II–DNA and R.*Eco*RI–DNA, (**C**) R.*Pvu*II–DNA and the apo-structure of R.*Bam*HI and (**D**) R.*Eco*RV–DNA (thick line) and R.*Eco*RI–DNA. The twisted β-sheet catalytic domain fits well among all four structures.

Table I. Statistics of crystallographic data collection

(A) Derivatives of iodinated oligonucleotides

Native	ITT1	ITT8	ITT9	ITT5
TGACCAGCTGGTC CTGGTCGACCAGT	TGACCAGCTGGUC CUGGTCGACCAGT	TGACCAGCTGGTC CTGGTCGACCAGT	TGACCAGCUGGTC CTGGUCGACCAGT	UGACCAGCTGGTC CTGGTCGACCAGU

(B) Data collection

Space group	$P2_12_12_1$			
Cell dimensions (Å)	a = 95.8	b = 86.3	c = 48.5	
Sample	Native	ITT1	ITT8	ITT9	ITT5
Observation (number)	45607	28737	40227	40137	30823
Unique (number)	10282	5960	8270	8919	7053
R merge (%)[a]	7.00	6.33	10.68	10.11	10.44
Redundancy	4.4	4.8	4.9	4.5	4.4
Crystals (number)	3	2	1[b]	2	3

(C) Phasing

Resolution range (Å)	Completeness (%)				
∞ − 3.00	92.7	70.7	91.7	92.5	83.4
Iodine sites (number)		2	2	2	2
Phasing power[c] at 3 Å					
Isomorphous		0.72	1.32	1.76	
Anomalous		1.38	1.13	2.55	
Figure of merit at 3 Å	0.729				

[a]R merge = $\Sigma|I - \langle I\rangle|/\Sigma\langle I\rangle$, where I = observed intensity and $\langle I\rangle$ = averaged intensity obtained from multiple observations of symmetry related reflections.
[b]One crystal was used with many translations for derivative ITT8.
[c]Phasing power = r.m.s. ($\langle F_H\rangle/E$), where F_H = heavy atom structure factor amplitude and E = residual lack of closure.

Table II. Statistics of refinement

Resolution range (Å)	Reflections (number)		*R* value (%)	
10.0–2.60	9799		18.9	
	Completeness (%)			
10.0–4.35	97.2			
4.35–3.50	98.5			
3.50–3.07	91.6			
3.07–2.80	69.3			
2.80–2.60	34.8			
Ramachandran plot quality assessment (Laskowski, 1993)				(%)
Residue in most favoured regions				82
Residue in additional favoured regions				18
Residue in generously allowed regions				0
Residue in disallowed regions				0
	Protein	DNA	Water	
Atoms	2584	492	36	
r.m.s. deviation from idealized covalent geometry				
bond length (Å)	0.012	0.031		
bond angle (°)	1.7	4.0		

What is particularly striking about the four structures we have to date, however, is the diversity they encompass despite the similarity of their enzymatic roles. Their subunit interfaces differ fundamentally. Their different DNA sequence recognition mechanisms are further reflected in the profound differences they exhibit in the structure of bound substrate DNA. Only the catalytic regions show obvious conservation and this only at the structural level. Further, the catalytic and sequence recognition regions are mutually interspersed at the primary sequence level, so a model of mosaic evolution would be difficult to accept. While the question of ancestry remains open, it is clear that these enzymes represent a family with shared structural elements, despite the lack of detectable sequence identity. As such, this unique group of enzymes may offer an important resource for helping to unravel the rules governing protein folding.

Materials and methods

Crystallization

R.*Pvu*II was expressed, purified and crystallized with DNA as described by Balendiran *et al.* (1994). The crystals were grown by mixing the complex solution (molar ratio of protein:DNA of 2.7:1) with mother liquor (8% PEG-4000, 0.3 mM EDTA, 50 mM sodium acetate, pH 4.5) and equilibrating the mixture against 1 ml of the latter solution at 16°C. The crystals formed in space group $P2_12_12_1$ with unit cell dimensions

a = 95.8 Å, b = 86.3 Å and c = 48.5 Å. There is one complex (two molecules of R.*Pvu*II and one duplex of the DNA) in each of the four asymmetric units and about 52% of solvent content.

X-ray data collection

X-ray diffraction data (Table I) were collected at 16°C or room temperature on a FAST area detector using the program MADNES (Messerschmidt and Pflugrath, 1987). An Elliott GX-21 rotating anode X-ray generator (Enraf-Nonius) produced monochromated CuK_α radiation (1.54 Å, operating at 45 KV, 95 mA). Crystals were aligned with their a-axis along the ω-axis, so that the isomorphous and anomalous data were measured simultaneously. The unit cell parameters were refined during data reduction. The data were evaluated by profile fitting (Kabsch, 1988). The data sets were divided into sectors for scaling and merging using the programs FS and PROTEIN (Steigemann, 1974; Weissman, 1982). The isomorphous and anomalous data were kept separate in the phase calculations.

Phasing

The heavy atom was introduced by covalent modification of the oligonucleotide, substituting an iodine atom for the methyl group of thymidine or cytosine. Iodinated oligonucleotides (i.e. T→U^I = 5-iododeoxyuridine, C→C^I = 5-iodocytosine) at two symmetrically related positions in the sequence were synthesized at New England Biolabs, Inc. Data for iodinated co-crystals were collected in the dark. The program package PHASES was used for refining heavy atom parameters, solvent leveling, two-fold density averaging, phase and map calculations and skeleton generation (Furey and Swaminathan, 1990). The iodine positions were first determined from difference Patterson syntheses and confirmed by the difference Fourier method. The heavy atom parameters were initially refined against centric data to 3 Å resolution. Though ITT5 positions do not refine along with the others, they are consistent with the fact that the iodine atoms are located in the unpaired 5′ overhangs and are apparently not well localized. Accordingly, ITT5 data were excluded from the phase calculations. The phases were then combined to give multiple isomorphous replacement with anomalous scattering (MIRAS) phases. Even though the data for derivatives extended beyond 3 Å resolution, the phases were only calculated to 3 Å because of the incompleteness of data and the drop in phasing power at higher resolution. The initial MIRAS phases were improved by four, four and eight cycles of solvent leveling (Wang, 1985), 40% solvent content, following each of three envelope determinations. The solvent-leveled map was used to refine the two-fold symmetry operator (ϕ = 2.9°, ψ = 132.9°, X_{ori} = 20.3 Å, Y_{ori} = 18.6 Å, Z_{ori} = 4.8 Å) and to construct the averaging mask using the MAPVIEW program of the PHASES package. The phases were further improved using 16 rounds of the averaging protocols suggested in the PHASES documentation. The electron density was averaged within the mask and the averaged density map was inverted to obtain new phases. In general, the secondary structural elements were obvious at this stage. The heavy atom parameters were further refined against the solvent-leveled, averaged protein phases to obtain better phases. The result clearly indicated that the isomorphous data from ITT1 at a resolution of 3 Å and the anomalous data from ITT8 at resolutions >3.9 Å were not consistent with the phases from the other data sets, because their phasing power dropped to a value <1.0; accordingly they were excluded from the phase calculations. A new MIRAS map was then computed, solvent-flattened and averaged. The final map, whose initial phase was calculated from only two derivatives, ITT9 (both isomorphous and anomalous data to 3 Å) and ITT8 (isomorphous data to 3 Å and anomalous data to 3.9 Å), was of higher quality than the first. The resulting map at 3 Å was used for automatic skeletonization.

Model building

The atomic model was built on an Elan-4000 IRIS Indigo (Silicon Graphics), using version 5.9.1 of the molecular modeling program 'O' (Jones *et al.*, 1991). The main chain skeletons account for most of the secondary structural elements, which confirmed that there are two very similar modules of protein and one duplex of oligonucleotide per asymmetric unit. The side chains were fitted to recognizable densities. In the final model the density for the side chain of Y94 is unclear and it was modeled as an Ala. The atomic model of both subunits was built and refined independently. The protein starts at amino acid 2 because the N-terminal methionine of the coding sequence is not visible in the density maps and has been shown to be removed *in vivo* (Tao and Blumenthal, 1992). The r.m.s. deviation between 156 C_α atoms of the final refined two subunits is 0.55 Å.

Refinement

The resulting model was refined using simulated annealing, positional refinement and restrained B-factor refinement with the program X-PLOR (Brunger, 1992). The X-ray data for refinement included *all* collected low resolution and high resolution reflections (Table II). This was coupled with a bulk solvent correction using X-PLOR. Two parameters for the solvent region, the solvent scattering density and an isotropic temperature factor, were determined by an iterative search for the value of one parameter that minimizes the R factor in the lowest resolution shell containing about 1000 reflections while keeping the protein parameters fixed. Thirteen rounds of solvent correction, least-square refinement using X-PLOR and model building using O allowed a better evaluation of the surface structure of the protein and placement of well-ordered water molecules by examination of difference Fourier maps, ($2F_o-F_c$, α_c) and (F_o-F_c, α_c). Coordinates have been deposited in the Brookhaven Protein Data Bank and are also available by e-mail (cheng@cshl.org).

Acknowledgements

We thank J.W.Pflugrath for advice on data collection, W.Furey for explaining his phase refinement and non-crystallographic symmetry averaging programs, R.Knott for making oligonucleotides, A.Aggarwal and M.Newman for the coordinates of R.*Bam*HI prior to publication, T.Malone for making some of the figures, R.M.Blumenthal for critical discussion and R.M.Blumenthal, A.S. Bhagwat and X.Zhang for critically reading the manuscript. Work at Cold Spring Harbor Laboratory was supported by NSF grant DMB9005579 to J.E.A. and a grant from the W.M.Keck Foundation and Cold Spring Harbor Laboratory Association to X.C.

References

Anderson,J.E. (1993) *Curr. Opin. Struct. Biol.*, **3**, 24–30.
Athanasiadis,A. and Kokkinidis,M. (1991) *J. Mol. Biol.*, **222**, 451–453.
Athanasiadis,A., Gregoriu,M., Thanos,D., Kokkinidis,M. and Papamatheakis,J. (1990) *Nucleic Acids Res.*, **18**, 6434–6435.
Balendiran,K., Bonventre,J., Knott,R., Jack,W., Benner,J., Schildkraut,I. and Anderson,J.E. (1994) *Proteins: Structure, Function Genet.*, **19**, 77–79.
Blumenthal,R.M., Gregory,S.A. and Cooperider,J.S. (1985) *J. Bacteriol.*, **164**, 501–509.
Brunger,A.T. (1992) *X-PLOR, Version 3.1.* Yale University Press, New Haven, CT.
Butkus,V., Klimasauskas,S., Petrauskiene,L., Maneliene,Z., Lebionka,A. and Janulaitis,A.A. (1987) *Biochim. Biophys. Acta*, **909**, 201–207.
Chen,D., Liu,Q., Chen,X., Zhao,X. and Chen,Y. (1991) *Nucleic Acids Res.*, **19**, 5703–5705.
Furey,W. and Swaminathan,S. (1990) *Am. Crystallogr. Assoc. Meeting Abstr., Ser. 2*, **18**, 73.
Gingeras,T.R., Greenough,L., Schildkraut,I. and Roberts,R.J. (1981) *Nucleic Acids Res.*, **9**, 4525–4536.
Jack,W.E., Terry,B.J., Modrich,P. (1982) *Proc. Natl Acad. Sci. USA*, **79**, 4010–4014.
Jeltsch,A., Alves,J., Maass,G. and Pingoud,A. (1992) *FEBS Lett.*, **304**, 4–8.
Jeltsch,A., Alves,J., Wolfes,H., Maass,G. and Pingoud,A. (1993) *Proc. Natl Acad. Sci. USA*, **90**, 8499–8503.
Jones,T.A., Zou,J.Y., Cowan,S.W. and Kjeldgard,M. (1991) *Acta Crystallogr.*, **A47**, 110–119.
Kabsch,W. (1988) *J. Appl. Crystallogr.*, **21**, 916–924.
Kim,S.-H. (1992) *Science*, **255**, 1217–1218.
Kim,Y., Grable,J.C., Love,R., Greene,P.J. and Rosenberg,J.M. (1990) *Science*, **249**, 1307–1309.
Laskowski,R.A. (1993) *J. Appl. Crystallogr.*, **26**, 283–291.
Messerschmidt,A. and Pflugrath,J.W. (1987) *J. Appl. Crystallogr.*, **20**, 306–315.
Newman,M., Strzelecka,T., Dorner,L.F., Schildkraut,I. and Aggarwal, A.K. (1994a) *Nature*, **368**, 660–664.
Newman,M., Strzelecka,T., Dorner,L.F., Schildkraut,I. and Aggarwal, A.K. (1994b) *Structure*, **2**, 439–452.
Park,Y. and Breslauer,K. (1991) *Proc. Natl Acad. Sci. USA*, **88**, 1551–1555.
Raumann,B.E., Rould,M.A., Pabo,C.O. and Sauer,R.T. (1994) *Nature*, **367**, 754–757.

Selent,U. *et al.*, (1992) *Biochemistry*, **31**, 4808–4815.
Somers,W.S. and Phillips,S.E.V. (1992) *Nature*, **359**, 387–393.
Steigemann,W. (1974) PhD Thesis, Technische Universität, Munchen, FRG.
Tao,T. and Blumenthal,R.M. (1992) *J. Bacteriol.*, **174**, 3395–3398.
Tao,T., Walter,J., Bennab,K.J., Cotterman,M.M. and Blenthal,R.M. (1989) *Nucleic Acids Res.*, **17**, 4161–4175.
Terry,B.J., Jack,W.E. and Modrich,P. (1985) *J. Biol. Chem.*, **260**, 13130–13137.
Wang,B.C. (1985) *Methods Enzymol.*, **115**, 90–112.
Weissman,L. (1982) In Sayre,D. (ed.), *Computational Crystallography*. Oxford University Press, Oxford, pp. 56–63.
Wilson,G.G. and Murray,N.E. (1991) *Annu. Rev. Genet.*, **25**, 585–627.
Winkler,F.K. *et al.*, (1993) *EMBO J.*, **12**, 1781–1795.

Received on April 20, 1994; revised on June 16, 1994