Nucleic Acids Research, 1995, Vol. 23, No. 7 1177–1183

Tyrosine 27 of the specificity polypeptide of *Eco*KI can be UV crosslinked to a bromodeoxyuridine-substituted DNA target sequence

A. Chen[1,2], L. M. Powell[1], D. T. F. Dryden[1], N. E. Murray[1,*] and T. Brown[2]

[1]Institute of Cell and Molecular Biology and [2]Department of Chemistry, University of Edinburgh, The King's Buildings, Edinburgh EH9 3JR, UK

Received December 8, 1994; Revised and Accepted February 24, 1995

ABSTRACT

The specificity (S) subunit of the restriction enzyme *Eco*KI imparts specificity for the sequence AAC(N_6)GTGC. Substitution of thymine with bromodeoxyuridine in a 25 bp DNA duplex containing this sequence stimulated UV light-induced covalent crosslinking to the S subunit. Crosslinking occurred only at the residue complementary to the first adenine in the AAC sequence, demonstrating a close contact between the major groove at this sequence and the S subunit. Peptide sequencing of a proteolytically-digested, crosslinked complex identified tyrosine 27 in the S subunit as the site of crosslinking. This is consistent with the role of the N-terminal domain of the S subunit in recognizing the AAC sequence. Tyrosine 27 is conserved in the S subunits of the three type I enzymes that share the sequence AA in the trinucleotide component of their target sequence. This suggests that tyrosine 27 may make a similar DNA contact in these other enzymes.

INTRODUCTION

The type IA restriction and modification (R–M) enzyme of *Escherichia coli* K12 (*Eco*KI) comprises three kinds of polypeptide, R (restriction), M (modification) and S (specificity) (1; for reviews see refs. 2–4). All three kinds of polypeptide are essential for restriction while only two of them, M and S, are required for modification (5,6). In the presence of *S*-adenosyl-L-methionine (AdoMet), both endonuclease and methyltransferase (mtase) recognize DNA that contains the recognition sequence AAC(N_6)GTGC (7,8). *Eco*KI methylates the N6 position of adenine at the underlined positions in the recognition sequence (9). Hemimethylated DNA is the preferred substrate for methylation (5,10,11) while unmodified DNA is restricted (12).

Many lines of evidence indicate that for type I R–M systems the S polypeptide imparts the specificity for a specific DNA sequence (13–15). Indeed it has been shown that the S polypeptide of each type I family (IA, IB or IC) has two variable DNA target recognition domains (TRDs) of negligible similarity, each of which recognizes one half of the bipartite recognition sequence (16–19). In the case of the type IA enzymes the amino recognition domain (ARD) recognizes the trinucleotide component of the recognition site while the carboxy recognition domain (CRD) recognizes the tetranucleotide component (16,17,20). In addition to the variable regions, the type IA S subunits have two relatively short regions which are well conserved within the family (21,22). The arrangement of conserved and variable regions within the type IB and IC families is shown in Figure 1 for comparison with type IA. Repeat sequences found in the conserved regions of type IA, IB and IC S polypeptides (18,23,24) have been implicated in interactions with the M subunits (25–27).

Amino acid sequence comparisons of S polypeptide recognition domains have not identified specific amino acids responsible for DNA sequence specificity (16,17). ARDs which recognize different sequences show negligible similarity, while comparison of ARDs (within a family) that recognize the same trinucleotide, reveals 80–90% identity (16,17). Comparison of ARDs recognizing the same trinucleotide between families, while showing much lower (44%) identity (16), still does not allow the pinpointing of important amino acids.

Crystal structures have been reported for several type II endonucleases, including the apo-form of *Bam*HI (28,29) and *Pvu*II (30), and complexes of *Eco*RI, *Eco*RV and *Pvu*II with their cognate DNAs (31–33). Recently the structures of two type II mtases have also been reported (34,35).

In the absence of structural information on the amino acids involved in DNA recognition in type I systems we have used a bromouracil-mediated photocrosslinking method (36–42) to identify amino acids in close proximity to the target sequence.

MATERIALS AND METHODS

Oligonucleotide synthesis and preparation of *Eco*KI methyltransferase

Complementary oligonucleotides (25mers) containing the *Eco*KI recognition sequence (AAC(N_6)GTGC) or derivatives of it were synthesized on an Applied Biosystems 380 B DNA Synthesizer utilizing β-cyanoethyl phosphoramidite chemistry (43). 5-Bromodeoxyuridine (BrdU) is an isosteric photoreactive analog of

* To whom correspondence should be addressed

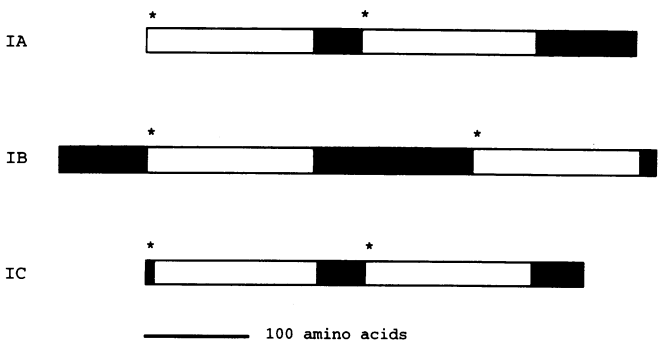

Figure 1. The organization of the S subunits of the three type I R–M system families. The regions conserved within a family are shown as shaded boxes while the ARD and CRD which recognize each part of the DNA target are shown as open boxes. Also shown is the repeated region (*) identified in the Results section in all type I S subunits.

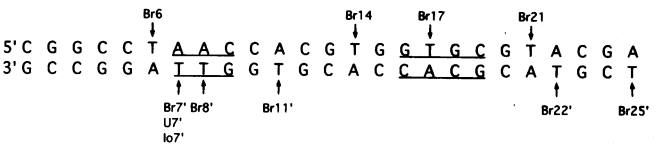

Figure 2. The 25 bp DNA duplex containing the *Eco*KI target sequence (underlined). This duplex is referred to as S25. The sites of substitution of thymine by modified bases are indicated and numbered from the 5′ end of the top strand with the addition of a prime (′) indicating a substitution on the bottom strand. The duplexes containing a single substitution of bromodeoxyuridine, deoxyuridine or iododeoxyuridine are designated Br6, Br7′, Br8′, Br11′, Br14, Br17, Br21, Br22′, Br25′, U7′ and Io7′. The duplex Br500 contains BrdU at all of the indicated sites.

thymidine (44). For 5-BrdU-substituted DNA, 5′-4,4′-dimethoxytrityl 5-bromo-2′-deoxyuridine-3′-*N,N*-diisopropyl cyanoethyl phosphoramidite was obtained from Link Technologies (UK). BrdU was incorporated either at a single position to give duplexes designated Br6, Br14, Br17, Br21, Br7′, Br8′, Br11′, Br22′ and Br25′ or at all available positions, designated duplex Br500, within the complementary oligonucleotides (Fig. 2). The unmodified duplex is called S25. The same DNA synthesis procedures were also used for the preparation of deoxyuridine and 5-iododeoxyuridine substituted DNA, duplexes U7′ and Io7′. All oligonucleotides were purified and analyzed by high performance liquid chromatography (HPLC). The oligonucleotides of the duplex were labelled respectively on the top strand, bottom strand, or both strands at the 5′-hydroxyl position using [γ-^{32}P]ATP (3000 Ci/mmol, Amersham) and T4 polynucleotide kinase (S. Bruce, ICMB). Unincorporated ATP was then removed using G25 spun column chromatography (45). Hybridization and DNA concentration determination were as described previously (46).

*Eco*KI mtase was prepared as described previously (6). Mtase stock solution was desalted by using a G25 Sephadex gel filtration column (Pharmacia, PD-10 column) and Tris–HCl buffer (20 mM Tris–HCl, pH 8.0, 100 mM NaCl). The concentration of mtase was determined by UV spectrophotometry using an extinction coefficient at 280 nm of 8.42 for a 1% (w/v) solution in a 1 cm pathlength cell (46).

Conditions for binding and UV crosslinking of substituted DNA

Gel retardation assays (46) were used to investigate the binding of the mtase with DNA oligonucleotides. To determine the optimum conditions for UV crosslinking, a solution of mtase (200 nM), AdoMet (100 μm) and DNA duplex (10 nM) in a total volume of 50 μl was put in a multiwell plate on ice and irradiated with a UV hand-held lamp. Different wavelengths of UV light and different irradiation times were investigated in order to obtain the optimum yield of crosslinked complex. In these experiments duplexes S25, Br7′, Br8′, Br11′, Br22′ and Br25′ were used. Crosslinking was examined by SDS–PAGE (47) with a 12% polyacrylamide gel of the mtase–DNA complexes, followed by silver staining (BioRad silver stain kit), or autoradiography and quantification of crosslinking efficiency by scintillation counting. To identify which one of the two DNA strands was crosslinked to protein, [γ-^{32}P]ATP was used to label one or other or both strands of the duplex. After the UV irradiation the crosslinked mixture was boiled for 2 min and cooled rapidly by quenching with liquid nitrogen before loading immediately on to a 10% polyacrylamide SDS–7 M urea gel. DNA duplexes are denatured by the boiling and quenching procedure, allowing the crosslinked DNA strands to be separated effectively from uncrosslinked DNA strands during electrophoresis on this gel.

For large scale experiments, DNA duplex Br7′ (20.4 μM), desalted mtase (8.8 μM) and AdoMet (70 μM) in a total volume of 4.6 ml were incubated for 15 min at 22°C. The reaction mixture was then spread to form a thin layer on the inside of the lid (12.5 cm × 8.5 cm) of a multiwell plate which was kept on ice for UV irradiation at 254 nm by UV handlamp set (Model UVGL-58, UVP International Ultra-Violet Products). Samples were directly irradiated by UV at a distance of 7 mm for 10 min. Crosslinking was monitored by autoradiography of SDS–PAGE gels of the irradiated samples and the efficiency of crosslinking was determined by scintillation counting of the appropriate gel slices.

Purification of large scale crosslinking complexes and trypsin digestion

The crosslinked complex was separated from the free DNA using a denaturing gel (9% polyacrylamide SDS–PAGE). The crosslinked band was cut from the gel, crushed and the DNA–protein complex eluted by shaking the crushed gel in 10 ml of elution buffer (0.5 M ammonium acetate, 1.0 mM EDTA, 0.1% SDS, 10.0 mM $MgCl_2$, pH 7.5 adjusted with NaOH) (48) at room temperature for 8 h. The crushed gel was removed by centrifugation. The solution of crosslinked complex was desalted on a G25 Sephadex gel filtration column, equilibrated with distilled water, before being freeze-dried.

The crosslinked complex was resuspended in 1 ml of ammonium acetate (200 mM, pH 8.5) or 10 mM Tris–HCl pH 8.0 and then digested by trypsin (Boehringer Mannheim). Possible chymotrypsin contamination of the trypsin stock solution (10 mg/ml trypsin in 100 mM Tris–HCl, 100 mM $CaCl_2$, pH 8.0) was inactivated using diphenyl carbamyl chloride (DPCC; Light & Co. Ltd) (49). A trypsin working solution was prepared by diluting the 10 mg/ml stock solution 10-fold with 0.2 M ammonium acetate pH 8.5 or 10 mM Tris–HCl pH 8.0. 1 ml of this 1 mg/ml trypsin working solution was added to the

resuspended crosslinked complex and incubated for 12 h at 37°C. The peptide–DNA crosslinked complex was purified from digested peptides and trypsin on an Econo-Pac Q ion exchange cartridge (BioRad). After sample loading the column was washed with salt-free phosphate buffer (20 mM KH_2PO_4, 5% ethanol, pH 6.8). A 100 ml gradient from 0 to 0.6 M KCl in phosphate buffer (pH 6.8), 5% ethanol was then applied with a flow rate of 48 ml/h. The radioactive fractions were determined with a Beckmann LS7000 scintillation counter by adding 1.0 µl of elution fractions to 2.5 ml of Ecoscint (National Diagnostics). Radiolabelled peptide–DNA fractions were pooled and desalted with a G25 Sephadex gel filtration column before freeze-drying. To obtain the shortest crosslinked peptide for sequencing, the purified digested sample was treated with trypsin a second time. The purified peptide–DNA was analyzed by SDS–PAGE and sequenced by Mr B. Dunbar and Prof. J. Fothergill at Aberdeen University and the WELMET service at Edinburgh University. The WELMET samples were resuspended in deionised water, filtered through a polyvinyl difluoride membrane in a Prospin sample preparation cartridge (Applied Biosystems). The complex was trapped on the filter, washed with deionised water and dried before sequencing (50).

RESULTS

Gel retardation experiments for the unmodified DNA duplex and all of the bromodeoxyuridine substituted duplexes showed that with 10 nM DNA, 200 nM mtase was sufficient to bind 84–94% of the DNA (Table 1). Therefore, any variation in crosslinking efficiency with these duplexes can be assigned to the efficiency of the photoreaction rather than altered binding affinity.

Table 1. Efficiency of binding and crosslinking of duplexes to mtase

Duplex	Duplex bound (%)		Duplex crosslinked (%)	
S25	84 ± 4	(5)	3.9 ± 0.2	(6)
Br500	92	(1)	12.2	(1)
Br6	83 ± 3	(9)	3.0	(1)
Br7′	79 ± 4	(5)	12.3 ± 0.2	(5)
Br8′	87 ± 7	(2)	4.3 ± 0.3	(2)
Br11′	77 ± 15	(2)	4.1 ± 0.2	(2)
Br14	74 ± 4	(5)	3.8	(1)
Br17	78 ± 4	(5)	3.0	(1)
Br21	75 ± 5	(5)	4.1	(1)
Br22′	84 ± 8	(2)	4.4 ± 0.1	(2)
Br25′	77 ± 13	(2)	4.3 ± 0.1	(2)

Binding and crosslinking efficiency are represented by the fraction of radiolabelled DNA, determined by scintillation counting, in the bound or crosslinked complex relative to the total amount of DNA. The number of measurements is shown in brackets.

Preliminary experiments showed that maximum crosslinking was achieved by 10 min irradiation at 254 nm (results not shown). The percentage crosslinking is 3–4.4% for all duplexes except duplexes Br500 and Br7′ (Table 1). The Br500 and Br7′ duplexes crosslinked with 12.2–12.3% efficiency. This strongly suggests that the Br7′ substitution is the site of crosslinking to the protein. It was confirmed that crosslinking occurred only to the bottom strand by radiolabelling each DNA strand separately and ensuring that the duplex was denatured during electrophoresis on 10% polyacrylamide SDS–7 M urea gels (results not shown). This implies that the 3–4.4% level of crosslinking observed with the other duplexes is due to crosslinking by the normal thymine base at position 7′. No crosslinking was observed when bovine serum albumin was substituted for the mtase, or when a non-specific duplex lacking the recognition sequence was used (results not shown).

The substitution of the thymine at position 7′ with uracil and iodouracil reduced the crosslinking efficiency to the mtase to ~23 and 30% respectively of that obtained with Br7′ (results not shown), i.e. not significantly different from unsubstituted DNA. Gel retardation experiments (Fig. 3) showed that the Br7′ duplex bound to the mtase as well as the unmodified duplex (dissociation constant, K_d <10 nM) but that the U7′ and Io7′ duplexes bound much more poorly with K_d >30 nM. The weaker binding affinities cannot, however, account for the lower crosslinking efficiency of these duplexes as, under the conditions used for crosslinking (680 nM mtase, 2.0 µM DNA, 60 µM AdoMet in 200 µl), the amount of complex formed for U7′ and Io7′ DNA was not significantly different from that for unsubstituted DNA (results not shown). The van der Waals radii of bromine (Br7′) and the methyl group (S25) are very similar (44) while those of hydrogen (U7′) and iodine (Io7′) are smaller and larger respectively. The differences in crosslinking efficiency are probably due, therefore, to the different size of the group substituted at the 5 position of the pyrimidine ring affecting the contacts at the DNA–protein interface. Therefore, in spite of iododeoxyuridine being intrinsically more photoreactive than BrdU (44), the large size of the iodine probably disrupts the DNA amino acid contacts needed for crosslinking.

Using the Br7′ duplex a large quantity of DNA–mtase crosslinked complex was prepared. The procedures for eluting the complex from the gel and desalting gave high recoveries of radiolabelled sample (90–97% and 89–93% for each step respectively). Digestion of the complex with trypsin and purification of the peptide–Br7′ resulted in 46–69% recovery for this step. The conditions used for the ion-exchange step eluted the peptide–Br7′ complex at 0.5 M KCl. The overall recovery of the final product was 2.0–2.5% of the total DNA used.

Figure 4 shows an analysis of the crosslinked complexes used for amino acid sequencing. Figure 4A shows a silver-stained denaturing gel of the mtase crosslinked to Br7′ duplex. Between the bands representing the S and M subunits of the mtase is a new band containing the radiolabelled duplex linked to the S subunit.

Figure 4B shows an autoradiograph of a SDS–20% acrylamide gel of the intermediate peptide–Br7′ complex (lane 4), and long peptide–Br7′ complex (lane 5), produced by digestion in the presence or absence of ammonium acetate respectively. The presence of ammonium acetate increased the solubility of the S-polypeptide–Br7′ complex and resulted in a more complete trypsin digest. Figure 4C shows the short peptide–Br7′ complex (lane 2) produced by twice digesting with trypsin, the long peptide–Br7′ complex (lane 3) and the S subunit–Br7′ complex (lane 4).

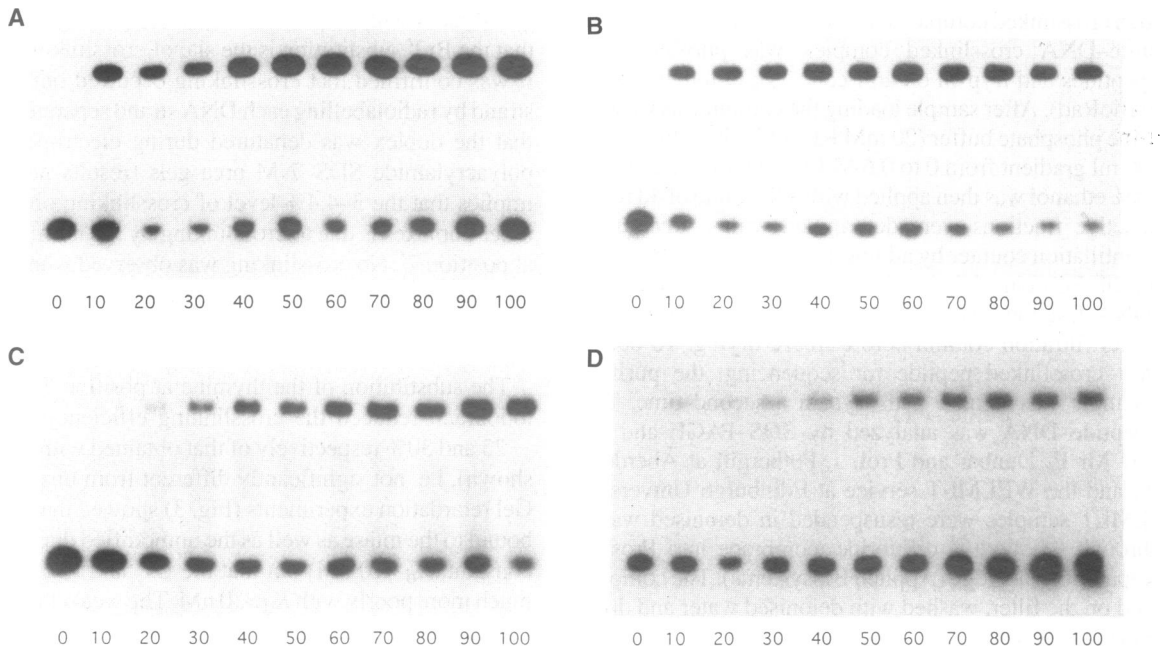

Figure 3. Binding of *Eco*KI mtase to labelled target sequences, as determined by gel retardation. Top band: DNA–mtase complex; lower band: free DNA. In all cases, 100 μM AdoMet and 0.1 nM target duplex DNA were used; the mtase concentration (nM) is noted below each lane. (**A**) Unmodified S25 duplex; (**B**) Br7′ duplex; (**C**) U7′ duplex and (**D**) Io7′ duplex.

Table 2. Amino acid sequences obtained from the crosslinked complexes containing duplex Br7′

*Eco*KI							
S subunit[a]	SAGKL	PEGWV	IAPVS	TVTTL	IRGVT	YKKEQ	AINY
Short peptide–Br7′					GVT	–KK	
Intermediate peptide–Br7′					–V–	–KKEQ	AINY
Long peptide–Br7′	– – –KL	PEG–V	IAPV–	TVTTL	I–GVT	– –KEQ	

[a]The S subunit in *Eco*KI lacks the N terminal methionine and the sequence starts at Ser2.
The first and third amino acids in the intermediate peptide–Br7′ complex and the first three amino acids of the long peptide–Br7′ complexes were not identifiable due to the presence of contaminants. The later sequencing cycles of the long peptide–Br7′ complex were difficult to interpret due to lack of material.
No amino acid corresponding to Tyr27 was identifiable in any of the peptide–Br7′ complexes sequenced.

The long peptide–Br7′ complex (Fig. 4B, lane 5; Fig. 4C, lane 3) and the intermediate peptide–Br7′ complex (Fig. 4B, lane 4) contained impurities, which made analysis of the first three amino acid sequencing cycles for the former and the first and third cycles for the latter impossible. This problem was avoided with the short peptide–Br7′ complex (Fig. 4C, lane 2) by absorbing it to a PVDF membrane and washing away the contaminants. The sequences obtained are shown in Table 2 along with the first 34 amino acids of the S subunit. The sequences obtained are uniquely found in the S subunit of *Eco*KI. Both the short and intermediate peptide–Br7′ complexes were produced by trypsin cutting after Arg23. The short peptide terminates after Lys29 but this site in the intermediate peptide has not been cut by trypsin, while in the long peptide there is no cutting after Lys5, Arg23 or Lys29. These incomplete digestions are possibly due to steric interference by the crosslinked DNA. All three peptides give blank sequencing cycles where Tyr27 is expected. Therefore, the bromodeoxyuridine at position 7′ in the DNA duplex has covalently crosslinked to Tyr27 in the S subunit.

The identification of the contact between Tyr27 and the first A–T base pair in the sequence AAC suggested that it might be conserved in other ARDs that recognize the AA dinucleotide within the trinucleotide target. Other ARDs which have a common dinucleotide sequence comprising CA or GA are also compared (Fig. 5A).

In the alignments shown in Figure 5A it is apparent that all of the ARDs have a conserved leucine (Leu21 in *Eco*KI). The two published alignments for *Sty*SBI are different (16,23), one containing the six residue gap after Leu21 (16) and the other not (23). We prefer the former alignment as this allows the line up of the *Sty*SBI Leu21 with the leucines of the other enzymes. Other alignments around Leu21 are probably equally valid and acceptable but we chose this line-up to provide us with a boundary between the LP-GWEW-containing structural regions of the S

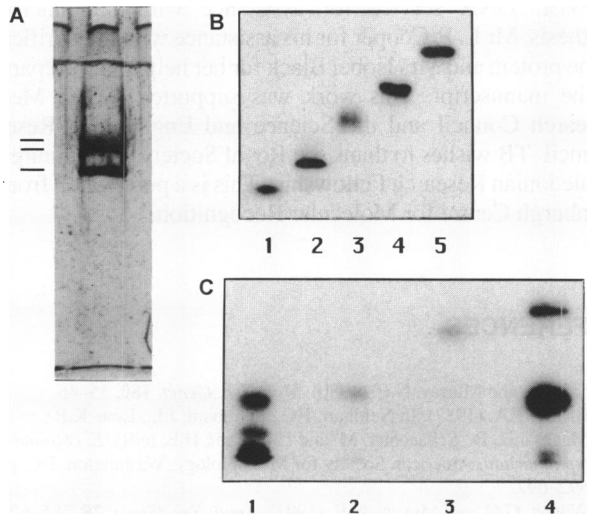

Figure 4. (**A**) A silver-stained SDS–12% polyacrylamide gel of mtase after crosslinking to duplex Br7′. The S subunit crosslinked to the duplex is visible between the S subunit (lower band) and the M subunit (upper band). (**B**) An autoradiogram of a 20% polyacrylamide SDS gel of the peptide–Br7′ crosslinked complexes after trypsin digestion and purification. Lanes 1, 2 and 3 are [γ-^{32}P]ATP, bottom strand of Br7′ and Br7′ duplex respectively. Lane 4 shows the intermediate peptide–Br7′ duplex complex obtained after trypsin treatment in the presence of ammonium acetate and lane 5 the long peptide–Br7′ duplex complex obtained after trypsin treatment in the absence of ammonium acetate. (**C**) An autoradiogram performed as in (B). The upper bands in lanes 1–4 are a ^{32}P-labelled Br7′ duplex, a double-digested short peptide–Br7′ crosslinked complex, a long peptide–Br7′ crosslinked complex and an S subunit–Br7′ crosslinked complex respectively. Additional bands are visible in lanes 1 and 4. In lane 1, the two bands running ahead of the ^{32}P-labelled Br7′ duplex are ^{32}P-labelled bottom strand of Br7′ and [γ-^{32}P]ATP. In lane 4 the two bands running ahead of the S subunit–Br7′ crosslinked complex are ^{32}P-labelled Br7′ duplex DNA and [γ-^{32}P]ATP, respectively.

subunit ARDs and the region containing conserved amino acids more likely to be important in DNA recognition, as discussed below.

The motif LP-GWEW is well-defined in all IB family enzymes and is retained in part in all of the other systems, except perhaps *Eco*BI. The small conserved region shown at the start of the IC family S polypeptides (Fig. 1) is also seen in the IA family sequences (Fig. 5A). This region in the IA family is apparent from previous alignments but was not commented on at that time (23). Sequences from the junction between the central conserved region and the CRD have also been aligned with the LP-GWEW motif (Fig. 5B). The repeat of the LP-GWEW motif in this region has been observed for members of the IB family (18), the IC family and *Eco*KI (24), but its presence in other type I systems, although apparent from earlier alignments (16,23), has also not been previously noted.

The conservation of components of the LP-GWEW motif in S subunits from all of the families regardless of their recognition site, and the fact that the motif is repeated, suggests that the amino acids shown in bold (but not underlined) prior to Leu21 of *Eco*KI have a structural role. After Leu21 other similarities occur only within the grouping of enzymes according to common dinucleotides in their recognition sequence (bold underlined amino acids in Fig. 5A), which suggests that these amino acids have a DNA recognition rather than a structural role. Of particular note are Tyr27, Asp39, Tyr40, Gly52 and Lys53, which are conserved in the two ARDs that recognise AAC, but also in the ARD of *Eco*R124I that recognises GAA, even though this system is in a different type I family from the other two. It should be noted that the second adenine in this trinucleotide is the substrate for methylation by *Eco*R124I (51). Therefore it is possible that the Tyr27 in the S subunit of *Eco*R124I also contacts the first A–T base pair in the GAA target sequence.

DISCUSSION

UV crosslinking studies have been widely used to investigate DNA–protein interactions. Many have simply exploited the intrinsic photoreactivity of the nucleotides, in particular thymine (52–55). Other approaches have relied on nucleotide analogs with enhanced photoreactivity (36–42,56–60). The most widely used of these has been BrdU in which the 5-methyl group of thymidine, which projects into the DNA major groove, is substituted by the similarly sized bromine atom (36–42). The crosslinking reaction proceeds following the loss of the bromine atom to give a free radical intermediate which can react with suitable acceptors, including amino acid side chains (60).

A

Type I R-M system	Trinucleotide recognition sites		Alignment of the first 60 amino acids of the S subunit ARDs (21 27 39,40 52,53)
(IA) *Eco*KI	AA*C		MSAGKLPEGWVIAPVSTVTT L IRGVT Y KKEQAINYLKD DY LPLIRANNIQN GK FDTTDLV
(IA) *Sty*SPI	AA*C		MNRGKLPEGWATAPVSTVTT L IRGVT Y KKEQALNYLQD DY LPIIRANNIQN GK FDTTDLV
(IC) *Eco*R124I	GAA*	MSEMSYLEKLLDGVEVEWLP	L GEITK Y EQPTKYLVKAK DY HDTYTIPVLTA GK TFILGYT
(IA) *Sty*SBI	GA*G		MSGGKLPEGWATSTINEMCN L M PKLKL D DDLDVGEMPMAGVPTTYLGKCNFETK
(IB) *Eco*AI	GA*G	(95) EKPPFELPDGWEWIT	L TRIAEI M PKIDVS D DEQEISFIPMPLISTKFDGSHEFEIK
(IB) *Eco*EI	GA*G	(95) EKPPFELPEGWEWIT	L SEIATI M PKIEVT D DEQEISFVPMPCISTRFDGAHDQEIK
(IA) *Eco*BI	TGA*	MSFNSTSKE	L IEQN I M GLLSIH D SWLPISMDSVANITNGEAFKSSEFNN
(IB) *Cfr*AI	GCA*	(95) DKPPFELPAGWEWVR	L GEAFYIEM G QSPSSQYYNQSEEGIPFFQGKADFGKKYPT
(IC) *Eco*prrI	CCA*	MSELSYLEKLMDGVEVEWLP	L SKVFNLRN G YTPSKTKKEPWANGDIPWFRMDDIRENGRI
(IC) *Eco*DXXI	TCA*	MSELSYLEKLLDGVEVEWVT	L GSMADIGT G SSNRQDESENGIYPFYVRSKNILKSDTFEF
(IA) *Eco*DI	TTA*	MSAGKLPVDWKTVELGELIK	L STGKLDANAADNDGQYPFFTCAESVSQINSWAFDTSAVL

B

Type I R-M system	Alignment of amino acids at start of S subunit CRDs
(IB) A	(389)LPEGWEW
(IB) E	(389)LPIGWEW
(IB) Cfr	(380)LPEGWEW
(IA) K	(208)LTEKWRN
(IA) D	(179)LTKEWRS
(IA) B	(209)LTKEDKD
	(230)IPEDWSE
(IA) SP	(208)LTTALRN
	(217)SLIGWHS
(IA) SB	(204)LTKELHK
	(223)IPSLWKI
(IC) R124	(200)KEGEVEW
(IC) DXXI	(196)KEGEVEW
(IC) prr	(216)KEDEVEG

Figure 5. (**A**) A comparison of the first 60 amino acids of the ARDs of type I R–M systems. The family to which each belongs is shown in brackets. Each group comprises ARDs which share a common dinucleotide including the methylation site (A*). The N terminal 95 amino acids of the IB family S subunits which are not part of the ARD are omitted. The amino acids in bold type, up to and including the conserved leucine (Leu21 in *Eco*KI), are repeated at the start of the CRD, and are proposed to have a structural role. The underlined bold amino acids indicate conservations within each dinucleotide recognition group, including the Tyr27 crosslinking site in *Eco*KI, which may be involved in recognizing the nucleotide preceding the methylation target. Other amino acids conserved within a group are underlined. (**B**) This shows amino acids at the start of the CRDs that line up with the amino acids at the start of the ARD preceding the conserved leucine (Leu21 in *Eco*KI). Two alignments are shown for *Eco*BI, *Sty*SPI and *Sty*SBI, the first in each case is as previously published (23), while the second is an alternative which is in better agreement with the *Eco*AI consensus, LPEGWEW. The number of amino acids preceding the shown sequence is given in brackets.

From our results we conclude that the methyl group of thymine 7′ is in close proximity to Tyr27 in the *Eco*KI S subunit ARD. Substitution of the methyl group with the similarly sized bromine has no effect on DNA binding and increases the crosslinking yield, but the substitution with the smaller hydrogen in the U7′ duplex or the larger iodine in the Io7′ duplex reduces both binding affinity and crosslinking efficiency. In the case of the bromodeoxyuridine-substituted oligonucleotides, crosslinking only occurred between bromine at position 7′ in the duplex and Tyr27, possibly because there are no suitable reactive amino acids near to the other substitution sites.

The amino acids crosslinked to DNA via BrdU in several DNA–protein complexes have been identified (38–40) and found to correspond to important amino acids in the complex structure. For example, for *Eco*RI, methionine 137 identified at the site of crosslinking is in a region of the molecule previously implicated in DNA binding and cleavage (38) and the subsequent co-crystal structure showed that amino acid residues Met137–Gly140 make important contacts to the two outer pyrimidines of the recognition site (28,31). In the *lac* repressor protein, amino acids that crosslinked to operator DNA substituted with BrdU are consistent with NMR evidence and molecular dynamic calculations (39), while for the basic region leucine zipper (bZIP) DNA-binding motif of GCN4, a yeast transcriptional activator, the demonstration that Ala238 crosslinked to a BrdU in its binding site (40), is in agreement with the crystal structure that showed contact of Ala238 with the corresponding thymine 5-methyl group in the GCN4 bZIP element–DNA complex (61).

The demonstration that Tyr27 in the *Eco*KI S subunit ARD is in close proximity to the DNA sequence recognition site is consistent with genetic work that correlated the ARD with the recognition of the trinucleotide part of the DNA target (17). The result also gives evidence for a contact with thymine in the major groove of the DNA, a conclusion consistent with methylation interference experiments, indicating that the mtase makes major groove contacts to both halves of the bipartite recognition sequence (62).

The ARDs of different type I systems show little similarity unless they recognise the same trinucleotide (16–18). However, there is some similarity for the first part of the ARDs from all of the families, with a motif that is repeated at the beginning of the C-terminal conserved region (Fig. 5; 24) suggesting a structural role for the amino acids shown in bold prior to Leu21 of *Eco*KI (Fig. 5A). After Leu21, the sequence comparisons of the first portions of ARDs that have a common dinucleotide sequence hint that some amino acids (underlined and bold in Fig. 5A) may be conserved to allow recognition of this dinucleotide part of the DNA target. In particular, it is possible that Tyr27 in the S polypeptide of the *Eco*R124I system plays the same role in contacting the AA dinucleotide sequence as it does in the *Eco*KI system. If this hypothesis could be confirmed by, for example, crosslinking experiments with *Eco*R124I, and mutagenesis of Tyr27 in *Eco*KI and *Eco*R124I, then it might suggest that the target recognition domains of the S subunits may have similar tertiary structure features for recognizing parts of their DNA targets.

ACKNOWLEDGEMENTS

We thank Prof. J. Fothergill and Mr B. Dunbar, University of Aberdeen and Welmet Protein Characterisation Facility, University of Edinburgh, for the peptide sequencing. We thank the OSWEL DNA service for assistance with oligonucleotide synthesis, Mr L. P. Cooper for his assistance with the purification of the protein and Mrs Isobel Black for her help in the preparation of the manuscript. This work was supported by the Medical Research Council and the Science and Engineering Research Council. TB wishes to thank the Royal Society of Edinburgh for a Caledonian Research Fellowship. This is a publication from the Edinburgh Centre for Molecular Recognition.

REFERENCES

1 Sain, B. and Murray, N.E. (1980). *Mol. Gen. Genet.* **180**, 35–46.
2 Bickle, T.A. (1987). In Neidhart, F.C., Ingraham, J.L., Low, K.B., Magasanik, B., Schaechter, M. and Umbarger, H.E. (eds), *E. coli and S. typhimurium*. American Society for Microbiology, Washington, DC, pp. 692–697.
3 Wilson, G.G. and Murray, N.E. (1991). *Annu. Rev. Genet.* **25**, 585–627.
4 Bickle, T.A. and Krüger, D.H. (1993). *Microbiol. Rev.* **57**, 434–450.
5 Suri, B., Nagaraja, V. and Bickle, T.A. (1984). *Curr. Top. Microbiol. Immunol.* **108**, 1–9.
6 Dryden, D.T.F., Cooper, L.P. and Murray, N.E. (1993). *J. Biol. Chem.* **268**, 13228–13236.
7 Kan, N.C., Lautenberger, J.A., Edgell, M.H. and Hutchison, C.A. (1979). *J. Mol. Biol.* **130**, 191–209.
8 Bickle, T.A., Brock, C. and Yuan, R. (1978). *Proc. Natl. Acad. Sci. USA* **75**, 3099–3103.
9 Bickle, T.A. (1982). In Linn, S.M. and Roberts, R.J. (eds), *Nucleases*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, pp. 85–108.
10 Vovis, G.F., Horiuchi, K. and Zinder, N.D. (1973). *Proc. Natl. Acad. Sci. USA* **71**, 3810–3813.
11 Burckhardt, J., Weisemann, J. and Yuan, R. (1981). *J. Biol. Chem.* **256**, 4024–4032.
12 Meselson, M. and Yuan, R. (1968). *Nature* **217**, 1111–1114.
13 Boyer, H.W. and Roulland-Dussoix, D. (1969). *J. Mol. Biol.* **41**, 459–472.
14 Glover, S.W. and Colson, C. (1969). *Genet. Res.* **13**, 227–240.
15 Fuller-Pace, F.V., Cowan, G. and Murray, N.E. (1985). *J. Mol. Biol.* **186**, 65–75.
16 Cowan, G.M., Gann, A.A.F. and Murray, N.E. (1989). *Cell* **56**, 103–109.
17 Fuller-Pace, F.V. and Murray, N.E. (1986). *Proc. Natl. Acad. Sci. USA* **83**, 9368–9372.
18 Kannan, P., Cowan, G.M., Daniel, A.S., Gann, A.A.F. and Murray, N.E. (1989). *J. Mol. Biol.* **209**, 335–344.
19 Gubler, M., Braguglia, D., Meyer, J., Piekarowicz, A. and Bickle, T.A. (1992). *EMBO J.* **11**, 233–240.
20 Nagaraja, V., Shepherd, J.C.W., Pripfl, T. and Bickle, T.A. (1985). *Nature* **316**, 371–372.
21 Gough, J. A. and Murray, N.E. (1983). *J. Mol. Biol.* **166**, 1–19.
22 Fuller-Pace, F.V., Bullas, L.R., Delius, H. and Murray, N.E. (1984). *Proc. Natl. Acad. Sci. USA* **81**, 6095–6099.
23 Gann, A.A.F., Campbell. A.J.B., Collins, J.F., Coulson, A.F.W. and Murray, N.E. (1987). *Mol. Microbiol.* **1**, 13–22.
24 Kneale, G.G. (1994) *J. Mol. Biol.* **243**, 1–5.
25 Meister, J., MacWilliams, M., Hübner, P., Jutte, H., Skrzypek, E., Piekarowicz, A. and Bickle, T.A. (1993). *EMBO J.* **12**, 4585–4591.
26 Abadjieva, A., Webb, M., Patel, J., Zinkevich, V. and Firman, K. (1994). *J. Mol. Biol.* **241**, 35–43.
27 Cooper, L.P. and Dryden, D.T.F. (1994). *J. Mol. Biol.* **236**, 1011–1021.
28 Newman, M., Strzelecka, T., Dorner, L.F., Schildkraut, I. and Aggarwal, A.K. (1994). *Nature* **368**, 660–664.
29 Newman, M., Strzelecka, T., Dorner, L.F., Schildkraut, I. and Aggarwal, A.K. (1994). *Structure* **2**, 439–452.
30 Athanasiadis, A., Vlasi, M., Kotsifaki, D., Tucker, P.A., Wilson, K.S. and Kokkinidis, M. (1994). *Nature Struct. Biol.* **1**, 469–475.
31 Kim, Y., Grable, J.C., Love, R., Greene, P. and Rosenberg, J.M. (1990) *Science* **249**, 1307–1309.
32 Winkler, F.K., Banner, D.W., Oefner, C., Tsernoglou, D., Brown, R.S., Heathman, S.P., Bryan, R.K., Martin, P.D., Petratos, K. and Wilson, K.S. (1993). *EMBO J.* **12**, 1781–1795.
33 Cheng, X., Balendiran, K., Schildkraut, I. and Anderson, J.E. (1994). *EMBO J.* **13**, 3927–3935.

34 Klimasauskas, S., Kumar, S., Roberts, R.J. and Cheng, X. (1994). *Cell* **76**, 357–369.
35 Labahn, J., Granzin, J., Schluckebier, G., Robinson, D.P., Jack, W.E., Schildkraut, I. and Saenger, W. (1994). *Proc. Natl. Acad. Sci. USA* **91**, 10957–10961.
36 Lin, S-Y. and Riggs, A.D. (1974). *Proc. Natl. Acad. Sci. USA* **71**, 947–951.
37 Ogata, R. and Gilbert, W. (1977). *Proc. Natl. Acad. Sci. USA* **74**, 4973–4976.
38 Wolfes, H., Fliess, A., Winkler, F. and Pingoud, A. (1986). *Eur. J. Biochem.* **159**, 267–273.
39 Allen, T.D., Wick, K.L. and Matthews, K.S. (1991). *J. Biol. Chem.* **266**, 6113–6119.
40 Blatter, E.E., Ebright, Y.W. and Ebright, R.H. (1992). *Nature* **359**, 650–652.
41 Ahmad, I. and Rao, D.N. (1994). *J. Mol. Biol.* **242**, 378–388.
42 Willis, M.C., LeCuyer, K.A., Meisenheimer, K.M., Uhlenbeck, O.C. and Koch, T.H. (1994). *Nucleic Acids Res.* **22**, 4947–4952.
43 Beaucage, L. and Caruthers, M.H. (1981). *Tetrahedra Lett.* **22**, 1859–1862.
44 Hutchinson, F. (1973). *Ann. Rev. Biophys.* **6**, 201–246.
45 Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual*, 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.
46 Powell, L.M., Dryden, D.T.F., Willock, D.F., Pain, R.H. and Murray, N.E. (1993). *J. Mol. Biol.* **234**, 60–71.
47 Laemmli, U.K. (1970). *Nature* **227**, 680–685.
48 Papavassiliou, A.G. (1994). In Kneale, G.G. (ed), *DNA–Protein Interactions: Principles and Protocols. Methods in Molecular Biology*, Vol. 30, Humana Press, New Jersey, pp. 43–78.
49 Erlanger, B.F., Cooper, A.G. and Cohen, W. (1966). *Biochem. J.* **5**, 190–196.
50 Hayes, J.D., Kerr, L.A. and Cronshaw, A.D. (1989). *Biochem. J.* **264**, 437–445.
51 Taylor, I., Watts, D. and Kneale, G. (1993). *Nucleic Acids Res.* **21**, 4929–4935.
52 Williams, K.R. and Konigsberg, W.H. (1991). *Methods Enzymol.* **208**, 516–539.
53 Paradiso, P.R., Nakashima, Y. and Konigsberg, W. (1979). *J. Biol. Chem.* **254**, 4739–4744.
54 Merrill, B.M., Williams, K.R., Chase, J.W. and Konigsberg, W.H. (1984). *J. Biol. Chem.* **259**, 10850–10856.
55 Cleghon, V. and Klessig, D.F. (1992). *J. Biol. Chem.* **267**, 17872–17881.
56 Evans, R.K., Johnson, J.D. and Haley, B.E. (1986). *Proc. Natl. Acad. Sci. USA* **83**, 5382–5386.
57 Meffert, R. and Dose, K. (1988). *FEBS Lett.* **239**, 190–194.
58 Meffert, R., Rathgeber, G., Schüfer, H.J. and Dose, K. (1990). *Nucleic Acids Res.* **18**, 6633–6636.
59 Nikiforov, T.T. and Connolly, B.A. (1992). *Nucleic Acids Res.* **20**, 1209–1214.
60 Welsh, J. and Cantor, C.R. (1985). *Trends Biochem. Sci.* **8**, 505–508.
61 Ellenberger, T.E., Brandl, C.J., Struhl, K. and Harrison, S.C. (1992). *Cell* **71**, 1223–1227.
62 Powell, L.M. and Murray, N.E. (1995) *Nucleic Acids Res.* **23**, 967–974.