

A prediction of the amino acids and structures involved in DNA recognition by type I DNA restriction and modification enzymes

Shane S. Sturrock and David T. F. Dryden*

Institute of Cell and Molecular Biology, The King's Buildings, University of Edinburgh, Mayfield Road, Edinburgh EH9 3JR, UK

Received June 3, 1997; Revised and Accepted July 22, 1997

ABSTRACT

The S subunits of type I DNA restriction/modification enzymes are responsible for recognising the DNA target sequence for the enzyme. They contain two domains of approximately 150 amino acids, each of which is responsible for recognising one half of the bipartite asymmetric target. In the absence of any known tertiary structure for type I enzymes or recognisable DNA recognition motifs in the highly variable amino acid sequences of the S subunits, it has previously not been possible to predict which amino acids are responsible for sequence recognition. Using a combination of sequence alignment and secondary structure prediction methods to analyse the sequences of S subunits, we predict that all of the 51 known target recognition domains (TRDs) have the same tertiary structure. Furthermore, this structure is similar to the structure of the TRD of the C5-cytosine methyltransferase, *HhaI*, which recognises its DNA target via interactions with two short polypeptide loops and a β strand. Our results predict the location of these sequence recognition structures within the TRDs of all type I S subunits.

INTRODUCTION

A major aim of many studies of sequence specific protein-DNA interactions has been to determine how certain sequences of amino acids can recognise, with great fidelity, a DNA target sequence. Structural analysis of protein-DNA complexes has shown how α helices, β strands and loops can be used to give sequence specificity (1-3).

DNA methyltransferases (mtases) of restriction/modification (R/M) systems use target recognition domains (TRD) of 50-150 amino acids to recognise their DNA target sequence (4). The TRD is the major determinant in DNA target specificity with separate catalytic domains being required for enzymatic activity. The crystal structures of two monomeric type II C5-cytosine mtases, *HhaI* and *HaeIII*, bound to their DNA targets show that the TRD uses a conserved structure comprising two loops and one β strand to accomplish sequence recognition (5,6). The amino acid sequences of TRDs of many different C5-cytosine DNA

mtases have been compared. The level of sequence identity in these comparisons is very low and confined to several very short amino acid sequences corresponding to the recognition region in the two crystal structures (7,8), however, experimental support has been obtained for the involvement of this region in DNA recognition by C5-cytosine mtases other than *HhaI* and *HaeIII* (8). The N6-adenine and C4-cytosine mtases also contain a TRD and a catalytic domain (9), however, no cocrystal structure of one of these enzymes with DNA has been solved. A model of DNA recognition by the *TaqI* N6-adenine mtase, whose structure is known in the absence of DNA, has been constructed (10,11).

All characterised type I R/M systems recognise N6-adenine methylation of a bipartite target sequence (12-14). They are large, oligomeric, multifunctional enzymes encoded by the *hsdR*, *M* and *S* genes, combining both restriction endonuclease (R) and modification mtase (M) subunits with a DNA specificity (S) subunit. Type I R/M systems of enteric bacteria have been grouped into four families based on subunit complementation, DNA hybridisation and antibody cross-reactivity experiments (14,15). The amino acid sequence identity is very high within a family for the R, M and parts of the S subunit outwith the TRDs (16-21). This is believed to reflect conservation of residues in the subunit interfaces and the nuclease and mtase catalytic sites.

The S subunits of type I R/M systems contain two TRDs of 150-180 amino acids (12-14). Each TRD is responsible for recognising one of the two parts of the bipartite DNA target. The amount of amino acid sequence conservation between TRDs is either below ~25% for TRDs recognising different targets, or 40-90% when a target is shared dependent on whether the S subunits are in a different or the same family. The remainder of the ~50 kDa S subunit contains amino acid sequences which show a high degree of conservation between type I systems. These regions are responsible for defining the length of the non-specific DNA spacer in between the two TRD target sequences (22) and for binding the M and R subunits (23-26). It is believed that each TRD fits into the major groove to recognise the DNA, and the M subunits are arranged on either side of the S subunit allowing them to encircle the DNA and gain access to the methylation targets (27,28). Methylation is predicted to occur via a base flipping mechanism in which the adenine base is displaced out of the DNA helix into the catalytic pocket of the M subunits. This implies that the M subunit is on the other side of the DNA helix from the S subunit.

*To whom correspondence should be addressed. Tel: +44 131 650 7053; Fax: +44 131 650 8650; Email: david.dryden@ed.ac.uk

Table 1. S subunits of type I restriction/modification systems

Family	Name ^a	DNA target if known	S subunit length	1st TRD approximate location	2nd TRD approximate location	Reference ^b
IA	<i>EcoKI</i>	AAC N6 GTGC	464	11-157	214-368	29
IA	<i>EcoBI</i>	TGA N8 TGCT	474	11-158	215-379	30,31,32
IA	<i>EcoDI</i>	TTA N7 GTCY	444	11-128	185-348	33
IA	<i>StyLTIH</i>	GAG N6 RTAYG	469	11-153	209-375	34
IA	<i>StySPI</i>	AAC N6 GTRC	463	11-157	214-367	34
IA	<i>EcoR5I</i>		>140	1-140		35,36
IA	<i>EcoR10I</i>		>131	1-131		35,36
IA	<i>EcoR13I</i>		>152	1-152		35,36
IB	<i>EcoAI</i>	GAG N7 GTCA	589	110-247	403-540	37,38
IB	<i>EcoEI</i>	GAG N7 ATGC	594	109-247	403-545	17
IB	<i>CfrAI</i>	GCA N8 GTGG	578	108-236	385-529	18
IB	<i>StySKI</i>	CGAT N7 GTTA	587	108-249	396-538	39
IB	<i>StySTI</i>		>146	1-146		36
IB	<i>EcoR17I</i>	ATR....	>138	1-138		35,36
IC	<i>EcoR124I</i>	GAA N6 RTCG	409	25-142	215-350	40
IC	<i>EcoDXXI</i>	TCA N7 RTTC	406	23-139	211-341	20
IC	<i>EcoprI</i>	CCA N7 RTGC	405	22-159	232-360	41
ID	<i>StySBLI</i>	CGA N6 TACC	434	1-153	229-405	15
ID	<i>EcoR9I</i>		464	1-188	264-435	35, N. Murray pers. comm.
ID	<i>KpnAI</i>		439	1-155	231-410	42, J. Ryu pers. comm.
IC?	<i>BsuCI</i>	GAY N7 TGGA	405	23-162	219-355	43, G. Xu and T. Trautner, pers. comm.
IC?	<i>MpuAI</i>		401	1-139	221-359	44
IC?	<i>MpuBI</i>		336	1-139	188-324	44
ID?	HI0216		385	20-138	198-333	45
ID?	HI1286		459	1-176	268-445	15,45
?	mj0130		425	23-161	231-371	46
?	mj1218		425	28-155	226-368	46
?	mj1531		425	28-170	241-371	46

^a*Eco*, *Escherichia coli*; *Sty*, *Salmonella enterica*; *Cfr*, *Citrobacter freundii*; *Bsu*, *Bacillus subtilis*; *Kpn*, *Klebsiella pneumoniae*; *Mpu*, *Mycoplasma pulmonis*; HI, *Haemophilus influenzae* gene number; mj, *Methanococcus jannaschi* gene number. The systems in *H. influenzae* and *M. jannaschi* are putative type I systems.

^bWhere the target sequence is known, the reference for this work is given.

The TRDs of type I S subunits recognise a wide variety of 3, 4 or 5 bp targets and it would be of interest to define which amino acids within the large and highly variable sequence of the TRDs are responsible for sequence specificity. In this paper we use amino acid sequence alignment combined with secondary structure prediction methods. The use of secondary structure predictions enhances the strength of the amino acid alignment making distant similarities more apparent. These alignments of the TRDs suggest that all have the same tertiary structure and that they are the products of divergent evolution. A comparison of the secondary structure predictions with the known structure of the TRD of the *HhaI* mtase shows a strong similarity which has allowed us to define potential DNA recognition loops for all of the type I TRDs and to model the tertiary structure of these domains in a manner amenable to experimental verification.

MATERIALS AND METHODS

Most nucleotide or amino acid sequences of the S subunits were obtained from published references or GenBank and a database

constructed which separated the sequences into TRDs and conserved spacer regions. The amino acid sequences for the S subunits of *BsuCI* and *KpnAI* were generously provided by Prof. T. Trautner and Dr G. Xu (Berlin) and Dr J. Ryu (Loma Linda). The locations of the TRDs in the S subunit sequence and, if known, their DNA target and type I family are given in Table 1. The amino terminal TRD and carboxyl terminal TRD are indicated by the suffix -1 or -2 appended to the type I system's name in Figure 1.

A database of TRDs was made by separating conserved and unconserved regions of the S subunits and discarding the conserved regions. This database was inverted (every member was compared with every other one) using *sss_align* (47, S. Sturrock and A. Coulson, manuscript in preparation), a new implementation of the Smith and Waterman algorithm (48) using the Dayhoff PAM scoring scheme (49). The Smith and Waterman algorithm is a mathematically rigorous and exhaustive method to optimally align a pair of sequences. Each TRD sequence, along with its closest homologues, was then sent to the PHD program (50,51)

and a secondary structure prediction acquired. PHD is a secondary structure prediction method which uses a neural network trained with a set of known tertiary structures combined with multiple sequence alignment. On average secondary structure is predicted with >70% confidence. Each prediction was then placed in a new database and again inverted using *sss_align* but this time including the secondary structure prediction as well as amino acid sequence. *sss_align* takes into account the reliability index assigned by PHD on a residue by residue basis. This database inversion was performed with the QVAL parameter set at 80, PAMS 150, GAPOPEN 20 and GAPEXTEND 4. The QVAL parameter sets the amount of weighting given to the secondary structure when performing the sequence alignment. A value of 0 means the alignment uses only secondary structure information while a value of 100, as used in the first database inversion above, indicates that only sequence information is used in the alignment. A value of 80 was found to give the optimal statistical significance to the alignments of type I TRDs. In this instance, *sss_align* performs better than normal sequence alignment methods in aligning two distantly related sequences because the addition of secondary structure information, whether derived from a real structure or a prediction as in this case, is used to help the alignment pass through regions of very low sequence identity. The output of *sss_align* was again used to cluster sequences of high similarity and overlap these clusters with others until nearly all the TRD sequences were successfully aligned. Some TRD sequences could not be inserted into this alignment by the program due to a lack of obvious homology and these were aligned manually. These sequences are indicated in Figure 1 by an asterisk after the TRD name. In addition, *sss_align* also aligned the known tertiary structure of *HhaI* mtase (5) with the *EcoKI*-1 TRD to provide a key for the prediction of the location of loops and strands involved in DNA recognition. In this case, settings of PAMS 250, GAPOPEN 8, GAPEXTEND 1 and QVAL 60 were used to obtain the optimal alignment. These values were used successfully in the CASP2 competition (see below). To obtain a measure of the statistical significance of our alignment of type I TRDs with *HhaI*, we enlarged our database of TRDs by merging it with the approximately 2000 unique sequences with known tertiary structures in the protein databank. This enlarged database was then searched using *sss_align* and the TRD of *HhaI* to find homologous structures.

sss_align can be accessed at http://www.icmb.ed.ac.uk/sss_align/. Using secondary structure information from known structures, *sss_align* has been shown to successfully align sequences with only 15% amino acid identity (personal communication A. Coulson; CASP2, Second meeting on the critical assessment of techniques for protein structure prediction on World Wide Web URL: <http://iris4.carb.nist.gov/casp2/>). *sss_align* also adjusts for the variation in the reliability of the secondary structure predictions by using the residue by residue reliability of PHD predictions. This allows the program to align sequences even if parts have incorrectly predicted secondary structure.

RESULTS

Normal sequence alignment methods have been applied to complete S subunit sequences in the past (16,52). These studies were hampered not only by the limited number of sequences available but also by the high degree of sequence similarity in the conserved regions of the subunits. These restricted areas of high

homology almost totally obscured any sequence similarity between TRDs except when the TRDs recognised identical DNA targets whereupon the similarity was so high as to preclude any prediction of amino acids involved in sequence recognition.

The result of applying *sss_align* (47) to the TRDs of type I S subunits is shown in Figure 1. In contrast to previous analyses, we were able to observe considerable sequence similarity between TRDs even if the TRDs recognised different DNA targets. The sequence identity between homologous TRDs ranged between 20 and 45% except for the TRDs which recognise the same DNA targets where identity was much greater (16,17). These sequence identities extended over ~50–100% of the TRD sequence. All of the TRDs had at least one homologue with this level of identity except those TRDs indicated by an asterisk in Figure 1. Further smaller segments of TRDs with high levels of identity were also found identifying more distantly related sequences. Sequence alignment combined with the input of secondary structure prediction clustered the TRDs into several groups of high homology. Sequence information alone would have been sufficient to perform this alignment but the secondary structure prediction increased the significance of the alignment and frequently extended the lengths of the aligned segments. We were then able to align the clusters of homologous sequences with each other by using their predicted secondary structures and any small segments of sequence identity which existed between the clusters. The TRDs without any homologous sequences were then aligned manually by trying to obtain the best end to end match of predicted secondary structures and short local sequence similarities. β strand 1 was used as the centre point for the manual alignment of all groups of sequences. Using this predicted β strand to 'lock' the sequences together, it is apparent that many other predicted secondary structure features, particularly loops 1 and 2 and β strand 2, then become aligned even when sequences are very distantly related. Close inspection of Figure 1 suggests that some of the sequence alignments could be further 'improved' by the manual introduction of small gaps or deletions. Overall, we believe that these results are suggestive of a common tertiary structure for the TRDs of type I S subunits.

We compared our alignment of all 51 TRDs with the known sequence and secondary structure of the TRD of the C5-cytosine mtase *HhaI* (Fig. 1) (5). To our surprise, the two loops and one β strand which are responsible for the recognition of the DNA target of *HhaI* matched very well with our alignment of type I TRDs if the β strand preceding the first recognition loop in *HhaI*, is aligned with β strand 1 of our alignment of type I TRDs. A database containing all known unique protein structures as well as the TRDs and their predicted secondary structures, was searched with the TRD of *HhaI*. Statistically significant matches were only found with type I TRDs. The best alignment, with only a 10^{-10} probability of occurring by chance, was against the first TRD of *EcoKI*. Figures of 14% sequence identity, 51% sequence conservation and 56% structural conservation over 73% of the total length of the TRD were obtained for the alignment shown at the top of Figure 1. Although the level of sequence identity between *HhaI* and *EcoKI*-1 is very low, a similar value was found in one of the successful applications of *sss_align* in the CASP2 competition (personal communication, A. Coulson). Therefore, we are very confident that the alignment between *HhaI* and the first TRD of *EcoKI* is correct.

Figure 2 shows the recognition of the DNA phosphate backbone and bases by part of the TRD of *HhaI* mtase. In *HhaI*,

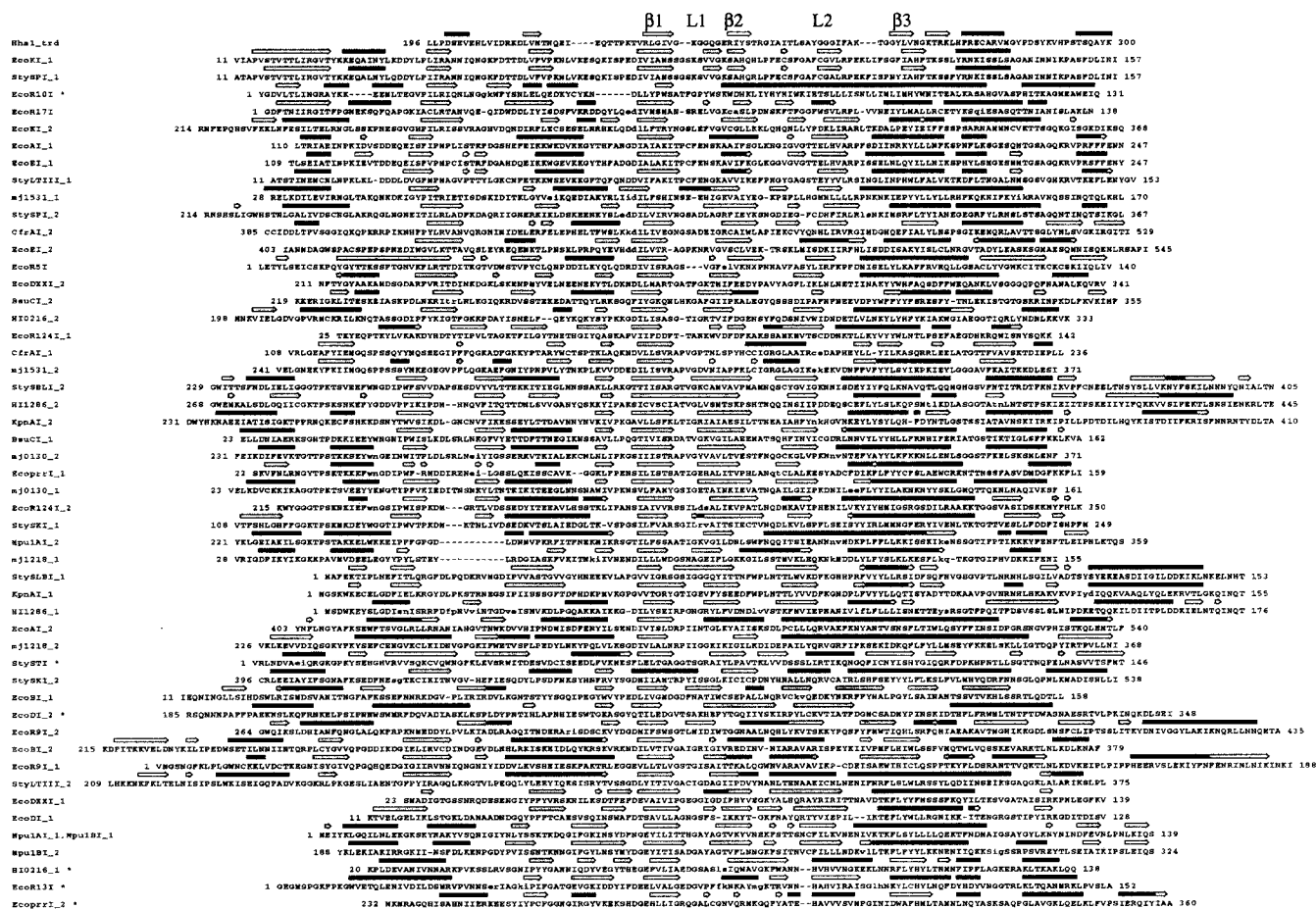


Figure 1. Multiple sequence alignments and predicted secondary structures for 51 TRDs from type I S subunits. β strands are shown as yellow arrows and α helices as red rectangles. The sequence and known secondary structure of the TRD of *HhaI* type II mtase are shown at the top of the diagram followed by the type I TRDs arranged such that they are closest relatives as determined by sss_align. β strands 1, 2 and 3, and loops 1 and 2 which are involved in DNA recognition by *HhaI* and predicted to have the same role in type I TRDs, are marked above the *HhaI* secondary structure. Strands 1, 2 and 3 comprise amino acids 228–231, 240–243 and 264–267 in *HhaI* respectively. In the absence of DNA, amino acids 250–253 in *HhaI* also form a β strand.

loop 1 (Val232–Glu239) fills the major groove and positions Gln237 into the gap left by the flipped out cytosine base, β strand 2 (Arg240–Tyr242) makes important base and phosphate contacts, and Thr250–Phe259, as part of the long loop 2, makes further backbone and base contacts.

The agreement in length and composition of strand 1 is good between *HhaI* and all of the type I TRDs. Loop 1 is generally predicted to be shorter and β strand 2 longer than equivalent structures in *HhaI*. However, in the prediction for *EcoKI*-1 for example, the three amino acid long strand 2 is preceded by an extra predicted strand which may suggest that the extra length of strand 2 in many of our predictions is due to a tendency for PHD to overpredict the length of a strand or to merge two strands together. In *HhaI*, strand 2 commences with Arg240 and it is apparent that an equivalent basic amino acid, e.g. Lys92 in *EcoKI*, is present in many of the type I TRDs, though usually in the middle of the longer predicted strand 2. Arg240 in *HhaI* is involved in base recognition and perhaps suggests a similar role for these basic residues in type I S subunits. Loop 2 is 21 amino acids long in the *HhaI*–DNA cocrystal structure but in the

absence of DNA, the loop is interrupted by a β strand at amino acids 250–253 (5,53). The existence of an equivalent extra β strand in the middle of loop 2 is predicted for many of the type I TRDs. In *HhaI* mtase, loop 2 terminates with another β strand, however, many of our predictions suggest that in type I TRDs, loop 2 is followed by an α helix. This may suggest that structure of type I TRDs deviates from that of *HhaI* at this junction. We propose that our alignment indicates that the TRDs of type I S subunits contain a DNA sequence recognition region consisting of β strand 2 and loops 1 and 2 with the same tertiary structure as part of the TRD of *HhaI* C5-cytosine mtase.

The only other type II mtase structure cocrystallised with DNA is of the *HaeIII* C5-cytosine mtase. The *HaeIII* TRD is slightly less ordered than that of *HhaI* but the overall fold of the polypeptide backbone in the DNA recognition region is the same (6). Although all biochemically characterised type I R/M systems methylate the N6 position of adenine and the *HhaI* and *HaeIII* mtases methylate the C5 position of cytosine, there is no reason why they cannot use the same protein structure to recognise their DNA target since it has been shown for a number of mtases that

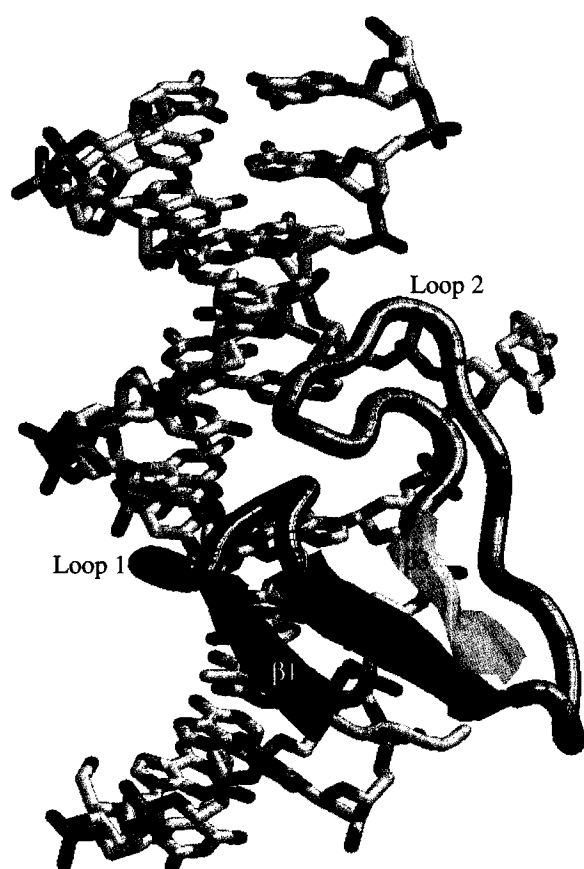


Figure 2. Part of the TRD of *HhaI* mtase bound to its DNA target with loops 1 and 2 highlighted in magenta and cyan respectively (5). β strands 1, 2 and 3 are shown in red, green and yellow respectively. The extrahelical cytosine base projects out of the minor groove into the catalytic site within the methylation domain. This methylation domain is present in the M subunits of type I R/M enzymes (28).

the nucleotide which is the target for methylation is not a major determinant of sequence specificity (54,55) and that sequence recognition can tolerate unusual base pairs (56). Our results make the experimentally testable prediction that amino acids in a well-defined and experimentally amenable region of the TRDs of type I S subunits are important for sequence recognition.

DISCUSSION

Combining the methods of multiple sequence alignment and secondary structure prediction within the *sss_align* program has facilitated the alignment of all 51 known or putative type I TRDs, overcoming the difficulties imposed by the large size of the TRDs and their very limited sequence conservation. The alignment bears a significant similarity to a short section responsible for DNA sequence specificity in the *HhaI* mtase. We suggest that this implies that all TRDs of type I S subunits are the products of divergent evolution with a conserved tertiary structure and that a small part of this structure, by analogy with *HhaI* mtase, is involved in DNA sequence recognition.

A variety of experiments such as UV-induced crosslinking to DNA (57), chemical modification of lysines (58) and random mutagenesis of TRDs (personal communication, M. O'Neill and N. E. Murray) have been applied to the best characterised type I R/M systems, *EcoKI* and *EcoR124I*, to identify amino acids involved in sequence recognition. These experiments provide preliminary support for our identification of a DNA binding region.

Chemical modification of *EcoR124I* showed that several lysines in the second TRD were susceptible to modification especially in the absence of bound DNA (58). Lysines 261, 297 and 327 within the TRD were particularly strongly modified. Lys297 is the most strongly modified residue and lies within the second proposed recognition loop. These three lysine residues are also conserved in the first TRD of *StySKI* which recognises the same DNA target as the second TRD of *EcoR124I* therefore supporting a role for them in sequence recognition (39). The other less strongly modified lysines in the second TRD may be required for non-specific DNA binding as they are not conserved in *StySKI* and lie outside of our predicted recognition region.

Random mutagenesis of the first TRD of *EcoKI* has so far changed 62 out of 150 amino acids (personal communication, M. O'Neill and N. E. Murray). Most of the mutations are silent, but five of seven mutations that impair restriction and modification are within the two putative recognition loops. The other two mutations occur shortly after the position of β 3 in Figure 1.

UV-crosslinking demonstrated that Tyr27 in the first TRD of *EcoKI* was in contact in the major groove with the 3' thymine base in the sequence complementary to the 5' AAC part of the *EcoKI* target (57). This residue is outside of our predicted recognition loops, however, it has been found that changing it to other amino acids has a minor effect on DNA specificity suggesting that it may be involved in a non-sequence specific interaction with the DNA (personal communication, M. O'Neill and N. E. Murray).

Genes similar to the *hsd* genes of enteric bacteria have now been found in non-enteric bacteria and archaeobacteria (see references in Table 1) indicating that type I R/M systems are widespread in nature. It has been suggested that diversity within genes such as those forming type I R/M systems would be advantageous to a bacterial population (59,60). Furthermore, the diversity in *hsd* gene sequences observed in enteric bacteria provides support for horizontal gene transfer and a very ancient origin for the *hsd* genes (19,21). The presence of type I R/M systems on conjugative plasmids would assist the spread of *hsd* genes by horizontal transfer (61). The existence of a common tertiary structure for TRDs, as implied by Figure 1, would support this model for the distribution of type I systems in nature. Gene duplication of TRDs and transfer of TRDs by recombination is evident, not only from genetic and sequencing experiments (16,18,44,52,62), but also from biochemical results on the domain structure of the S subunit (23–26). Recombination was responsible for the generation of two new type I target specificities, *StySQI* and *EcoR124/3I* (22,63,64), and evidence for recombination of a short stretch of the *hsdS* gene between *E.coli* B and *S.enterica* serovar *Potsdam* has been found (19). It is possible that other recombination events could encompass the short region within the TRD which we have predicted to be involved in DNA recognition, thereby allowing the generation of new specificities. These experiments suggest that the type I S subunit is a fusion of two half S subunits each containing one TRD to give a 2-fold rotationally symmetric arrangement of the TRDs and a bipartite DNA target (24,25,27,28,65). Horizontal gene transfer has also

been proposed for the type II R/M systems (66). The range of organisms in which type I systems have been found or postulated, and their diversity within species such as *E.coli* and *S.enterica*, could also suggest that a large pool of TRDs existed before the evolution of different bacterial species. Therefore, it may be possible to have similar TRDs in different species even without invoking horizontal transfer, if they both carried with them the same range of TRDs when the species diverged (67).

If our alignments are realistic, then the similarity between TRDs of type I N6-adenine mtases and the TRDs of C5-cytosine mtases may extend further to many, if not all, TRDs of type II N6-adenine mtases, type III mtases and other mtases which do not fit current classifications. This would support the proposal (68) that all mtases have evolved from a common ancestor consisting of a small monomeric TRD, such as that still found in *Aqui* mtase (69) and *EcoHK31I* mtase (70), associated with a separate catalytic subunit. It has been proposed that the mtase catalytic subunit may have developed from early DNA repair enzymes which use the same base flipping method to gain access to their target base as the mtases (71). The normal rate of mutation and gene duplication events coupled with the selection pressure within a bacterial population to expand the range of DNA target sequences, has virtually obscured this common origin. A conserved tertiary structure within TRDs implies that it may eventually be feasible to derive the amino acid recognition code used by TRDs to recognise DNA sequences as is currently being revealed for zinc finger-DNA recognition (3).

NOTE ADDED IN PROOF

Pasteurella haemolytica also appears to contain a type I system belonging to the ID family. See S. K. Highlander and O. Garza (1996), *Gene*, **178**, 89–96 for the gene sequences and alignment of the S subunit with that of HI1286. The TRDs of this system fit into the alignment scheme shown in Figure 1.

ACKNOWLEDGEMENTS

We wish to thank Professor Noreen Murray, Dr Andrew Coulson and our colleagues in their laboratories, particularly Dr Mary O'Neill, for provision of unpublished data and many useful discussions. We also thank Professor Thomas Trautner and Dr Guoliang Xu (Berlin), and Dr Junichi Ryu (Loma Linda) for the provision of unpublished sequences and other information. This work would not have been possible without the support of The Royal Society and The Darwin Trust. David Dryden thanks the Royal Society for a University Research Fellowship and Shane Sturrock thanks the Biochemical and Biological Sciences Research Council for a studentship.

REFERENCES

- Harrison, S. C. (1991) *Nature*, **353**, 715–719.
- Luisi, B. (1995) In Lilley D. M. J. (ed.), *DNA-protein: Structural Interactions*. Oxford University Press, pp. 1–48.
- Choo, Y. and Klug, A. (1997) *Curr. Opin. Struct. Biol.*, **7**, 117–125.
- Wilke, K., Rauhut, E., Noyer-Weidner, M., Lauster, R., Pawlwek, B., Behrens, B. and Trautner, T. A. (1988) *EMBO J.*, **7**, 2601–2609.
- Klimasauskas, S., Kumar, S., Roberts, R. and Cheng, X. (1994) *Cell*, **76**, 357–369.
- Reinisch, K. M., Chen, L., Verdine, G. L. and Lipscomb, W. N. (1995) *Cell*, **82**, 143–153.
- Cheng, X. and Blumenthal, R. M. (1996) *Curr. Biol.*, **4**, 639–645.
- Lange, C., Wild, C. and Trautner, T. A. (1996) *EMBO J.*, **15**, 1443–1450.
- Malone, T., Blumenthal, R. M. and Cheng, X. (1995) *J. Mol. Biol.*, **253**, 618–632.
- 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.
- Schluckebier, G., Labahn, J., Granzin, J., Schildkraut, I. and Saenger, W. (1995) *Gene*, **157**, 131–134.
- Bickle, T. A. and Kruger, D. H. (1993) *Microbiol. Rev.*, **57**, 434–450.
- King, G. and Murray, N. E. (1994) *Trends Microbiol.*, **2**, 465–469.
- Barcus, V. A. and Murray, N. E. (1995) In Baumberg, S., Young, J. P. W., Saunders, S. R. and Saunders E. M. H. (eds), *Population Genetics of Bacteria*. Society for General Microbiology Symposium **52**, pp. 31–58.
- Titheradge, A. J. B., Terment, D. and Murray, N. E. (1996) *Mol. Microbiol.*, **22**, 437–447.
- 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.
- Cowan, G. M., Gann, A. A. F. and Murray, N. E. (1989) *Cell*, **56**, 103–109.
- Kannan, P., Cowan, G. M., Daniel, A. S., Gann, A. A. F. and Murray, N. E. (1989) *J. Mol. Biol.*, **209**, 335–344.
- Sharp, P. M., Kelleher, J. E., Daniel, A. S., Cowan, G. M. and Murray, N. E. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 9836–9840.
- Gubler, M., Braguglia, D., Meyer, J., Piekarowicz, A. and Bickle, T. A. (1992) *EMBO J.*, **11**, 233–240.
- Murray, N. E., Daniel, A. S., Cowan, G. M. and Sharp, P. M. (1993) *Mol. Microbiol.*, **9**, 133–143.
- Price, C., Lingner, J., Bickle, T. A., Firman, K. and Glover, S. W. (1989) *J. Mol. Biol.*, **205**, 115–125.
- Abadjeva, A., Patel, J., Webb, M., Zinkevich, V. and Firman, K. (1993) *Nucleic Acids Res.*, **21**, 4435–4443.
- Meister, J., MacWilliams, M., Hubner, P., Jutte, H., Skrzypek, E., Piekarowicz, A. and Bickle, T. A. (1993) *EMBO J.*, **12**, 4585–4591.
- Cooper, L. P. and Dryden, D. T. F. (1994) *J. Mol. Biol.*, **236**, 1011–1021.
- Webb, M., Taylor, I. A., Firman, K. and Kneale, G. G. (1995) *J. Mol. Biol.*, **250**, 181–190.
- Kneale, G. G. (1994) *J. Mol. Biol.*, **243**, 1–5.
- Dryden, D. T. F., Sturrock, S. S. and Winter, M. (1995) *Nature Struct. Biol.*, **2**, 632–635.
- Kan, N. C., Lautenberger, J. A., Edgell, M. H. and Hutchison III, C. A. (1979) *J. Mol. Biol.*, **130**, 191–209.
- Lautenberger, J. A., Kan, N. C., Lackey, D., Linn, S., Edgell, M. H. and Hutchison III, C. A. (1978) *Proc. Natl. Acad. Sci. USA*, **75**, 2271–2275.
- Ravetch, J. V., Horiuchi, K. and Zinder, N. D. (1978) *Proc. Natl. Acad. Sci. USA*, **75**, 2266–2270.
- Somer, R. and Schaller, H. (1979) *Mol. Gen. Genet.*, **168**, 331–335.
- Nagaraja, V., Steiger, M., Nager, C., Hadi, S. M. and Bickle, T. A. (1985) *Nucleic Acids Res.*, **13**, 389–399.
- Nagaraja, V., Shepherd, J. C. W., Pripfl, T. and Bickle, T. A. (1985) *J. Mol. Biol.*, **182**, 579–587.
- Barcus, V. A., Titheradge, A. J. B. and Murray, N. E. (1995) *Genetics*, **140**, 1187–1197.
- Thorpe, P. H. (1995) *The DNA specificity of type I restriction and modification enzymes*. Ph. D. Thesis, University of Edinburgh.
- Kroger, M. and Hobom, G. (1984) *Nucleic Acids Res.*, **12**, 887–899.
- Suri, B., Shepherd, J. C. W. and Bickle, T. A. (1984) *EMBO J.*, **3**, 575–579.
- Thorpe, P. H., Terment, D. and Murray, N. E. (1997) *Nucleic Acids Res.*, **25**, 1694–1700.
- Price, C., Shepherd, J. C. W. and Bickle, T. A. (1987) *EMBO J.*, **6**, 1493–1497.
- Tyndall, C., Meister, J. and Bickle, T. A. (1994) *J. Mol. Biol.*, **237**, 266–274.
- Lee, N. S., Rutebuka, O., Arakawa, T., Bickle, T. A. and Ryu, J. (1997) *J. Mol. Biol.*, **272**, 1–7.
- Xu, G., Willert, J., Kapfer, W. and Trautner, T. A. (1995) *Gene*, **167**, 59.
- Dybvig, K. and Yu, H. (1994) *Mol. Microbiol.*, **12**, 547–560.
- Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., et al. (1995) *Science*, **269**, 496–512.

- 46 Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., et al. (1996) *Science*, **273**, 1058–1073.
- 47 Sturrock, S. S. (1997) *Improved tools for protein tertiary structure prediction*. Ph. D. Thesis, University of Edinburgh.
- 48 Smith, T. F. and Waterman M. S. (1981) *J. Mol. Biol.*, **147**, 195–197.
- 49 Dayhoff, M. O., Schwartz, R. M. and Orcutt, B. C. (1978) *Atlas of Protein Sequence and Structure*, vol. 5, suppl. 3, pp. 345–352. National Biomedical Research Foundation, Washington DC.
- 50 Rost, B. and Sander, C. (1993) *J. Mol. Biol.*, **232**, 584–599.
- 51 Rost, B., Sander, C. and Schneider, R. (1994) *CABIOS*, **10**, 53–60.
- 52 Argos, P. (1985) *EMBO J.*, **4**, 1351–1355.
- 53 Cheng, X., Kumar, S., Posfai, J., Pflugrath, J. W. and Roberts, R. (1993) *Cell*, **74**, 299–307.
- 54 Klimasauskas, S. and Roberts, R. (1995) *Nucleic Acids Res.*, **23**, 1388–1395.
- 55 Yang, A. S., Shen, J.-C., Zingg, J.-M., Mi, S. and Jones, P. A. (1995) *Nucleic Acids Res.*, **23**, 1380–1387.
- 56 Smith, S. S., Kan, J. L. C., Baker, D. J., Kaplan, B. E. and Dembek, P. (1991) *J. Mol. Biol.*, **217**, 39–51.
- 57 Chen, A., Powell, L. M., Dryden, D. T. F., Murray, N. E. and Brown, T. (1995) *Nucleic Acids Res.*, **23**, 1177–1183.
- 58 Taylor, I. A., Webb, M. and Kneale, G. G. (1996) *J. Mol. Biol.*, **258**, 62–73.
- 59 Levin, B. R. (1988) *Phil. Trans. Roy. Soc. London, Ser. B*, **319**, 459–472.
- 60 Korona, R. and Levin, B. R. (1993) *Evolution*, **47**, 556–575.
- 61 Tyndall, C., Lehnerr, H., Sandmeier, U., Kulik, E. and Bickle, T. A. (1997) *Mol. Microbiol.*, **23**, 729–736.
- 62 Gough, J. A. and Murray, N. E. (1983) *J. Mol. Biol.*, **166**, 1–19.
- 63 Fuller-Pace, F. V., Bullas, L. R., Delius, H. and Murray, N. E. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 6095–6099.
- 64 Fuller-Pace, F. V. and Murray, N. E. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 9368–9372.
- 65 MacWilliams, M. P. and Bickle, T. A. (1996) *EMBO J.*, **17**, 4775–4783.
- 66 Jeltsch, A. and Pingoud, A. (1996) *J. Mol. Evol.*, **42**, 91–96.
- 67 Maynard-Smith, J., Smith, N. H., O'Rourke, M. and Spratt, B. G. (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 4384–4388.
- 68 Wilson, G. G. and Murray, N. E. (1991) *Annu. Rev. Genet.*, **25**, 585–627.
- 69 Karreman, C. and de Waard, A. (1990) *J. Bacteriol.*, **172**, 266–272.
- 70 Lee, K.-F., Kam, K.-M. and Shaw, P.-C. (1995) *Nucleic Acids Res.*, **23**, 103–108.
- 71 Roberts, R. J. (1995) *Cell*, **82**, 9–12.