The nucleotide sequence recognised by the *Escherichia coli* D type I restriction and modification enzyme

Valakunja Nagaraja, Martin Stieger, Christoph Nager, Sheikh M.Hadi and Thomas A.Bickle*

Microbiology Department, Biozentrum, University of Basel, Klingelbergstrasse 70, CH-4056 Basel, Switzerland

ABSTRACT

A type I restriction endonuclease from a new isolate of Escherichia coli (E. coli E166) has been purified and characterised. The enzyme, EcoD, has a recognition sequence similar in overall structure to the previously determined type I enzyme sequences, an exception being that it is degenerate. The sequence is

5'-T-T-A-N-N-N-N-N-N-N-G-T-C-Y-3'
3'-A-A-T-N-N-N-N-N-N-N-C-A-G-R-5'
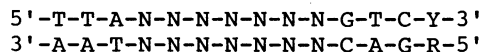
where Y is a pyrimidine, R is a purine and N can be any nucleotide. The enzyme methylates adenosyl residues in both strands of the DNA that are separated by ten base pairs, suggesting that the enzyme interacts with DNA along one face of the helix making contacts in two successive major grooves.

INTRODUCTION

Type I restriction enzymes are complicated, multifunctional proteins with a complex mechanism of action (1,2). The enzymes interact with specific DNA sequences to methylate them or, under different reaction conditions, they are stimulated by the interaction to cleave the DNA randomly at distances that can be several thousand base pairs (bp) from the interaction site (3). The type I restriction enzymes EcoK and EcoB from the E. coli strains K12 and B are by far the best studied examples (reviewed in refs. 1 and 2); however, type I restriction modification (R-M) systems have been discovered in several other strains of E. coli and species of Salmonella. All of the enzymes contain three subunits coded by three genes organised into two transcriptional units. The genes are called hsdR, hsdM and hsdS coding for the restriction, modification and sequence recogni-

tion subunits of the enzyme (4-6).

There is strong evidence that most of the type I R-M systems found in different bacteria are genetically related to each other (reviewed in 1). Further, DNA hybridisation experiments using probes derived from the E. coli K12 hsd region have suggested a close similarity between the hsdR and hsdM genes of E. coli K12 and those of several other systems, confirming their allelic nature. Interestingly enough, these experiments showed that there was no such strong homology between the different hsdS genes (7).

A new type I R-M system, EcoD, from a clinical isolate of E. coli, was described recently and shown to be allelic to the other enzymes from E. coli (7). More recently, the hsdS gene sequence of this system has been determined and compared to that of the EcoK and EcoB hsdS genes (8). These sequences are between 1,300 and 1,450 bp long and are homologous in only two short regions, one about 100 and the other 250 bp long; these short stretches are highly conserved in all three genes, no homologies can be found in the rest of the sequences. Thus these related restriction systems, though very similar with respect to their hsdM and hsdR genes, have divergent hsdS gene sequences, enabling them to recognise different DNA sequences. We have purified the EcoD enzyme and report here the experiments that we carried out to elucidate the DNA recognition sequence.

MATERIALS AND METHODS
Bacterial strains, phages and plasmids

E. coli strain E166 which produces EcoD was kindly supplied by N.E. Murray. RSF2124, an E. coli strain carrying a colE1 plasmid with a Tn3 insertion was a gift from D. Sherratt. The cells were grown in a yeast extract-tryptone medium. Phage T7 was from F. Studier and was the strain used for the genomic sequence (9). λcI857S7 was used to prepare EcoD modified and nonmodified DNA.

Enzymes and DNA

All type II restriction enzymes were purchased from New
England Biolabs or Boehringer. Phage λ and T7 DNAs were pre-
pared by phenol extraction of phage particles. The plasmid
p4014 was a gift from C. Sengstag. Other plasmid DNAs were
purified from cleared lysates by CsCl-ethidium bromide equi-
librium centrifugation (10).

Purification of EcoD

The enzyme was purified by a procedure that we recently
developed for other type I enzymes and which is described in
detail elsewhere (11). In short, cell free extracts were made
free from nucleic acids by polyethylenimine precipitation
followed by ammonium sulphate fractionation of the supernatant.
The enzyme was purified by a series of column chromatography
steps on DEAE-Sephacel, gel filtration on Sephacryl S-200,
heparin agarose and lastly on ATP-agarose. The purification
was monitored by the DNA dependent, S-adenosylmethionine
(AdoMet) dependent ATPase activity of the enzyme as described
in (12) as well as by immunoblotting (13) using antibodies
raised against the hsdR and hsdM subunits of EcoK as previously
described (7,11).

In vitro DNA methylation

Methylation of DNA by EcoD was carried out using ^{3}H-methyl-
AdoMet (Amersham, 64-74 Ci/mmol) as described in (14) with some
modifications (11).

Other methods

Electrophoresis of proteins on sodium dodecyl sulphate
polyacrylamide gels was according to (15). DNA fragments were
separated on either 1% agarose or 4% polyacrylamide gels de-
pending on the expected fragment size. DNA in the gels was de-
tected by staining with ethidium bromide and ^{3}H-DNA bands by
fluorography (16).

RESULTS

Properties of EcoD

The enzyme EcoD resembles EcoB and EcoK in its subunit

structure, functions and cofactor requirements. It exhibits non-modified DNA and AdoMet dependent ATPase activity, DNA cleavage and modification methylation activities. Both ATP and AdoMet are required for restriction which is followed by massive ATP hydrolysis.

Determination of the EcoD DNA recognition sequence

Until recently, the recognition sequences for only EcoK (17), EcoB (18-21) and EcoA (14) were known. In all these cases, the recognition sequence is split into two specific domains separated by a spacer of non specific sequence. The EcoK and EcoB recognition sequences were elucidated by sequencing mutated sites. For EcoA we developed a computer aided strategy that is described in detail elsewhere (11,14). In brief, DNA of known sequence is methylated by the modification methylase activity of the enzyme using ^{3}H-methyl-AdoMet as the methyl donor. The enzyme methylates specific adenosyl residues within the recognition sequence and each labelled site is precisely mapped to a small region in the DNA by a combination of digestion with various type II restriction enzymes, gel electrophoresis and fluorography to locate the labelled DNA bands. This mapping data is then fed to a computer which is instructed to search for sequences common to all of the labelled regions but absent from the rest of the DNA. The putative sequences thus obtained are then tested by predicting where the enzyme ought to methylate other sequenced DNA molecules.

An example of one such analysis for the enzyme EcoD is shown in Figure 1. In this case, SV40 DNA (22) was methylated with EcoD, digested with various type II restriction enzymes and analysed on a polyacrylamide gel. EcoD has only one site in SV40 DNA since only one band is labelled in each of the enzyme digests. From the data shown in Figure 1, the site could be mapped to within a 152 bp region; further analysis (not shown) with the enzymes FokI and Fnu4HI allowed it to be narrowed down to an 80 bp region (from the DdeI site at 2293 to the FokI site at 2373). Since this information alone was not sufficient for the computer analysis, we tested a number of
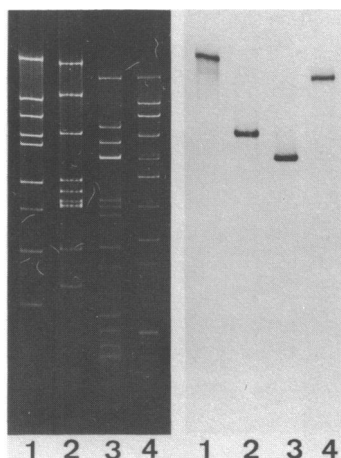
Figure 1. Mapping of the EcoD site in SV40 DNA. SV40 DNA was methylated in vitro with EcoD and ^{3}H-methyl AdoMet (Amersham, 64-72 Ci/mmol) as described (11). Labelled DNA was then cleaved with different type II restriction enzymes and the fragments were separated on a 4% polyacrylamide gel which was then stained with ethidium bromide and photographed, dried and fluorographed. The left panel shows the stained gel and the right panel the fluorogram. Lanes 1, RsaI digestion, a fragment extending from 1385-2989 is labelled; lanes 2, HaeIII digestion, label is in the fragment extending from 2177-2716; lanes 3, DdeI digestion, the coordinates of the labelled fragment are 2293-2710; lanes 4 show a BstNI digestion and a fragment extending from 1453-2445 is labelled.

other sequenced DNA molecules in our collection to see whether they were substrates for EcoD. The following DNAs did not contain sites for the enzyme: pBR322, pBR325, pSHI44, p3201, M13, G4, ØX174, Cauliflower Mosaic Virus, yeast 2μ DNA, IS1, IS2, IS3, IS4 and IS10. These sequences together with SV40 were extracted from the EMBL DNA sequence data bank and the 80 bp region of SV40 that contains an EcoD site was screened against the rest of the DNA sequences as described above. The analysis produced three sequences as candidates for the EcoD site: 5'-TTA(N)$_7$GTCT-3', 5'-CAC(N)$_6$TGTC-3' and 5'-GAA(N)$_7$TGTG-3'. In other words, these were the only three sequences that could be found within the 80 bp of SV40 DNA that were absent from the other sequenced DNAs, a total of about 50,000 bp.
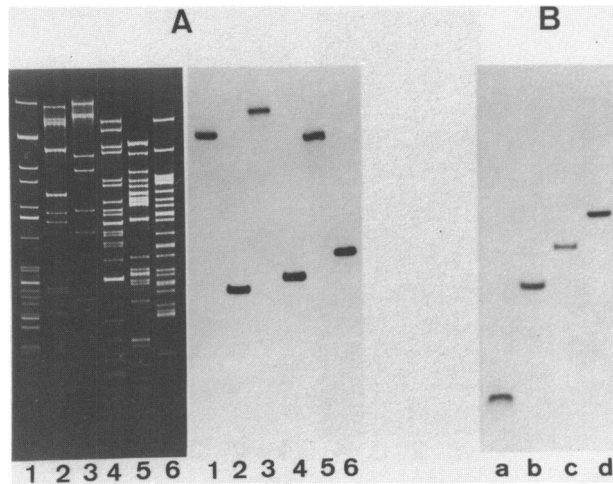
Figure 2. Mapping the EcoD site in plasmid p4014 DNA. The DNA
was labelled and analysed as described in the Legend to Figure 1.
Panel A shows a 4% polyacrylamide gel (left) with the corres-
ponding fluorogram (right). The enzymes used and the extent of
the labelled fragments are as follows: Lane 1, MspI 9548-456;
lane 2, TaqI 9379-9574; lane 3, RsaI 9031-511; lane 4, HinfI
9494-9706; lane 5, HaeIII 9152-14; lane 6, DdeI 9465-9723.
Panel B shows the fluorogram of an 8% polyacrylamide gel. Lane
a, MspI and TaqI 9548-9574; lane b, MspI and Fnu4HI 9548-9605;
lane c, MnlI 9536-9612 and lane d, Sau3AI 9481-9570.

Among the DNA molecules that proved to contain an EcoD
site was the plasmid p4014. This plasmid has the BamHI-EcoRI
fragment of λ DNA extending from 39168 to 34499 (23) cloned
into EcoRI and BamHI cleaved pBR322. The plasmid also contains
an IS2 element within the λ sequences (C. Sengstag, unpublished
information). The mapping of this site is shown in Figure 2; it
could be localised to within 26 base pairs. These 26 base pairs,
however did not contain any of the three candidate EcoD sequen-
ces found in SV40. This result led us to explore the possibili-
ty that the EcoD recognition sequence might be degenerate at
one or more positions. We took the three candidate sequences
from SV40 and instigated a computer search in which each posi-
tion within the sequences was systematically allowed to "wobble".
This analysis brought out a single sequence that was compatible
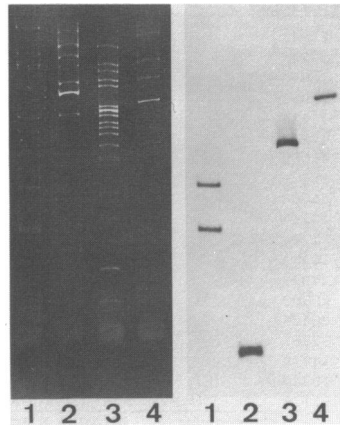with all of the data: 5'-TTA(N)$_7$GTCY-3' where Y can be either

Figure 3. Confirmation of the EcoD recognition sequence. The plasmid pRSF 2124 was methylated with EcoD on a 4% polyacryl-amide gel as described in the Legend to Figure 1. The left of the Figure shows the stained gel and the right the fluorogram. The type two restriction enzymes used were in lane 1, FnuDIII, lane 2 TaqI, lane 3 DdeI and lane 4 HaeII.

pyrimidine. In SV40, Y is thymidine whereas in p4014 it is cytosine.

To test this sequence we did another computer search and found that the transposon Tn3 (24,25) contains the sequence at position 3224 (the version with T in the wobble position). Moreover, this sequence contains a site for the type II restriction enzyme FnuDII within the non-specific spacer. We predicted therefore that when a plasmid containing Tn3 is methylated with EcoD and then digested with FnuDII, the radioactive methyl groups should be split between two different DNA fragments. Figure 3 demonstrates that digestion of a Tn3 containing plasmid with FnuDII results in two labelled bands while digestion with other enzymes gives only one at the predicted position.

Phage λ DNA (23) was predicted to contain the sequence at five positions, two of them with T and three of them with C in the wobble position. Experimental analysis of the computer prediction showed that λ DNA contains five EcoD sites which are found on the predicted restriction fragments. Similar experi-

DNA	position		sequence	
SV40	2294	C	TTA GAAAGCG GTCT	G
Tn3	3224	G	TTA TGCGCGG GTCT	C
λ	26850	T	TTA TCTTACT GTCT	T
	29960	G	TTA CTAACGG GTCT	T
	38734	A	TTA CGCTCCT GTCC	G
	47222	C	TTA TAGAAAC GTCC	T
	48069	T	TTA CAATATC GTCC	T
T7	556	C	TTA TCAGGCT GTCT	A
	6500	T	TTA AAGCCGT GTCT	T
	10147	A	TTA CGTCTTC GTCT	G
	12774	C	TTA GCGAGCA GTCT	A
	15353	T	TTA ACCCTTC GTCT	C
	24935	G	TTA ACGGTTG GTCT	T
	26809	G	TTA CCGCTGA GTCT	C
	28304	A	TTA CCTCCCC GTCT	T
	6776	T	TTA ACACAGG GTCC	G
	14871	G	TTA CCACAAC GTCC	T
	30930	A	TTA CCCAAGC GTCC	C
	35176	T	TTA AGAACGA GTCC	A
	Consensus:		TTANNNNNNNGTCY	

Figure 4. EcoD recognition sites. The regions of DNA containing
the 19 EcoD sites discussed in the text are shown. The FnuDII
cleavage site within the Tn3 sequence is underlined.

ments with phage T7 showed the presence of 12 sites as pre-
dicted (not shown).

All of the EcoD recognition sequences that we have ana-
lysed are shown in Figure 4 together with their flanking
nucleotides. An examination of these sequences shows that the
only conserved elements are the ones indicated. All four bases
can be found at all positions of the non-specific spacer as
well as flanking the sequence. Thus, the maximum ambiguity
allowed is the pyrimidine wobble at the last specified posi-
tion.

DISCUSSION

The EcoD recognition sequence has a general structure
strikingly similar to the other type I sequences described so
far. It has a trinucleotide and a tetranucleotide domain se-
parated by a 7 bp spacer of non-specific sequence. The EcoB
and EcoK sequences have the same overall structure, the only
difference being in the length of the spacer, 6 bp for EcoK
and 8 bp for EcoB (Fig. 5). Comparing the specific parts of
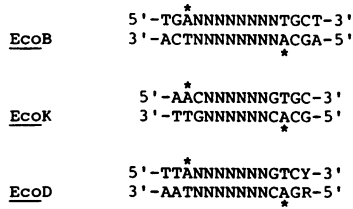these three sequences, we find no elements common to all

```
                                                *
                          5'-TGANNNNNNNNNTGCT-3'
           EcoB           3'-ACTNNNNNNNNACGA-5'
                                              *


                                            *
                          5'-AACNNNNNNNGTGC-3'
           EcoK           3'-TTGNNNNNNCACG-5'
                                           *


                                           *
                          5'-TTANNNNNNNNGTCY-3'
           EcoD           3'-AATNNNNNNNCAGR-5'
                                          *
```

Figure 5. The recognition sequences of EcoK, EcoB and EcoD.
The methylated adenosyl residues of the modified sites are in-
dicated with asterisks.

except that the sites of methylation and the distance between

the methylated bases is the same in all cases. For EcoK, it

has been shown that the second A in the top strand and the

only A in the lower strand of the sequence as written in

Figure 5 are those that are methylated in the modification

reaction (1) and the distance between the methylated bases is

thus ten base pairs. The methylated bases have also been de-

termined for EcoB (26) and they are also ten base pairs apart.

For EcoD, the results shown in Figure 3 demonstrate that the

enzyme methylates both domains of the sequence and this ex-

cludes that the methylation could be on cytosine residues.

The results presented in Table 1 show that EcoD methylates both

Table 1.   EcoD methylates adenines in both strands of DNA

Nucleotide added	% ^{3}H released
1. none	79.8
2. 100 mM dGTP	15.0
3. 100 mM dCTP	18.8
4. 100 mM dTTP	52.6

Tn3 DNA was methylated with EcoD in vitro and then cleaved with
FnuDII, which cuts in the non-specific part of the EcoD site
(Fig. 4). The DNA was then treated with T4 DNA polymerase. In
the absence of dNTPs the exonuclease activity of the enzyme
removes the methylated nucleotides (line 1). The presence of
dGTP or dCTP (lines 2 and 3) protects the methylated nucleo-
tides from exonuclease action while the addition of dTTP pro-
tects one half of the radioactivity from exonuclease action.
Inspection of the sequence (Fig. 4) shows that dTTP must have
protected the top strand of the trinucleotide domain which con-
tains only one methylatable residue, an adenosine.

strands of the DNA and that the methylation is, as predicted, on adenosyl residues. These are, the only invariant adenosine in the tetranucleotide domain and the adenosine in the top strand of trinucleotide domain as drawn in Figure 5. The methyl-ated residues are again ten base pairs apart. We have recently determined the recognition sequences of several other type I enzymes (11) and the same ten bp spacing is found for all of them. These findings immediately suggest the way in which the enzyme binds to its recognition sequence in DNA. The protein must bind along one side of the DNA helix, interacting with the specific parts of the sequence in two successive major grooves (the methylation site is in the major groove) and with the non-specific part of the sequence in the intervening minor groove.

The EcoD recognition sequence differs from those of EcoK or EcoB in that it is degenerate at one position. This degener-acy may be advantageous to strains carrying the enzyme in that satistically one expects to find more EcoD sites than those for EcoK or EcoB.

*To whom correspondence should be addressed

REFERENCES
1. Bickle, T.A. (1982) in The Nucleases, Linn, S.M. and
    Roberts, R.J. Eds. pp. 85-108, Cold Spring Harbor Press,
    New York.
2. Yuan, R. (1981) Annu. Rev. Biochem. 50, 285-315.
3. Bickle, T.A., Brack, C. and Yuan, R. (1978) Proc. Natl.
    Acad. Sci. U.S.A. 75, 3099-3103.
4. Boyer, H.W. and Roulland-Dussoix, D. (1969) J. Mol. Biol.
    41, 459-472.

5. Glover, S.W. and Colson, C. (1969) Genet. Res. 13, 227-240.
6. Hubacek, J. and Glover, S.W. (1970) J. Mol. Biol. 50, 111-127.
7. Murray, N.E., Gough, J.A., Suri, B. and Bickle, T.A. (1982) EMBO J. 1, 535-539.
8. Gough, J.A. and Murray, N.E. (1983) J. Mol. Biol. 166, 1-19.
9. Dunn, J.J. and Studier, F.W. (1983) J. Mol. Biol. 166, 477-535.
10. Clewell, D. and Helinski, D.R. (1969) Proc. Natl. Acad. Sci. U.S.A. 62, 1159-1166.
11. Nagaraja, V., Shepherd, J.C.W., Pripfl, T. and Bickle, T.A. J. Mol. Biol., in press.
12. Hadi, S.M., Bickle, T.A. and Yuan, R. (1975) J. Biol. Chem. 250, 4159-4164.
13. Towbin, H., Staehelin, T. and Gordon, J. (1979) Proc. Natl. Acad. Sci. U.S.A. 76, 4350-4353.
14. Suri, B., Shepherd, J.C.W. and Bickle, T.A. (1984) EMBO J. 3, 575-579.
15. Laemmli, U.K. (1970) Nature (Lond.) 227, 680-685.
16. Chamberlain, J.P. (1979) Anal. Biochem. 98, 132-135.
17. Kan, N.C., Lautenberger, J.A., Edgell, M.H. and Hutchison, C.A. III (1979) J. Mol. Biol. 130, 191-209.
18. Ravetch, J.V., Horiuchi, K. and Zinder, N.D. (1978) Proc. Natl. Acad. Sci. U.S.A. 75, 2266-2270.
19. Lautenberger, J.A., Kan, N.C., Lackey, D., Linn, S., Edgell, M.H. and Hutchison, C.A. III (1978) Proc. Natl. Acad. Sci. U.S.A. 75, 2271-2275.
20. Lautenberger, J.A., Edgell, M.H., Hutchison, C.A. III and Godson, G.N. (1979) J. Mol. Biol. 131, 871-875.
21. Sommer, R. and Schaller, H. (1979) Mol. Gen. Genet. 168, 331-335.
22. Reddy, V.B., Thimmappaya, B., Dhar, R., Subramanian, K., Zain, B.S., Pan, J., Ghosh, P.K., Celma, M.L. and Weissman, S.M. (1978) Science 200, 494-502.
23. Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F. and Petersen, G.B. (1982) J. Mol. Biol. 162, 729-773.
24. Heffron, F., McCarthy, B.J., Ohtsubo, H. and Ohtsubo, E. (1979) Cell 18, 1153-1163.
25. Sutcliffe, J.G. (1978) Proc. Natl. Acad. Sci. U.S.A. 75, 3737-3741.
26. Van Ormondt, H., Lautenberger, J.A., Linn, S. and de Waard, A. (1973) FEBS Lett. 33, 177-180.