THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 256, No. 5, Issue of March 10, pp. 2143-2153, 1981
*Printed in U.S.A.*

# Sequence Analysis of the DNA Encoding the *Eco* RI Endonuclease and Methylase*

(Received for publication, October 22, 1980)

**Patricia J. Greene‡, Madhu Gupta, and Herbert W. Boyer**

*From the Howard Hughes Medical Institute and the Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, California 94143*

**William E. Brown**

*From the Department of Biological Sciences, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213*

**John M. Rosenberg§**

*From the Department of Biological Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania 15260*

**The *Eco* RI endonuclease and methylase recognize the same hexanucleotide substrate sequence. We have determined the sequence of a fragment of DNA which encodes these enzymes using the chain-termination method of Sanger (Sanger, F., Nicklen, S., and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U. S. A.* 74, 5463-5467). The amino acid sequences of both enzymes were derived from the DNA sequence. The coding regions selected include the only open translational frames of sufficient length to accommodate the enzymes. They coincide with previously established gene boundaries and orientation. The predicted amino acid sequences correlate well with analyses of the purified protein. Comparison of the nucleotide and protein sequences reveals no homology between the endonuclease and methylase which might provide insight into the origin of the restriction-modification system or the mechanism of common substrate recognition. Based on secondary structure predictions, the two enzymes also have grossly different molecular architecture. The base composition of the sequence is 65% A + T, and the codon usage is significantly different from that observed in several *Escherichia coli* chromosomal genes. In some cases, frequently selected codons are recognized by minor tRNA species. A spontaneous mutation in the endonuclease gene was isolated. Serine replaces arginine at residue 187. In crude extracts, *Eco* RI specific cleavage is ~0.3% wild type.**

The *Eco* RI restriction endonuclease and modification methylase, together with their DNA substrate, provide a model system for probing the molecular mechanisms of sequence-specific DNA-protein interactions. The two enzymes recognize the same nucleotide sequence d(−GAATTC−) (1, 2) but differ in several aspects of interaction with the substrate (see Ref. 3 for a recent review). The endonuclease and methylase have been purified to homogeneity and x-ray diffraction analysis of the crystalline endonuclease is underway (4, 5). We report here the sequence of the DNA which encodes the *Eco* RI endonuclease and methylase and the amino acid sequences of both enzymes deduced from the DNA sequence.

The source of DNA for this analysis is the plasmid, pMB1, and derivatives constructed by recombination *in vitro*. pMB1 was obtained from the original clinical strain of *Escherichia coli* found to have the *Eco* RI host specificity. It is closely related to ColE1. Both plasmids encode colicin immunity and production and have identical replication properties. pMB1 is larger by approximately two kilobases, the amount required to accommodate the *Eco* RI genes. Examination of heteroduplexes of pMB1 with ColE1 by electron microscopy revealed no mismatched regions except the *Eco* RI genes (6). The approximate boundaries of the *Eco* RI genes have been determined by subcloning restriction fragments of pMB1, and the direction of transcription has been shown to be endonuclease to methylase (4) (Fig. 1).

The 2234-base pair DNA sequence presented here was determined by the dideoxyribonucleotide chain-termination method of Sanger (7). Single-stranded template DNA was obtained by cloning the *Eco* RI genes in M13mp5, a single strand bacteriophage vector developed by Messing and his coworkers (8, 9).

In the accompanying paper, Newman *et al.* (10) report a sequence analysis of the *Eco* RI genes contained in pMB4, a derivative of pMB1 which determines ampicillin resistance but not colicin production. These two plasmids have been maintained as separate laboratory lines for approximately 10 years (6, 11). The DNA sequence in the accompanying paper was determined by the method of Maxam and Gilbert (12). The two DNA sequences, obtained in different laboratories using different plasmids as sources of DNA and different methods of DNA sequencing, are identical in both the coding and noncoding regions. We are, therefore, confident that the nucleotide sequence and the predicted amino acid sequences are correct.

## MATERIALS AND METHODS[1]

Figs. 1 and 2 depict the plasmids and phage which carry the *Eco* RI genes from pMB1.

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ Recipient of Grant GM 25729-03 from the National Institutes of Health.

§ Recipient of Grant GM 25671 from the National Institutes of Health.

[1] Portions of this paper (including "Materials and Methods," Figs. 6 to 8, Tables IV and V, and additional references) are presented in miniprint at the end of this paper. Miniprint is easily read with the aid of a standard magnifying glass. Full size photocopies are available from the Journal of Biological Chemistry, 9650 Rockville Pike, Bethesda, Md. 20014. Request Document No. 80M-2241, cite author(s), and include a check or money order for $7.60 per set of photocopies. Full size photocopies are also included in the microfilm edition of the Journal that is available from Waverly Press.

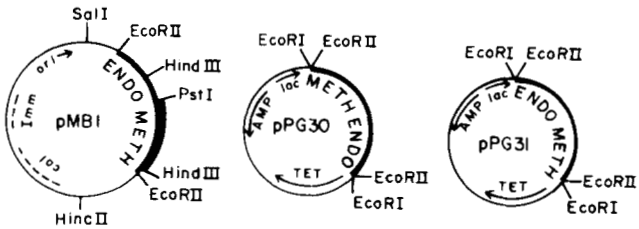


FIG. 1. ***Eco* RI-containing plasmids.** The approximate boundaries of the *Eco* RI genes were determined by subcloning restriction fragments of pMB1 (4, 6, 13). The 2300-base pair *Eco* RII fragment indicated by the *inner heavy line* contains all of the information necessary for normal expression of both genes. The 1280-base pair *Pst*I-*Hin*dIII fragment indicated by the *outer heavy line* contains the methylase gene. pPG30 and pPG31 contain the 2300-base pair *Eco* RII fragment inserted into the *Eco* RI site of pBH20 (a pBR322 derivative) (14) in opposite orientations relative to the *lac* promoter. The *arrows* indicate the direction of transcription from the *lac*, tet, and amp promoters. pPG30 and pPG31 were constructed by filling in the *Eco* RI ends of the vehicle and the*Eco* RII ends of the fragment using T4 DNA polymerase. The resulting blunt-ended fragments were joined using T4 DNA ligase (4). The direction of transcription of the *Eco* RI genes was determined by measuring the effect of the *lac* promoter on the specific activity of the *Eco* RI enzymes in strains containing these plasmids. In strains containing pPG31, the endonuclease and methylase levels are both elevated by growth in glycerol and by the addition of isopropyl-1-thio-β-D-galactopyranoside. Enzyme levels are not affected by these growth conditions in pPG30 or pMB1. *ENDO*, endonuclease; *METH*, methylase.

SEQUENCING STRATEGY

M13-RI recombinant phage were used to prepare template DNA. A *Hin*dIII fragment which spans the *Eco* RI genes was isolated from a partial digest of pPG30 and inserted into the *Hin*dIII site of M13mp5 (Fig. 2*A*). Isolates with each of the orientations of the fragment relative to the phage DNA were obtained to provide templates of both strands of the *Eco* RI genes. Opposite orientations were identified by assessing the ability of phage DNA from two M13-RI isolates to form double-stranded DNA in the region of interest. M13-RI recombinant phage are unstable, eliminating at high frequency a segment of DNA approximately the same size as the insert encoding the *Eco* RI enzymes. In spite of this problem, we were able to obtain high quality template for sequence analysis. The cloning experiments, analysis of the recombinant phage, and template preparation are described in the miniprint.

Primers were isolated from restriction endonuclease digests of pMB1 and pPG30 DNA as described in the miniprint. A synthetic primer, 5′ p-C-C-C-A-G-T-C-A-C-G-A-C-G-T-T-OH 3′, which hybridizes to M13mp5 adjacent to the cloning site, was a gift from Dr. Roberto Crea (Genentech, Inc.). The hybridization site of this sequence is shown in Fig. 2*A*. Restriction sites used to prepare primers and a summary of sequencing experiments are shown in Fig. 3.

The absolute orientation of generated sequence was deduced by using primers with different restriction sites at the ends. For example, when the smaller *Hin*dIII-*Pst* I fragment was used as a primer with one template orientation, the reaction mixtures were divided in half and digested with either *Pst* I or *Hin*dIII. Readable sequence was obtained with only one set of cleavage products. When sequence could be read following *Pst* I cleavage, the orientation, designated a, corresponded to the direction of transcription of the *Eco* RI genes. When sequence could be read following *Hin*dIII cleavage, the ori-

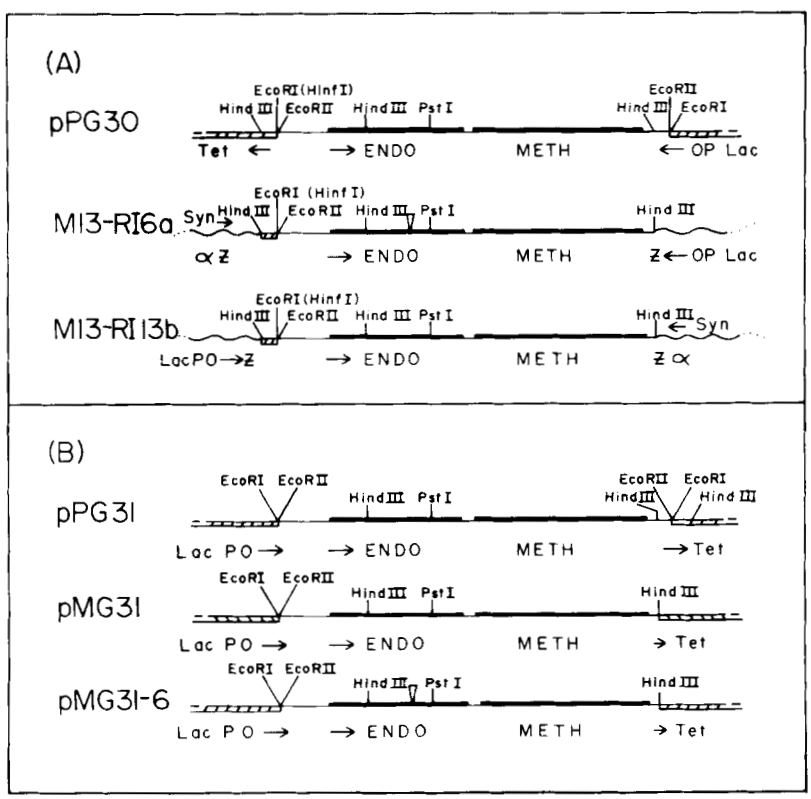


FIG. 2. **Recombinant plasmids and phage carrying the *Eco* RI genes from pMB1.** The distances between restriction sites near the end of the *Eco* RII fragment from pMB1 are exaggerated to show details of gene transfer. The endonuclease and methylase genes are indicated by *heavy lines*. *Arrows* show the direction of transcription from the *lac*, tet, and *Eco* RI promoters. DNA originating from pBH20, the vehicle used to construct pPG30 and pPG31, is indicated by *hatched lines*. *ENDO*, endonuclease; *METH*, methylase. *A*, derivation of M13-RI hybrid phage: M13mp5, designated by *wavy lines*, contains a *Hin*dIII site in the DNA encoding the *lac*Z α peptide (8, 9). The *Hin*dIII fragment from pPG30, which was inserted into this site, lacks approximately 40 nucleotides from the methylase end of the *Eco* RII fragment of pMB1 and contains 34 nucleotides from pBH20. The two orientations of the *Eco* RI genes relative to the M13 are designated a and b. *B*, derivation of pMG31 and pMG31-6: pMG31 is analogous to pPG31 except for deletion of a *Hin*dIII fragment of approximately 70 base pairs. pMG31-6 was constructed by isolating the 1685-base pair *Hin*dIII fragment from MI3-RI6a RF DNA and inserting it back into pPG31. MI3-RI6a contains a spontaneous mutation in the endonuclease gene marked (▽ in M13-RI6a and pMG31-6.

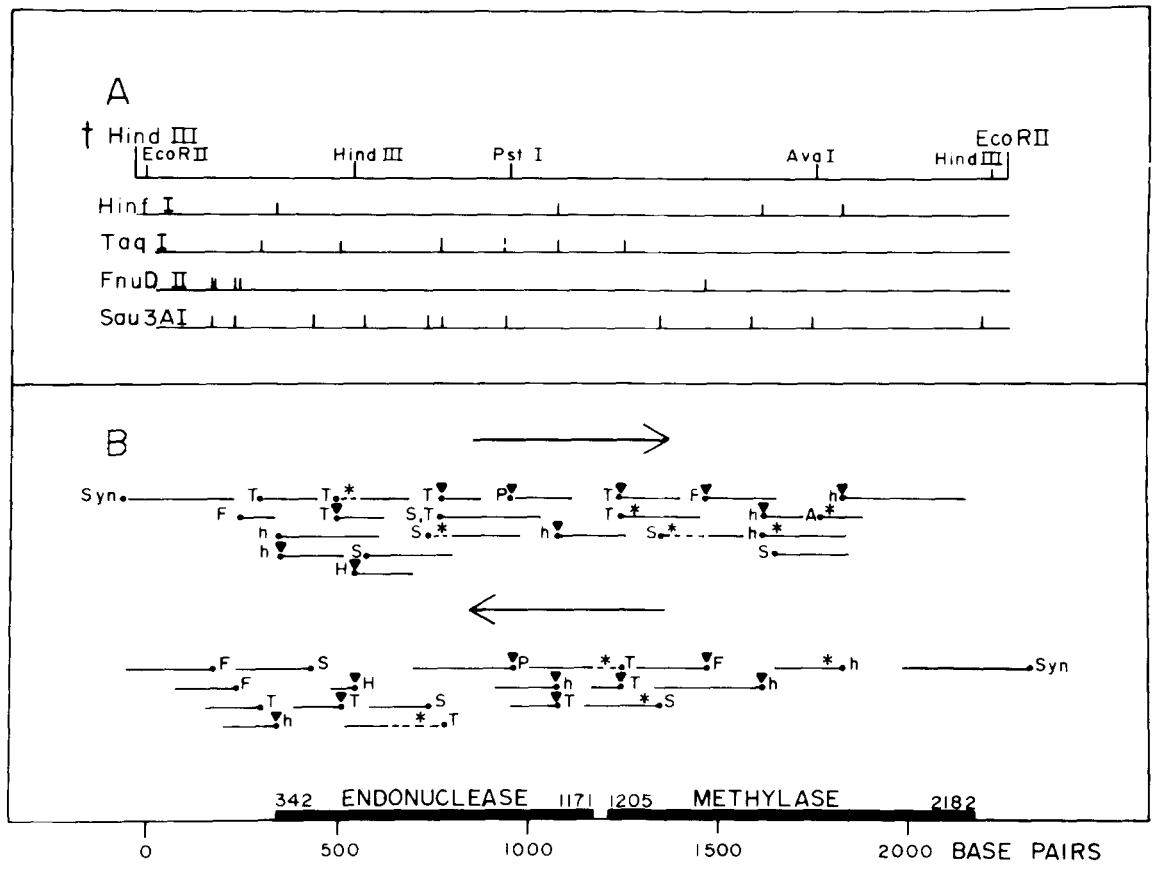


FIG. 3. **Summary of sequencing strategy.** Nucleotide numbering starts with the *Eco* RII site of pMB1. *A*, restriction sites used to make primers. The *Hin*dIII site at the *left* marked † occurs only in pPG30, and the DNA between the *Hin*dIII and *Eco* RII sites originates from pBH20 (via pPG30). pMB1 DNA was digested with *Eco* RII, and the 2300-base pair fragment carrying the *Eco* RI genes was isolated. This was further digested with the following enzymes either singly or in combination: *Hin*fI, *Taq* I, *Fnu*DII, and *Sau*3aI. pPG30 DNA was triply digested with *Hin*dIII, *Ava* I, and *Pst* I; and the four fragments which span the *Eco* RI genes were isolated. These were used directly as primers or further digested with the enzymes listed above. The *Taq* I site indicated by a *dotted line* overlaps a methylated *Mbo* I site and is not cleaved by *Taq* I (15). *B*, sequencing experiments. The start of a DNA segment is indicated by ● and a letter which designated the restriction site: *A*, *Ava* I; *F*, *Fnu*DII (*Tha* I); *H*, *Hin*dIII; *h*, *Hin*fI; *S*, *Sau*3AI (*Mbo* I); *T*, *Taq* I; *Syn*, synthetic. The length of the *solid line* corresponds to the length of sequence read. Long primers were cleaved off prior to electrophoresis. DNA segments read from cleavage points are designated ▼. Several primers were digested with Exonuclease III; * designates sequence read from these, and the *dotted line* indicates the distance from the 5′ end of the primer that readable sequence began.

entation was designated b. These assignments were further confirmed by use of the synthetic primer.

## RESULTS AND DISCUSSION

We have determined the sequence of a 2234-nucleotide DNA fragment from pMB1 by the sequencing strategy outlined. The nucleotide sequence together with the amino acid sequences of the *Eco* RI endonuclease and methylase are presented in Fig. 4. Most of the DNA sequence was determined on both templates with more than one primer. In sections where only one primer was available, or only one strand was used as template, reactions were performed more than once with the same primer. The sequence in lower case letters represents DNA from the M13 cloning vector and from pBH20 (via pPG30). The numbering of the nucleotide sequence begins with DNA originating from pMB1.

### *The Amino Acid Sequences*

*Agreement of Predicted Sequences with Experimental Observations*—The methionines selected as start codons for the endonuclease and methylase precede the only open-reading frames of the correct length to accommodate the subunit molecular weights of the two proteins. The size of the endonuclease predicted from the nucleotide sequence is 31,065 daltons compared with 29,500 daltons measured by sodium dodecyl sulfate gel electrophoresis (16). The size of the methylase predicted from the nucleotide sequence is 38,050 daltons compared with 36,000 daltons measured by sodium dodecyl sulfate gel electrophoresis (16). Other potential reading frames in either orientation have frequent translational stop codons with no open translational frame starting from an initiation codon specifying a peptide over 49 amino acids. The direction of transcription and translation corresponds to the direction predicted from measuring the effect of the *lac* promoter on levels of *Eco* RI endonuclease and methylase in strains containing pPG30 and pPG31. The boundaries of the translated segments fall within the gene boundaries established by subcloning experiments (Fig. 1) (4, 6, 13).

The amino acid sequences derived from the nucleotide sequence correlate well with available data derived from analyses of the proteins. $NH_2$-terminal sequence analysis by Hsiang[2] confirms the start point selected for the endonuclease. There is reasonable agreement between previously published amino acid compositions of the endonuclease and methylase and the amino acid compositions predicted from the nucleotide sequence (Table I). Further substantiation of the predicted amino acid sequence of the endonuclease was accomplished by comparing the composition of peptides obtained

[2] M. Hsiang, personal communication.

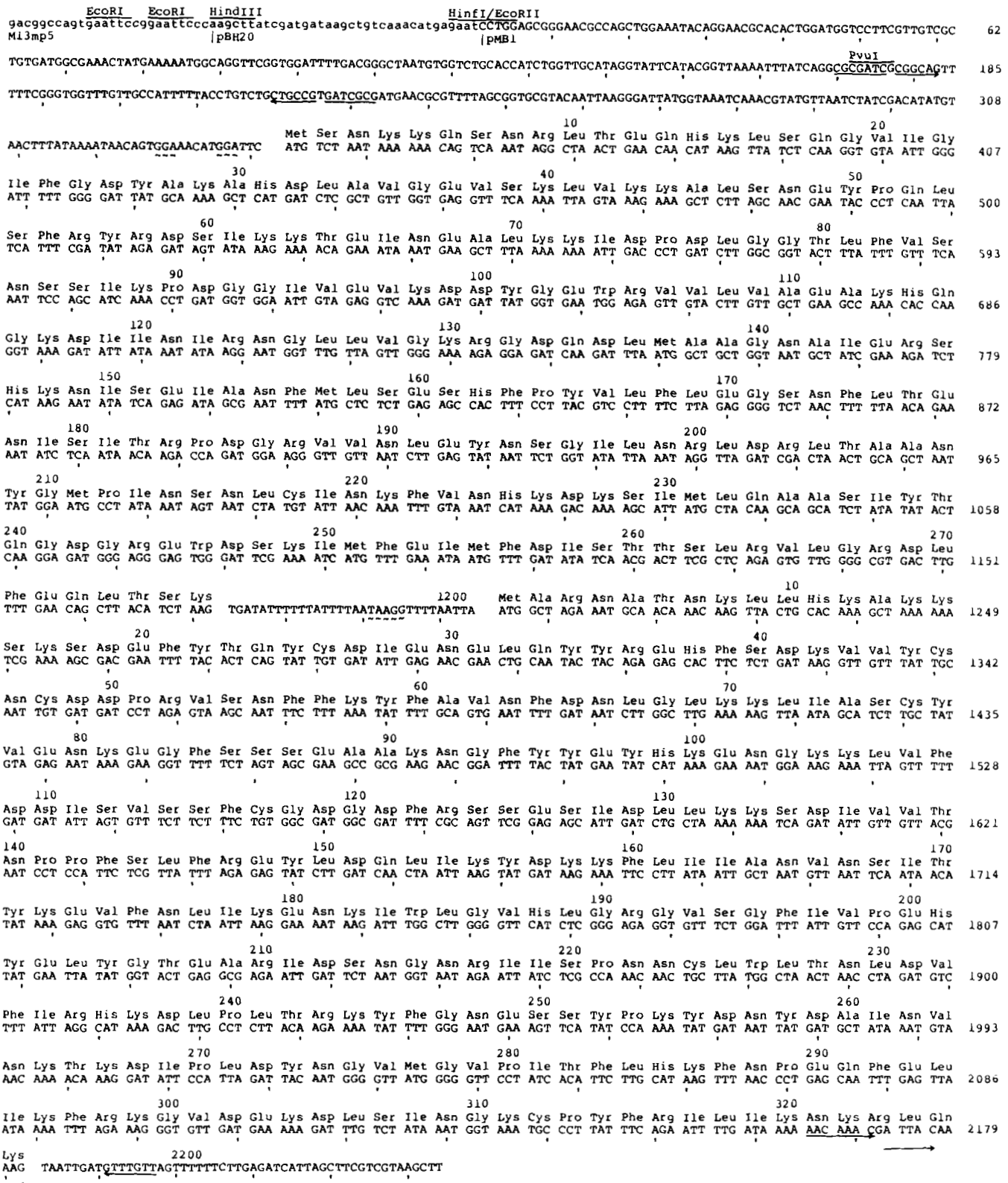


FIG. 4. **Nucleotide sequence of the DNA cloned in M13mp5 and amino acid sequences of the *Eco* RI endonuclease and methylase.** Sequence originating from M13mp5 and pBH20 (via pPG30) is depicted in *lower case letters*. The boundaries are marked with | . Two *Eco* RI sites precede the *Hin*dIII site of M13mp5. The *Eco* RI site expected in pPG30 has been converted to a *Hin*fI site by loss of one base pair. Nucleotide numbering starts with sequence originating from pMB1 (*upper case letters*). Amino acid numbering starts with the initial methionine of each protein. Homology with the 3' end of 16 S ribosomal RNA is indicated with *wavy underlines*. Inverted repeats discussed in the text are underlined with *arrows*.

TABLE I

*Amino acid composition of the Eco RI endonuclease and methylase*

| | Endonuclease | | Methylase | |
|---|---|---|---|---|
| | Predicted from nucleotide sequence[a] | Reported[b] | Predicted from nucleotide sequence[a] | Reported[b] |
| Ala | 15 | 15.6 | 10 | 9.8 |
| Arg | 14 | 12.9 | 13 | 12.0 |
| Asp | 18 | 40.0 | 23 | 61.1 |
| Asn | 20 (38) | | 28 (51) | |
| Cys | 1 | 1.8 | 7 | 6.5 |
| Glu | 17 | 29.0 | 20 | 25.7 |
| Gln | 9 (26) | | 5 (25) | |
| Gly | 21 | 16.8 | 17 | 17.8 |
| His | 6 | 5.6 | 7 | 6.6 |
| Ile | 23 | 21.9 | 23 | 19.5 |
| Leu | 27 | 28.2 | 28 | 26.0 |
| Lys | 22 | 21.1 | 35 | 30.4 |
| Met | 6 | 4.7 | 1 | 0.9 |
| Phe | 11 | 16.2 | 23 | 29.7 |
| Pro | 6 | 4.7 | 1 | 0.9 |
| Ser | 24 | 18.9 | 23 | 29.7 |
| Thr | 10 | 9.9 | 9 | 8.4 |
| Try[c] | 2 | | 2 | |
| Tyr | 8 | 7.6 | 21 | 17.9 |
| Val | 16 | 15.5 | 19 | 16.5 |

[a] $NH_2$-terminal methionine not included.

[b] Reported mole per cent values (16) are multiplied by 276 residues for the endonuclease and 325 for the methylase, the totals obtained from the predicted sequence excluding the $NH_2$-terminal methionine.

[c] Tryptophan was not determined.

TABLE II

*Amino acid composition of cyanogen bromide peptides*

| Amino acids | Peptide positions | | | | | |
|---|---|---|---|---|---|---|
| | 138–157 | | 252–255 | | 256–277 | |
| | Predicted | Found | Predicted | Found | Predicted | Found |
| Ala | 4 | 3.6 | | | | 0.1 |
| Arg | 1 | 1.7 | | | 2 | 2.5 |
| Asp | 3 | 3.3 | | 0.2 | 2 | 2.0 |
| Glu | 2 | 2.0 | 1 | 1.1 | 2 | 2.1 |
| Gly | 1 | 1.1 | | 0.1 | 1 | 1.2 |
| His | 1 | 1.0 | | | | 0.1 |
| Ile | 3 | 3.1 | 1 | 1.0 | 1 | 1.0 |
| Leu | | 0.2 | | 0.3 | 4 | 3.9 |
| Lys | 1 | 1.2 | | | 1 | 1.0 |
| Met[a] | 1 | 0.5 | 1 | 0.7 | | |
| Phe | 1 | 1.0 | 1 | 0.8 | 2 | 1.8 |
| Ser | 2 | 1.8 | | 0.2 | 3 | 3.6 |
| Thr | | | | 0.2 | 3 | 2.7 |
| Val | | 0.1 | | | 1 | 1.0 |

[a] Methionine is determined as homoserine in the first two peptides; none is found in the third, indicating it is the COOH-terminal peptide.

after cyanogen bromide cleavage of the protein with the corresponding peptides predicted from the nucleotide sequence. Table II gives the compositions of three peptides including residues 138–157 near the middle and 252–277 at the COOH terminus of the protein. The close match between the measured and predicted values confirms the postulate that the peptides predicted from the nucleotide sequence exist in the purified protein. These data strengthen our confidence that the derived amino acid sequences are correct. Hsiang's $NH_2$-terminal sequence analysis indicates that the methionine is not present in the purified endonuclease.[2] Comparison of the predicted and measured amino acid compositions suggests that the $NH_2$-terminal methionine is also removed from the methylase.

Newman *et al.* (10) present additional data verifying the predicted amino acid sequences. They find that the $NH_2$-terminal alanine as well as the methionine is absent in purified methylase.

*Search for Sequence Homologies*—Since the endonuclease and methylase recognize a common substrate sequence, and the presence of the endonuclease without a functional methylase is apparently lethal, Boyer *et al.* (16) postulated that the two enzymes might have evolved from a common ancestral gene and suggested that sequence homology would be found between the two proteins in regions involved in site recognition. Subsequently, physical characterization of the enzymes and studies on their reaction mechanisms have revealed extensive differences in the way the two enzymes recognize the common substrate sequence (3, 4). Comparison of the nucleotide and amino acid sequences of these two enzymes reveals no extensive regions of sequence homology, and it now seems likely that recognition of the same DNA sequence by these two enzymes represents convergent evolution from unrelated precursors. This is not the only example of two nonhomologous proteins recognizing the identical DNA sequence. The products of the *cI* and *cro* genes of phage λ bind to the same set of operator sequences but show no obvious sequence homology (17).

The amino acid sequences of the endonuclease and methylase do contain a number of common tri- and dipeptides, and Newman *et al.* (10) have identified a region with five out of nine homologous amino acids. Solution of the tertiary structure will reveal whether any of these have functional homology.

*Secondary Structure Predictions*—The translated sequences of the proteins were analyzed for expected secondary structure using the method of Chou and Fasman (18, 19). Newman *et al.* (10) present similar secondary structure predictions based on their identical sequences. While their predictions differ from ours in detail, the general impressions gained from both are the same, particularly the substantial differences between the endonuclease and methylase. The results of our analysis are presented in the miniprint (Tables IV and V and Fig. 7 and 8) together with a discussion of the differences between the two predictions. Since the current state of the art of secondary structure prediction is imperfect (20, 21), these predictions should be viewed with appropriate caution, and are mainly useful in suggesting further experiments such as the location of sites for specific mutagenesis. With this caveat in mind, we have drawn the following conclusions.

1) The endonuclease and methylase exhibit gross differences in predicted molecular architecture. The former contains significantly more predicted α helix than the methylase, while the latter contains substantially more amino acid residues which are predicted to be neither α helix nor β sheet. Specifically, the endonuclease is predicted to be 35% α helix and 26% β sheet. Newman *et al.* (10) predicted 39% and 17% for α helix and β sheet, respectively. There are two experimental measurements of α helix and β sheet in the endonuclease: Goppelt *et al.* (22) obtained values of 36% α helix and 21% β sheet, while Newman *et al.* (10) report 31% and 13%, respectively. There is very good agreement of all these results for α helical content. The β sheet values are more variable, but all values indicate that the endonuclease contains more α helix than β sheet. Our prediction for the methylase is 19% α helix and 30% β sheet; Newman *et al.* (10) predict 18% and 27% for α helix and β sheet, respectively. In both predictions, the β sheet content is greater than the α helix content. Newman *et al.* (10) report circular dichroism results of 11% α helix and 5% β sheet for this enzyme; however, they regard these values with caution. Indeed, caution is required in the

interpretation of all these secondary structure estimations. Even so, the differences between the two molecules appears significant.

2) *Eco* RI endonuclease can be divided into two domains of dissimilar architecture. The region from residues 1 to 174 is predicted to consist of 44% $\alpha$ helix and 18% $\beta$ sheet in alternating segments; this is a well known motif in protein structures, forming the basis of the domains which bind nucleotides in a variety of oxidation-reduction enzymes (23). Residues 175–277 are predicted to contain 20% $\alpha$ helix and 39% $\beta$ sheet. Besides the differences in amount of predicted $\alpha$ helix and $\beta$ sheet, the $NH_2$- and COOH-terminal regions appear different in a more subtle sense. The predictions are quite consistent in the $NH_2$-terminal region whereas considerable ambiguity is encountered in the COOH-terminal region; there are competing tendencies to predict $\alpha$ helix and $\beta$ sheet in the same region. The differences in statistical associations of amino acid residues in these two regions of the molecule support the prediction of two domains, even if details of the predicted secondary structure do not prove to be accurate. The prediction of two structural domains in the endonuclease is interesting since there is considerable data to suggest that other DNA recognition proteins, for example, the *lac* and λ repressors and CAP, are divided into two functional domains (17, 24). Limited proteolysis and site-directed mutagenesis will be employed to further analyze the endonuclease for the presence of different functional domains.

Both enzymes exhibit clusterings of amino acids which are potential points of contact with substrate. The most obvious of these are several clusters of basic residues, which might be expected to interact with phosphate moieties in the backbone of DNA. Many of these occur in regions predicted to be devoid of secondary structure, *i.e.* in loops between secondary structure elements. In enzymes with known structure, similar loops commonly form the active sites. The remaining basic clusters in the *Eco* RI enzymes can be found in regions predicted to form $\alpha$ helices, primarily at the COOH end of those helices. These clusters could also be playing a structural role since there is a statistical tendency for helices to exhibit this charge distribution (18, 19). Both enzymes also contain clusters which are rich in hydroxyls: in the endonuclease, Ser(84)-Asn-Ser-Ser and Ser(259)-Thr-Thr-Ser; in the methylase, Ser(85)-Ser-Ser, Tyr(95)-Tyr-Glu-Tyr, Ser(112)-Val-Ser-Ser, and Ser(144)-Ser-Glu-Ser.

*Internal Homologies in the Amino Acid Sequences*—One of the most striking features of the endonuclease amino acid sequence is an iterative tetrapeptide beginning at residue 250, where the sequence reads: Ile-Met-Phe-Glu-Ile-Met-Phe-Asp. The DNA which encodes this region is equally iterative with only two base changes, one in the wobble position of the Ile and the other leading to the Glu-Asp substitution. This is suggestive of a duplication event in the history of the endonuclease. This octapeptide is predicted to form a strand of $\beta$ sheet. Furthermore, commencing at residue 230, there is a similar tetrapeptide, Ile-Met-Leu-Gln, which is also predicted to lie within a strand of $\beta$ sheet. If this prediction is accurate, the methionines and acidic groups would project out from one side of the sheet and the hydrophobic isoleucines and phenylalanines would project from the other side, probably towards the interior of the molecule. The residues immediately preceding both these regions are also similar in character, consisting of aspartate, lysine, and serine in different order. Without further analysis, the function of the region cannot be assigned; however the occurrence of this periodicity in predicted $\beta$ sheets is tantalizing since the distance between amino acid side chains in $\beta$ sheets is 3.4 Å (25), which means that the repetitive amino acids occur at multiples of the base-stacking distance in helical DNA.

A notable feature of the methylase sequence is the repeat of the tripeptide, Leu-Ile-Lys, four times beginning at residues 153, 177, 294, and 318. The first pair and the last pair are each separated by 21 amino acids. Three of the tripeptides fall within predicted $\alpha$ helices while the fourth overlaps a predicted $\beta$ sheet strand. It remains a matter of speculation whether this pattern has a significant relationship to the function of the methylase.

Additional limited homology in the amino acid sequences adjacent to the two pairs and underlying genetic homologies are discussed by Newman *et al.* (10).

### *The Nucleotide Sequence*

In comparing the nucleotide sequence derived from the two template orientations represented by M13-RI6a and M13-RI13b, a perfect match was obtained with two exceptions. One of these proved to be a mutation which will be discussed later. The other unclear area was found in the noncoding region 168–181. This 14-base pair region contains 12 G-C pairs and the following restriction sites: one *Hha* I, two *Fnu*DII (*Tha* I), one *Fnu*4HI, and one *Pvu* I site which is also a *Sau*3AI (*Mbo* I) site. The presence of at least one *Fnu*DII and the *Hha* I and *Pvu* I sites was established by analysis with these enzymes. The *Pvu* I site is at the center of a 10-base pair sequence with 2-fold symmetry. This kind of symmetrical sequence rich in GC residues can form secondary structure which leads to compression and poor resolution on sequencing gels (26) and may also interfere with enzymatic chain elongation. In this case, we were able to read sequence through the *Sau*3AI site on both templates. From the appearance of autoradiographs of the sequencing gels, it was clear that the sequence immediately following the site could not be read accurately. Therefore, the sequence presented was established by combining the apparently normal regions from each template. An inverted repeat which involves most of these residues is marked in Fig. 4. Nucleotides 170–183 and 220–233 would pair (in mRNA) with an energy of −35.2 kcal (27). The function of this noncoding region of the sequence is not known, but this is the most impressive free energy change of any of the possible stem and loop structures in this sequence.

*Base Composition and Codon Usage*—The DNA sequence shown in Fig. 4 is 65% A + T, significantly higher than that of ColE1 (28) and *E. coli* (29), the host cell for both pMB1 and ColE1. Furthermore, a shift in base composition occurs near residue 260, so that residues 1–260 are 49% A + T, similar to ColE1 and *E. coli*. The codon usage in the endonuclease and methylase genes reflects the high A + T content. There is a strong preference for codons which end with A or T, and in the case of arginine and leucine, where the choice is available, a preference is also manifest for those codons which begin with A or T. This pattern of codon usage is quite different from that seen in a number of *E. coli* chromosomal genes, and especially the ribosomal protein genes (Table III) (30–36). Post and Nomura (30) have argued that codon preference in the ribosomal protein genes generally corresponds to the most abundant tRNA species, reflecting the need for efficient translation of these proteins. In most cases, wobble pairing allows translation of the *Eco* RI codons by major tRNA species; however, exceptions occur. AGA and AGG for arginine and AUA for isoleucine are frequently used codons in the *Eco* RI sequence which are represented by minor tRNA species and are almost never selected in a number of other *E. coli* gene sequences (Table III). The frequent use of codons represented by minor tRNA species may affect the translational efficiency of these genes. Betlach *et al.* (6) speculated that pMB1 arose

TABLE III

*Codon usage (in frequency per 1000)*

| CODON | | ENDO + METH | RIBOSOMAL[1] | COLI MIX[2] |
|---|---|---|---|---|
| Arg | CGU | 2 | 42 | 19 |
| | CGC | 2 | 22 | 17 |
| | CGA | 5 | 0 | 4 |
| | CGG | 0 | 1 | 2 |
| | AGA | 27 | 1 | 2 |
| | AGG | 10 | 0 | 0 |
| Leu | CUU | 18 | 4 | 8 |
| | CUC | 7 | 3 | 8 |
| | CUA | 15 | 0 | 4 |
| | CUG | 5 | 61 | 70 |
| | UUA | 33 | 4 | 10 |
| | UUG | 13 | 3 | 9 |
| Ser | UCU | 27 | 22 | 19 |
| | UCC | 2 | 19 | 13 |
| | UCA | 17 | 1 | 7 |
| | UCG | 10 | 2 | 10 |
| | AGU | 10 | 4 | 5 |
| | AGC | 13 | 6 | 13 |
| Thr | ACU | 13 | 28 | 10 |
| | ACC | 0 | 19 | 23 |
| | ACA | 15 | 3 | 2 |
| | ACG | 3 | 2 | 13 |
| Pro | CCU | 18 | 4 | 4 |
| | CCC | 0 | 1 | 7 |
| | CCA | 10 | 5 | 8 |
| | CCG | 0 | 29 | 19 |
| Ala | GCU | 22 | 68 | 15 |
| | GCC | 3 | 12 | 28 |
| | GCA | 12 | 38 | 22 |
| | GCG | 5 | 25 | 42 |

| CODON | | ENDO + METH | RIBOSOMAL | COLI MIX |
|---|---|---|---|---|
| Gly | GGU | 25 | 40 | 32 |
| | GGC | 7 | 32 | 37 |
| | GGA | 13 | 1 | 4 |
| | GGG | 18 | 0 | 5 |
| Val | GUU | 35 | 43 | 23 |
| | GUC | 5 | 7 | 13 |
| | GUA | 13 | 38 | 13 |
| | GUG | 5 | 15 | 28 |
| Lys | AAA | 65 | 70 | 40 |
| | AAG | 30 | 25 | 9 |
| Asn | AAU | 60 | 4 | 13 |
| | AAC | 20 | 29 | 25 |
| Gln | CAA | 18 | 7 | 15 |
| | CAG | 5 | 24 | 31 |
| His | CAU | 15 | 4 | 7 |
| | CAC | 7 | 8 | 7 |
| Glu | GAA | 33 | 53 | 32 |
| | GAG | 28 | 14 | 20 |
| Asp | GAU | 60 | 13 | 25 |
| | GAC | 8 | 28 | 18 |
| Tyr | UAU | 37 | 4 | 11 |
| | UAC | 12 | 12 | 14 |
| Cys | UGU | 7 | 1 | 6 |
| | UGC | 7 | 5 | 6 |
| Phe | UUU | 43 | 8 | 30 |
| | UUC | 13 | 15 | 22 |
| Ile | AUU | 35 | 15 | 33 |
| | AUC | 10 | 34 | 35 |
| | AUA | 22 | 1 | 1 |
| Met | AUG | 12 | 22 | 25 |
| Trp | UGG | 7 | 4 | 6 |

[1] See Ref. 30.

[2] Codon usage for *lac*I (31), *lac*Y (32), trypA (33), recA (34, 35), lipoprotein (36), and part of RNAP (30).

by translocation of a segment of DNA into a common ancestor of ColE1. The features of the DNA sequence noted here suggest that the *Eco* RI genes may originate from a species whose DNA has high A + T content. We are extending our sequence analysis of pMB1 and comparing it to ColE1 DNA sequence in order to determine the boundaries of nonhomology.

*Transcriptional and Translational Signals*—We have previously suggested that the endonuclease and methylase are coordinately controlled. The level of both enzymes is altered by conditions which affect the *lac* control system when the endonuclease is adjacent to the *lac* promoter (pPG31) but not when the orientation is reversed (pPG30), and the magnitude of the effect is similar for both enzymes (4). Accordingly, the DNA sequence preceding the endonuclease initiation codon was examined for sequences homologous to known promoters. Several partial homologies occur, but none is convincing enough to select a promoter by inspection of the sequence alone. The potential promoter sequences most distant from the initiation of translation are probably located in DNA homologous to ColE1, but this does not automatically exclude them as possible choices. The potential promoters (Fig. 5) include five sequences that match reported Pribnow boxes and four more with the three most conserved bases (see Refs. 37 and 38 for recent reviews). Of these, only the Pribnow box beginning at position 156 has significant homology with the −35 region, and, in this case, the spacing of two homologous regions is one nucleotide less than that found in any published wild type promoter sequence. There is a sequence with very good homology to the consensus −35 region beginning at residue 200, but the accompanying Pribnow box would have to be an uncharacteristic -TGCCGT- or -CGTGAT-.

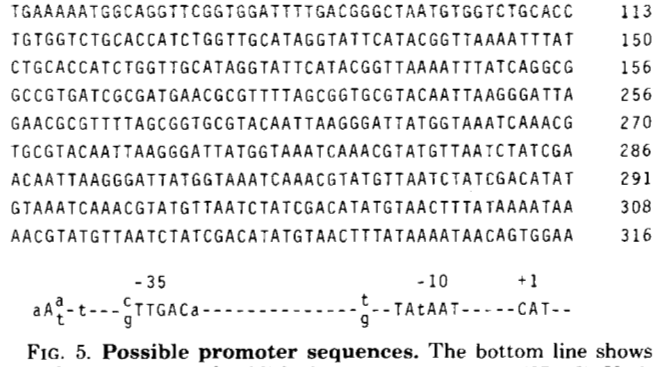


FIG. 5. **Possible promoter sequences.** The bottom line shows homologous regions of published promoter sequences (37, 38). Variation in spacing between the Pribnow sequence (−10) and the −35 region in wild type promoters is two more or one less than shown; variation in spacing between the Pribnow sequence and the messenger initiation site (+1) is ±2. *Capital letters* designate bases which occur in 61% or more of the sequences, *small letters* designate frequency between 46% and 61%. Two *small letters* designate a combined frequency of ≥70%. The *number* following each of the possible promoter sequences indicates the position of the first nucleotide of the Pribnow box, designated *. The sequences at positions 156, 256, 270, 286, and 316 match published Pribnow boxes. Those at positions 113, 150, 291, and 308 match the three most frequently occurring bases TA---T.

The mode of regulation of the *Eco* RI genes has not been elucidated. The restriction-modification system provides a

barrier to foreign DNA entering a cell containing both enzymes, but when a plasmid carrying the *Eco* RI genes enters a host cell with unmodified DNA, host DNA is restricted with low efficiency. The order of translation of the genes would appear to present a source of difficulty for such a plasmid. However, since the endonuclease requires at least two subunits while the methylase is active as a monomer (3), the methylase is functional first in spite of the translational order. The earlier presence of active methylase achieved by the subunit structure of the enzymes may be sufficient to account for the survival of the genes. Alternatively, there may be a separate promoter for the methylase. When different fragments (*Hin*dIII, *Hin*dIII-*Pst* I, and*Bgl* II-*Hin*cII), which contain the methylase but not the endonuclease gene, are cloned in a variety of sites and orientations, the resulting hybrid plasmids always confer the *Eco* RI modification phenotype on the host cell (4, 6, 13). Methylase specific activity was measured in one of these clones and found to be 5% of normal (6). Early expression of this amount of enzyme could be functionally important for the restriction-modification system. We are currently examining promoter activity in the regions preceding both methylase and the endonuclease genes.

The DNA sequence following each of the structural genes was examined for RNA polymerase chain-termination signals (38). A series of T residues occurs following both the endonuclease and methylase stop codons. Neither T series is preceded by a GC-rich region as is characteristic of some transcription-termination sequences. The sequence near the end of the methylase gene has two hyphenated inverted repeats which could form secondary structure in the message preceding the series of Ts. Unlike other messenger-termination sequences, these overlap the coding sequence. The possibility remains that the true termination sequence occurs beyond the sequence presented here.

Homology with the 3′ end of 16 S rRNA sequence (39) is indicated in Fig. 4 for both genes. The trinucleotide -GGA- which occurs twice, 6 and 14 base pairs before the endonuclease ATG, and the pentanucleotide -TAAGG- which occurs 14 nucleotides before the methylase ATG, are likely ribosome-binding sites. It has been suggsted that the termination codons UAA and UGA may also be part of the recognition signal for ribosomal initiation (40). One UAA codon begins at position 320 preceding the endonuclease Shine-Dalgarno sequence. Four UAA codons occur in the region of the methylase initiation site.

The sequence of the intercistronic region is over 90% A + T; in the coding strand, 19 of 32 nucleotides are Ts. The presence of stems and loops in intercistronic regions with the ribosome-binding site exposed in the loops has been noted in a number of sequences (32, 41). Residues 1144–1155 and 1221–1232 could pair to form a stem and loop structure ($\Delta G = -16.8$ kcal) (27). In scanning the entire sequence for possible secondary structure, however, the low potential for the intercistronic region to participate in any stem is more impressive than the presence of the stem and loop noted.

### *A Spontaneous Mutation in the Eco RI Endonuclease Gene*

The sequences obtained from the two templates, M13-RI6a and M13-RI13b, differed at residue 902. The M13-RI6a sequence was GCG and the complementary sequence obtained from M13-RI13b was CCC. Autoradiographs of sequencing gels from each template were clear and unambiguous. Therefore, template DNA was prepared from two other clones, M13-RI3a and M13-RI22b, representing independent isolates of the two templates orientations. The sequences read from both of these templates match the sequence obtained from M13-RI13b. M13-RI6a contains a spontaneous mutation in the *Eco* RI endonuclease gene which replaces the arginine at residue 187 with a serine. Endonuclease activity was measured in crude extracts of strain 71-18 infected with M13-RI6a and M13-RI3a. Specific cleavage at *Eco* RI sites was undetectable in M13-RI6a/71-18 under assay conditions in which about 1% of the activity of M13-RI3a/71-18 would have been detectable. Extracts of the two strains contained comparable amounts of *Eco* RI methylase.

The specific activity of the *Eco* RI enzymes is lower in the M13-RI strains than in plasmid-containing strains. Furthermore we have been unable to measure *Eco* RI restriction of λ phage in the M13-RI strains even when a substantial amount of *Eco* RI endonuclease activity is measurable *in vitro*. Therefore, we transferred the mutant sequence back into a plasmid derivative of pPG31 for further study of the mutant endonuclease.

Two plasmids, pMG31 and pMG31-6 (Fig. 2*B*), were constructed. pMG31-6 contains the 1685-base pair *Hin*dIII fragment carrying the mutant sequence from M13-RI6a, and pMG31 contains the analogous fragment from pPG31. Both plasmids differ from pPG31 (Figs. 1 and 2*B*) in that the small *Hin*dIII fragment which spans one of the *Eco* RI-*Eco* RII junctions has been deleted.

The activity of the *Eco* RI enzymes in *E. coli* strain 294 (42) carrying pMG31 or pMG31-6 was measured after growth in minimal medium containing glycerol + isopropyl-1-thio-β-D-galactopyranoside. These growth conditions yield the highest specific activity of the *Eco* RI enzymes in pPG31/294 (4). The specific activity of the *Eco* RI methylase in the two strains is approximately equivalent, 860 and 1090 units/mg of protein, in pMG31 and pMG31-6, respectively. Endonuclease activity is measured in relative numbers because of the nature of the assay used. We could not use quantitative assays which rely on measuring conversion of circular to linear forms in plasmids with one *Eco* RI site (43, 44), or on methods which quantitate cleavage by measuring the increase in 5′ termini (45), since these methods do not distinguish *Eco* RI specific cuts from nonspecific cuts. *Eco* RI endonuclease activity is high enough in normal strains that nonspecific background cleavage is not a problem. However, to assess activity in the mutant, we had to rely on the appearance of specific DNA fragments. Quite good relative values can be assigned by comparing different times of digestion and different dilutions of the extracts. The mutant endonuclease clearly cleaves DNA at *Eco* RI sites. The level of enzyme activity is 0.3% of the level observed in pMG31. Further characterization of the mutant endonuclease is underway.

*Acknowledgments*—We thank Patricia Clausen for preparing the manuscript and Sonja Bock for editing the sequence in the computer to produce Fig. 4. We also thank Keith Backman and Jon Lawrie for comments on the manuscript and Dr. Paul Modrich for interesting and useful discussions of results prior to publication.

REFERENCES

1. Hedgpeth, J., Goodman, H. M., and Boyer, H. W. (1972) *Proc. Natl. Acad. Sci. U. S. A.* **69,** 3448–3452
2. Dugaiczyk, A., Hedgpeth, J., Boyer, H. W., and Goodman, H. M. (1974) *Biochemistry* **13,** 503–512
3. Modrich, P. (1979) *Q. Rev. Biophys.* **12,** 315–369
4. Rosenberg, J. M., Boyer, H. W., and Greene, P. J. (1980) in *The Site Specific Restriction Endonucleases* (Chirikjian, J. G., ed), Elsevier-North Holland, in press
5. Rosenberg, J. M., Dickerson, R. E., Greene, P. J., and Boyer, H. W. (1978) *J. Mol. Biol.* **122,** 241–245
6. Betlach, M., Hershfield, V., Chow, L., Brown, W., Goodman, H. M., and Boyer, H. W. (1976) *Fed. Proc.* **35,** 2037–2043
7. Sanger, F., Nicklen, S., and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–5467
8. Messing, J., Gronenborn, B., Müller-Hill, B., and Hofschneider,

P. H. (1977) *Proc. Natl. Acad. Sci. U. S. A.* **74,** 3642–3646
9. Gronenborn, B., and Messing, J. (1978) *Nature* **272,** 375–377
10. Newman, A. K., Rubin, R. A., Kim, S.-H., and Modrich, P. (1981) *J. Biol. Chem.* **256,** 2131–2137
11. Goodman, H. M., Greene, P. J., Garfin, D. E., and Boyer, H. W. (1977) in *Nucleic Acid-Protein Recognition* (Vogel, H. J., ed), pp. 239–259, Academic Press, New York
12. Maxam, A. M., and Gilbert, W. (1977) *Proc. Natl. Acad. Sci. U. S. A.* **74,** 560–564
13. Bolivar, F., Rodriguez, R. L., Greene, P. J., Betlach, M. C., Heyneker, H. L., Boyer, H. W., Crosa, J. H., and Falkow, S. (1977) *Gene* **2,** 95–113
14. Itakura, K., Hirose, T., Crea, R., Riggs, A. D., Heyneker, H. L., Bolivar, F., and Boyer, H. W. (1977) *Science* **198,** 1056–1063
15. Backman, K. (1980) *Gene* **11,** 169–171
16. Boyer, H. W., Greene, P. J., Meagher, R. B., Betlach, M. C., Russel, D., and Goodman, H. M. (1975) *Fed. Eur. Biochem. Soc. Symp. (Budapest)* **34,** 23–37
17. Ptashne, M., Jeffrey, A., Johnson, A. D., Maurer, R., Meyer, B. J., Pabo, C. O., Roberts, T. M., and Sauer, R. T. (1980) *Cell* **19,** 1–11
18. Chou, P. Y., and Fasman, G. D. (1974) *Biochemistry* **13,** 211–222
19. Chou, P. Y., and Fasman, G. D. (1974) *Biochemistry* **13,** 222–245
20. Schulz, G. E., Barry, C. D., Friedman, J., Fasman, G. D., Chou, P. Y., Finkelstein, A. V., Lim, V. I., Ptifsyn, O. B., Kabat, E. A., Wu, T. T., Levitt, M., Robson, B., and Nagano, K. (1974) *Nature* **250,** 140–142
21. Matthews, B. W. (1975) *Biochim. Biophys. Acta* **405,** 442–451
22. Goppelt, M., Pingoud, A., Maas, G., Moyer, H., Koster, H., and Frank, R. (1980) *Eur. J. Biochem.* **104,** 101–107
23. Rossman, M. G., Liljas, A., Branden, C. I., and Banaszak, L. J. (1975) in *The Enzymes* (Boyer, P. D., ed) Vol. XI, pp. 61–102, Academic Press, New York
24. Ogata, R. T., and Gilbert, W. (1979) *J. Mol. Biol.* **132,** 709–728
25. Church, G. M., Sussman, J. L., and Kim, S.-H. (1977) *Proc. Natl. Acad. Sci. U. S. A.* **74,** 1458–1462
26. Sanger, F., and Coulson, A. R. (1978) *FEBS Lett.* **87,** 107–110
27. Tinoco, I., Borer, P. N., Dengler, D., Levine, M. D., Uhlenbeck, O. C., Crothers, D. M., and Grolla, J. (1973) *Nature New Biol* **246,** 40–41
28. Falkow, S. (1975) in *Infectious Multiple Drug Resistance* (Lagnado, J. R., ed), pp. 98–143, Pion Limited, London
29. Shapiro, H. S. (1970) *CRC Handbook of Biochemistry Selected Data For Molecular Biology* (Sober, H. A., ed) 2nd Ed, pp H80–H88, The Chemical Rubber Co., Cleveland
30. Post, L. E., and Nomura, M. (1980) *J. Biol. Chem.* **255,** 4660–4666
31. Farabaugh, P. J. (1978) *Nature* **274,** 765–769
32. Büchel, D. E., Gronenborn, B., and Müller-Hill, B. (1980) *Nature* **283,** 541–545
33. Nichols, B. P., and Yanofsky, C. (1979) *Proc. Natl. Acad. Sci. U. S. A.* **76,** 5244–5248
34. Horii, T., Ogawa, T., and Ogawa, H. (1980) *Proc. Natl. Acad. Sci. U. S. A.* **77,** 313–317
35. Sancar, A., Stachelek, C., Konigsberg, W., and Rupp, W. D. (1980) *Proc. Natl. Acad. Sci. U. S. A.* **77,** 2611–2615
36. Nakamura, K., Pirtle, R. M., Pirtle, I. L., Takeishi, K., and Inouye M. (1980) *J. Biol. Chem.* **255,** 210–216
37. Siebenlist, U., Simpson, R. B., and Gilbert, W. (1980) *Cell* **20,** 269–281
38. Rosenberg, M., and Court, D. (1979) *Annu. Rev. Genet.* **13,** 319–353
39. Shine, J., and Dalgarno, L. (1974) *Proc. Natl. Acad. Sci. U. S. A* **71,** 1342–1346
40. Atkins, J. F. (1979) *Nucleic Acids Res.* **7,** 1035–1041
41. Selker, E., and Yanofksy, C. (1979) *J. Mol. Biol.* **130,** 135–143
42. Backman, K., Ptashne, M., and Gilbert, W. (1976) *Proc. Natl Acad. Sci. U. S. A.* **73,** 4174–4178
43. Greene, P. J., Betlach, M. C., Boyer, H. W., and Goodman, H. M (1974) in *DNA Replication (Methods in Molecular Biology)* (Wickner, R. B., ed) Vol. 7, pp. 87–111, Marcel Dekker, Inc. New York
44. Lackey, D., and Linn, S. (1980) *Methods Enzymol.* **65,** 26–28
45. Berkner, K. L., and Folk, W. R. (1980) *Methods Enzymol.* **65,** 28–36

Additional references are found on p. 2153

SUPPLEMENTARY MATERIAL TO

Sequence Analysis of the DNA Encoding the EcoRI Endonuclease and Methylase

by Patricia J. Greene, Madhu Gupta, Herbert W. Boyer, William E. Brown and John M. Rosenberg

## MATERIALS AND METHODS

Enzymes — HsuI (HindIII), EcoRII, and PstI were prepared as described (46). T4 polynucleotide ligase was prepared as described by Panet et al. (47). HinfI, TaqI, Sau3aI, FnuDII, AvaI, and Exonuclease III were purchased from New England Biolabs. DNA polymerase I, Klenow fragment A, was purchased from Boehringer.

Strains — Plasmids and phage are depicted in Figs. 1 and 2 and are described in ref. 4, 8, and 9. E. coli 294 (EndoI, $r_k^-m_k^+$, $B_1^-$, $pro^-$, $hsr^-$, $hsm^+$) (42), GM31 ($dcm^-$, $B_1^-$, $gal^-$, $ara^-$, $lac^-$, $xyl^-$, $thr^-$, $leu^-$, tonA, tsx, str) (48) and 71-18 (Δ[lac, pro], $F'lacI^QZ\Delta M15pro^+$) (8) were used as background strains.

Electrophoresis of DNA — Analytical and preparative electrophoresis of restriction fragments was carried out on vertical agarose or acrylamide slab gels in 90 mM Tris borate, 1 mM EDTA, pH 8.3 (Tris-borate) (49). Single-stranded phage DNA was analyzed as described by J. Messing by electrophoresis on 1% agarose gels in 40 mM Tris, 20 mM NaAcetate, pH 8.2, 18 mM NaCl, 2 mM EDTA (Tris acetate-NaCl).[1] Sequencing gel electrophoresis was carried out as described by Sanger and Coulson (26).

Plasmid DNA — Plasmids were maintained in the $dcm^-$ strain, GM31. Plasmid DNA was prepared as described by Bolivar and Backman (50).

Sequencing procedures — Nucleotide sequence was determined using the dideoxyribonucleotide triphosphate chain termination method of Sanger et al. (7) with the following modifications. The reaction mixtures contained all four [$\alpha^{32}$P] deoxynucleotide triphosphates (400 Ci/mmol, Amersham), and the chase mixture contained additional DNA polymerase I, Klenow fragment A.

Preparation of primers — Restriction endonuclease digests of plasmid DNA were fractionated by electrophoresis on polyacrylamide slab gels in Tris-borate and stained in ethidium bromide (1 μg/ml). DNA bands were visualized under ultraviolet light and the acrylamide containing the desired fragments excised. DNA fragments of larger than 80 base pairs were recovered by electroelution in dialysis bags in 0.025 M Tris acetate, pH 8.0, 2 mM EDTA, at 40 V for 2-4 h depending on the fragment size (51). Smaller fragments were recovered by finely mincing the acrylamide and soaking in 0.3 M $NH_4OAc$, 0.01 M $Mg(OAc)_2$, 0.01% sodium dodecyl sulfate (SDS), 1 mM EDTA at 37°C for 4 h (12). The solutions of DNA obtained by both of these methods were filtered through a siliconized glass wool plug in 1000 μl pipetman tip, phenol extracted once, chloroform extracted twice, and the DNA recovered by ethanol precipitation. Precipitates were washed once with 95% ethanol and dissolved in distilled water. Recovery of the larger fragments, estimated by comparison to standards on polyacrylamide gels stained in ethidium bromide, was 50-80%. Recovery of fragments too small to analyze by this method was assumed to be 50%.

Purified fragments were sometimes digested with Exonuclease III (52,53) before use as primers. The exact reaction conditions had to be determined empirically for each fragment. Typically 1 pmol of a 200 base pair fragment in a final volume of 10 μl of 50 mM Tris-HCl, pH 8, 1 mM $MgCl_2$, 1 mM DTT was incubated with 4 units of ExoIII at 37°C for 15 to 30 min. Two primers of 30 and 50 base pairs were treated with 2 units of enzyme for 2 and 5 min respectively. Sequence obtained from fragments treated in this way usually started approximately 12 nucleotides from the 5' end of the fragment, but this distance was sometimes much greater (Fig. 3). Some fragments were so resistant to Exonuclease III treatment that useful primers could not be obtained by this treatment.

Computer analysis — The DNA sequence was analyzed using programs from the Computation Laboratory, Department of Biochemistry and Biophysics, University of California, San Francisco.[2]

Cyanogen bromide peptides — Cyanogen bromide cleavage of the endonuclease was performed as described (55), and the peptides were isolated from preparative maps (56).

Measurement of EcoRI endonuclease and methylase specific activity in pMG31 and pMG31-6 — 80 ml minimal medium (M-9 salts supplemented with $B_1$, casamino acids, and 0.4% glycerol) plus IPTG at a concentration of $2 \times 10^{-3}$ M and 2 μg/ml tetracycline, was inoculated with 2.5 ml of an overnight culture of 294 carrying pMG31 or pMG31-6 (Fig. 2B). After 5-7 h of growth at 37°C, the cultures were harvested and the cell pellets were frozen. Cell pellets were suspended in 10 mM $K_2HPO_4$-$KH_2PO_4$, 0.2 M NaCl, 0.1 mM EDTA, 7 mM βME, pH 7.0 and lysed by sonication, taking care to maintain the temperature below 10°C. The sonicates were centrifuged at 18,000 RPM in a Sorvall SS34 rotor for 30 min at 4°C. The supernatants were assayed for EcoRI endonuclease and methylase, and the protein concentration was determined by the method of Lowry (55). Methylase assays were as previously described (58). Endonuclease assays utilized DNA from pML21 (59), a plasmid with two EcoRI sites, so that the appearance of characteristic DNA fragments could be used to distinguish EcoRI specific cleavage from non-specific cleavage. The relative amount of enzyme activity in the extracts was estimated by comparing digests at different dilutions and different times of digestion.

## USE OF M13-RI RECOMBINANT PHAGE TO OBTAIN TEMPLATE DNA

Cloning the EcoRI genes in M13mp5 — Methods for using M13mp5 as a cloning vehicle have been described in detail (8,9).[1] M13mp5 encodes part of the β-galactosidase gene which complements a partial lac deletion in the host strain (E. coli 71-18). This segment of the β-galactosidase gene has been engineered to contain a HindIII site. Host cells infected with vector phage are $lac^+$. Cloning DNA in the HindIII site results in phage infected cells which are $lac^-$. Recombinant phage are distinguished from parental phage by growing phage infected cells on plates containing 5-bromo-4-chloro-3-indolyl-β-galactoside, an indicator dye which turns blue when hydrolyzed by β-galactosidase. M13mp5 was digested with HindIII and ligated to a HindIII fragment which carries the EcoRI genes (Fig. 2A). The ligation mixture was used to transform 71-18. Twenty-three colorless plaques were picked and grown overnight. Phage DNA was analyzed by agarose gel electrophoresis for the presence of the desired size insert. Phage from 11 of the 23 plaques contained a large insert, seven contained a smaller insert, and five were the same size as M13mp5.

Replicating form DNA from the 11 clones with the large insert was isolated and digested with HindIII. These clones contained the 580 and 1685 base pair HindIII fragments expected from a complete HindIII digest of the cloned DNA.

Orientation of the inserts in M13-RI hybrid phage — Recombinant phage containing the EcoRI genes in opposite orientations were identified by mixing phage DNA from two clones and analyzing for the presence of duplex DNA (54).[3] One isolate, M13-RI6a, was selected as a standard strain. A 50 μl sample from an overnight culture of each of the ten other clones was centrifuged and mixed with 50 μl of culture medium from M13-RI6a. The mixtures were made 0.1% SDS, extracted with phenol at 65°C, extracted twice with chloroform, and ethanol precipitated. The pellets were washed with 70% ethanol and dissolved in 20 μl of 6 mM Tris-HCl, 6 mM $MgCl_2$, 6 mM βME, 50 mM NaCl, pH 7.5 (HinfI reaction buffer), and incubated 15 min at 65°C to promote hybridization. Duplex DNA was detected by sensitivity to cleavage by HinfI (Fig. 6). Five strains were the same orientation as M13-RI6a and were not digested by HinfI. Five were the opposite orientation and yielded a restriction pattern corresponding to four of the five fragments in a HinfI digest of EcoRII fragment from pMB1. The 440 and 336 base pair fragments occur at the ends of the EcoRII fragment and are not expected in a HinfI digest of the template. The presence of the 336 base pair fragment was explained when the sequence of the template was determined (see Fig. 4). The EcoRI end, which had been filled in for the construction of pPG30, lost one base pair, resulting in the formation of a HinfI site at the junction with the EcoRII fragment.

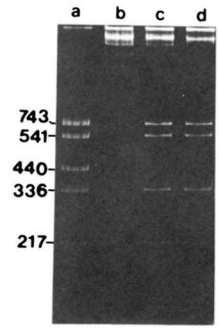


Fig 6. Polyacrylamide gel analysis of template orientation. Mixtures of phage DNA were incubated with HinfI. Those containing inserts of opposite orientation hybridize and are cleaved by HinfI. Isolates of the same orientation as M13-RI6a are also designated "a". Those of opposite orientation are designated "b". Lane a is a HinfI digest of the EcoRII fragment from pMB1. Fragment sizes in base pairs are indicated at the left. b, M13-RI6a + M13-RI3a; c, M13-RI6a + M13-RI13b; d, M13-RI6a + M13-RI22b.

Instability of M13-RI recombinant phage — Nine of the 11 cultures of phage containing the EcoRI genes also contained M13mp5 size phage visible on the agarose gel. In an attempt to obtain pure hybrids, the 11 cultures were plaque-purified. When one colorless plaque from each plated culture was picked, grown overnight and analyzed by agarose gel electrophoresis, eight had lost the insert and contained only M13mp5 size phage which had lost the ability to render the host strain $lac^+$. The other three retained a mixture of phage with the inserted EcoRI genes plus M13mp5 size phage. Since the hybrid phage replicate more slowly than M13mp5, we screened the plates again for the smallest size colorless plaques. By picking a number of these plaques, we were able to obtain 1.0 ml cultures in which only hybrid phage were visible on agarose gels. However, a mixed population always resulted when a large volume of phage infected cells was grown.

The rate of loss of inserts in filamentous phage has been measured in the case of fd carrying antibiotic resistance markers. A 2.7 kilobase fragment encoding resistance to two antibiotics was lost at a relatively high rate (54). Our inability to obtain preparative volumes of pure hybrid phage is probably the result of a similar or even greater instability. M13mp5 hybrids may be particularly unstable because of the sequence around the HindIII site. Most of the phage in our cultures which have lost the insert co-migrate with M13mp5 during agarose gel electrophoresis, as if the deletion event occurs at or very near the insertion site. Two EcoRI sites precede the HindIII site (see Fig. 4) and four EcoRI sites follow it.[4] A HindIII fragment inserted into this site is bordered by a 22 residue inverted repeat. This could form a stem structure which might frequently prevent copying of the insert.

In most cases the mixed population of phage DNA is not a problem in the sequencing reactions since primers prepared from the parent plasmid will hybridize only with the inserted sequence. However, we planned to use a synthetic primer which primes from M13mp5 sequence. If the contaminating smaller phage arise from a single deletion event near the HindIII site, both M13mp5 sequence and insert sequence would be copied. Therefore the template DNA was purified as described below.

Preparation of single-stranded template DNA — Two isolates, M13-RI6a and M13-RI13b, representing the two orientations of the EcoRI genes were grown as described.[1] The cultures were centrifuged, and phage DNA was recovered from the supernatant as described.[1] The DNA was further purified by centrifugation through linear gradients of 5-20% sucrose in 0.1 M Tris-HCl, pH 7.5, 1 M NaCl, in the Beckman SW40 rotor at 25 K, 4°C for 16 h. The two sizes of DNA did not form well separated peaks; however, by analyzing gradient fractions on 1% agarose Tris acetate-NaCl gels, fractions containing essentially pure template DNA were identified. Approximately 600 μg of pure template DNA was obtained from a one liter culture.

Single strand DNA was also prepared from 1 ml cultures of two other strains, M13-RI3a and M13-RI22b, representing independent isolates of the two insert orientations. The cultures were centrifuged, and the supernatants were adjusted to 0.5M NaCl, 4% polyethyleneglycol 6000 (PEG), incubated 10 min at 0°C, centrifuged in the eppendorf for 10 min and the pellets dissolved in 100 μl of 10 mM $MgSO_4$. Phage were reprecipitated by the addition of NaCl to 0.5 M and PEG to 4% and centrifuged in the eppendorf. The pellets were dissolved in 150 μl of 10 mM Tris-HCl, pH 7.5, 1 mM EDTA, phenol extracted at 60°C for 5 min, and chloroform extracted. The DNA was ethanol precipitated, dissolved in 20 μl of $H_2O$, and 5 μl was used for sequence determination.

## SECONDARY STRUCTURE PREDICTION

The method of Chou and Fasman (18,19) was used to predict the α helix and β sheet content of the endonuclease and methylase. We did not feel justified in predicting β turns given the difficulties encountered in predicting the major secondary structure elements in other proteins (20,21). The average potentials were calculated for each predicted element of secondary structure (Tables IV and V). The predictions are represented schematically in Figs. 7 and 8 for the endonuclease and methylase respectively. In many cases, considerable ambiguities were encountered, i.e. regions with both high α and high β potential. In these situations, "boundary probabilities" (18) were employed and/or potentials were calculated for several overlapping regions of the polypeptide chain in order to resolve the ambiguity. The criteria used to resolve each such ambiguity are given in the comments to the tables. We have also assigned a letter grade indication of the internal consistency of each prediction: an "A" grade indicates that the region clearly and unambiguously met all the prediction criteria. A "B" grade indicates that some ambiguity was encountered but that it was resolved in a straightforward fashion; e.g. in a region that had both α and β nuclei, one was clearly dominant. A "C" grade indicates that considerable ambiguity was encountered, and relatively indirect criteria had to be used to resolve it. In some cases, these grades were modified downward when the individual prediction appeared to present structural difficulties (see individual comments). These ambiguities were encountered in the carboxyl terminal region of the endonuclease and throughout the entire methylase sequence. The predictive methods used by Newman et al. (10) differ from ours in several respects: Slightly different data bases were used; they predicted β turns while we did not; and they used computerized probability calculations while we used symbolic clustering (19) to identify secondary structure elements. Each of the enumerated factors contributed towards the discrepencies, but their effects were minor. Most of the discrepencies occur in regions which have both high α and high β potential as discussed above, and our conclusion is that these differences represent a genuine limitation of the predictive method. Since the endonuclease has been crystallized (4,5), it will be interesting to compare these predictions with the X-ray structure when it becomes available.

[1] J. Messing, personal communication

[2] Described in "DNA Analysis Programs Manual" available by writing to Hugo Martinez, Department of Biochemistry, UCSF, San Francisco, California 94143.

[3] R. Hallewell, personal communication

[4] P. J. Greene, unpublished data

TABLE IV

Secondary Structure Predictions for EcoRI Endonuclease

| Region | Number of residues | Predicted structure | $\langle P_\alpha \rangle$ | $\langle P_\beta \rangle$ | Grade | Comment |
|---|---|---|---|---|---|---|
| Leu 10 - Gln 18 | 9 | α | 1.16 | 0.95 | B | 1 |
| Val 20 - Phe 24 | 5 | β | 0.96 | 1.39 | B | 1 |
| Ala 28 - Ser 47 | 20 | α | 1.17 | 0.97 | A- | 2 |
| Gln 52 - Tyr 57 | 6 | β | 0.97 | 1.11 | A- | |
| Glu 65 - Ile 73 | 9 | α | 1.19 | 0.89 | A+ | |
| Gly 79 - Val 83 | 5 | β | 0.99 | 1.23 | A* | |
| Ile 94 - Asp 99 | 6 | α | 1.14 | 1.12 | C | 3, 5 |
| Glu 103 - Gln 115 | 13 | α | 1.24 | 1.00 | B- | 4, 5 |
| Ile 119 - Gly 129 | 11 | β | 0.92 | 1.16 | B | 5 |
| Asp 133 - Ala 139 | 7 | α | 1.22 | 1.09 | B | 5 |
| Glu 152 - Phe 163 | 12 | α | 1.15 | 0.95 | A | |
| Tyr 165 - Leu 169 | 5 | β | 1.11 | 1.33 | C | 6 |
| Ile 179 - Arg 183 | 5 | β | 0.88 | 1.21 | A- | |
| Arg 187 - Leu 191 | 5 | β | 1.03 | 1.23 | C- | 7 |
| Leu 201 - Ala 207 | 7 | α | 1.17 | 1.04 | C | 8 |
| Tyr 209 - Ile 213 | 5 | β | 0.79 | 1.20 | C | 9 |
| Lys 221 - Lys 226 | 6 | α | 1.06 | 0.96 | B | 10 |
| Ile 230 - Gln 240 | 11 | β | 1.09 | 1.25 | C+ | 10, 11 |
| Ile 250 - Ile 258 | 9 | β | 1.13 | 1.31 | C | 12 |
| Leu 263 - Gly 267 | 5 | β | 1.03 | 1.16 | A | |
| Leu 270 - Lys 277 | 8 | α | 1.15 | 0.98 | B | |

Comments on Table IV

1. These two elements of secondary structure are very clear in their individual predictions but are crowded together.
2. This long helix could be two helices, kinked in the middle at Gly 36.
3. Prediction is very ambiguous here since $\langle P_\alpha \rangle \approx \langle P_\beta \rangle$. The helix was preferred because of the slightly higher probability and because of the presence of Glu 96, a residue rarely found in β sheets.
4. There is a strong tendency towards β in the middle of this region. The α helix was chosen because of the nucleus outside the ambiguous region, i.e. residues 110 to 115, and because of statistical preferences: Val and Leu are very often found near the center of α helices (these are responsible for the ambiguity), Glu is a good N-terminal residue, and Lys, His and Gln are the three most common C-terminal α helical residues.
5. The secondary structure elements in these regions are a little crowded, i.e. the loops are too short. However, it is not clear which predictions should be modified to relieve the crowding, hence they remain with this caveat.
6. The overlapping region Val 166 - Glu 177 also has high α potential ($\langle P_\alpha \rangle$ = 1.11 and $\langle P_\beta \rangle$ = 0.98). This possibility was not accepted because there was no α nucleation center outside the ambiguous region (note that the preceding helix cannot propagate through Pro 164), the overall β potential was much higher and the boundaries were clearly those of a β sheet. The close proximity to the preceding helix is a further problem here.
7. This prediction is extremely ambiguous since the region Arg 187 to Glu 192 gives $\langle P_\alpha \rangle$ = 1.11, $\langle P_\beta \rangle$ = 1.07. The method is saying there is secondary structure here, but cannot discriminate which kind. The β was chosen because it has the highest probability score.
8. This region has a more diffuse α nucleus than usually required, but it clearly has high α potential (composition $H_4i_3$). Hence the prediction, but with a low grade.
9. There are two problems here, one is the presence of Pro 211, which is rare in β sheets; the other is the crowding with the previous prediction. Hence the low grade.
10. The region from 221 to 240 contained many ambiguities. There is an α nucleus in the 221 to 226 region and a β nucleus between 237 and 240. Both could propagate into the middle. The helix propagation was stopped at Lys 226 because the region from Asn 224 to Ser 228 is a continuous stretch of six hydrophilic residues, four of which are charged, i.e. it was considered unlikely that over one turn of helix would be continuously charged.
11. The segment Ile 230 to Ala 235 contains both α and β nuclei, but the latter is statistically dominant ($\langle P_\alpha \rangle$ = 1.23, $\langle P_\beta \rangle$ = 1.34). There is also the previously noted β nucleus (237-240). The boundary residues are much more consistent with a β sheet.
12. The region has both α and β potential, especially between Met 251 and Phe 256 ($\langle P_\alpha \rangle$ = 1.20, $\langle P_\beta \rangle$ = 1.29). However, β clearly dominates the statistics. The region Ile 254 to Ile 258 has an unambiguous β nucleus ($\langle P_\alpha \rangle$ = 1.06, $\langle P_\beta \rangle$ = 1.39). The boundary residues are clearly those of a β sheet. Hence this prediction.

TABLE V

Secondary Structure Predictions for EcoRI Methylase

| Region | Number of residues | Predicted structure | $\langle P_\alpha \rangle$ | $\langle P_\beta \rangle$ | Grade | Comment |
|---|---|---|---|---|---|---|
| Leu 9 - Lys 17 | 9 | α | 1.16 | 0.87 | B- | 1 |
| Phe 21 - Cys 26 | 6 | β | 0.85 | 1.23 | A | |
| Ile 28 - Gln 33 | 6 | α | 1.22 | 0.87 | C | 2 |
| Val 43 - Cys 48 | 6 | β | 0.86 | 1.31 | A | |
| Phe 56 - Val 62 | 7 | β | 1.09 | 1.21 | C | 3 |
| Leu 72 - Val 78 | 7 | β | 1.01 | 1.25 | C | 4 |
| Phe 84 - Ala 90 | 7 | α | 1.13 | 0.81 | C | 5 |
| Asn 92 - Tyr 96 | 5 | β | 1.06 | 0.72 | B | |
| Leu 106 - Ile 111 | 6 | β | 1.09 | 1.23 | C- | 6 |
| Ser 127 - Leu 131 | 5 | β | 1.09 | 1.11 | C | |
| Asp 135 - Thr 139 | 5 | β | 1.02 | 1.37 | A | |
| Pro 142 - Lys 155 | 14 | α | 1.06 | 1.03 | C | 7 |
| Phe 160 - Ala 164 | 5 | β | 1.18 | 1.33 | B- | |
| Glu 173 - Glu 180 | 8 | α | 1.18 | 0.96 | B- | |
| Ile 183 - Val 187 | 5 | β | 1.03 | 1.29 | B- | |
| Val 193 - Val 198 | 6 | β | 0.95 | 1.29 | B- | 8 |
| Pro 199 - Leu 204 | 6 | α | 1.14 | 0.73 | B- | 8 |
| Cys 224 - Ile 234 | 11 | β | 1.07 | 1.21 | B- | 9 |
| Leu 241 - Phe 246 | 6 | β | 0.96 | 1.10 | C- | 10 |
| Val 276 - Leu 284 | 9 | β | 0.99 | 1.30 | B | |
| Pro 289 - Lys 299 | 11 | α | 1.12 | 0.92 | A- | |
| Tyr 314 - Ile 319 | 6 | β | 0.98 | 1.38 | A | |

Comments on Table V

1. The prediction is quite consistent, but the continuous stretch of hydrophilic and charged residues, which would wrap around the entire circumference of the helix, is a problem.
2. The overlapping region Leu 32 - Arg 36 is a β nucleus ($\langle P_\alpha \rangle$ = 0.90, $\langle P_\beta \rangle$ = 1.19). This was rejected because the α potential was somewhat higher and the amino acids on the C terminus are commonly found after α helices. A further problem is the crowding with the previous prediction.
3. The region also has significant α potential, however the β clearly dominates.
4. The overlapping region Leu 69 - Ala 74 is an α nucleus ($\langle P_\alpha \rangle$ = 1.21, $\langle P_\beta \rangle$ = 1.08); however, the β prediction was based on the existence of a clear β nucleus outside the ambiguous region.
5. The α nucleus was not complete, however the region clearly has strong α potential.
6. This prediction is tenuous because the two Asp residues in succession would project out from both sides of β strand - forcing one towards the interior of the protein.
7. The region from Leu 150 to Lys 155 is ambiguous and has both α and β potential. The α assignment was based on the strong α nucleus between Phe 143 and Glu 148 ($\langle P_\alpha \rangle$ = 1.12, $\langle P_\beta \rangle$ = 0.94). One of the two contiguous N-terminal prolines was assigned to the helix and one to the N-terminal non-helical region more or less arbitrarily.
8. Each of these predictions seems acceptable, but the fact that they are immediately adjacent renders the predictions somewhat tenuous.
9. The region has considerable α potential but no α nucleus.
10. The adjacent Arg 243 - Lys 244 renders this prediction very tenuous.

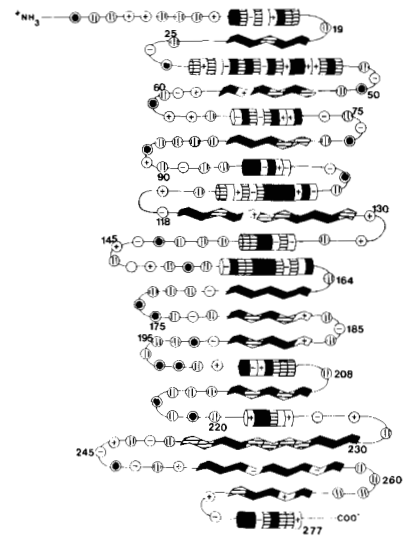


Fig 7. The predicted secondary structure of EcoRI endonuclease is represented schematically: Amino acid residues represented by a disk are part of a predicted α helix, those represented by irregular (folded) hexagons are part of a strand of β sheet, while residues which are not a part of a defined secondary structure are represented as circles. The shading scheme represents the chemical character of the amino acid residue: Heavily shaded residues are hydrophobic, hence usually found in the interior of a protein (i.e., Phe, Trp, Tyr, Ile, Leu, Met and Val). Charged residues, which are usually found on the exterior of the protein are blank, except for the sign of the charge (i.e., Glu and Asp are minus, Lys and Arg are plus while His is blank). Those residues which are commonly found in varied locations in proteins are represented by bars (i.e., Asn, Gln, Cys, Ser, Thr, Gly, Ala and Pro).

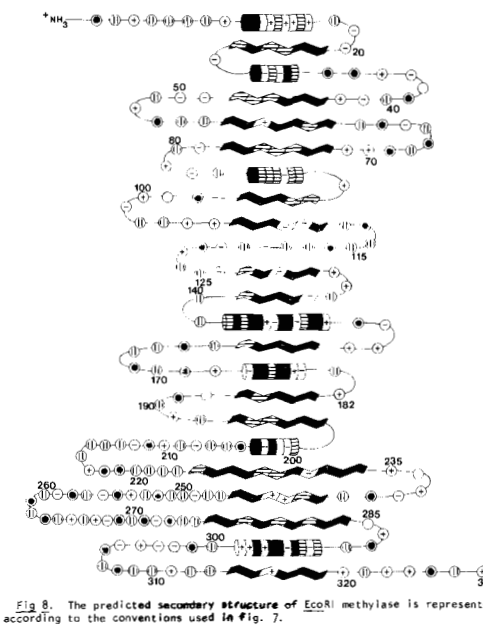


Fig 8. The predicted secondary structure of EcoRI methylase is represented schematically according to the conventions used in Fig. 7.

References

Note: References 1-45 appear after the parent paper.

46. Greene, P.J., Heyneker, H.L., Bolivar, F., Rodriguez, R.L., Betlach, M.C., Covarrubias, A.A., Backman, K., Russel, D.J., Tait, R., and Boyer, H.W. (1978) Nucleic Acids Res. 5, 2373-2380
47. Panet, A., van de Sande, S.H., Zoerwen, P.C., Khorana, H.G., Raae, A.J., Lillehaug, J.R., and Kleppe, K. (1973) Biochemistry 12, 5045-5050
48. Marinus, M.G. (1973) Mol. Gen. Genet. 127, 47-55
49. Maniatis, T. (1975) Biochemistry 14, 3787-3794
50. Bolivar, F., and Backman, K. (1979) Methods Enzymol. 68, 245-267
51. McDonell, M.W., Simon, M.N., and Studier, F.W. (1977) J. Mol. Biol. 110, 119-146
52. Rogers, S.G., and Weiss, B. (1980) Methods Enzymol. 65, 201-216
53. Smith, A.J.H. (1980) Methods Enzymol. 65, 560-580
54. Herrmann, R., Neugebauer, K., Pirkl, E., Zentgraf, H., and Schaller, H. (1980) Mol. Gen. Genet. 177, 231-242
55. Givol, D., and Porter, R.R. (1965) Biochem. J. 97, 32C
56. Aromatorio, D.K., Parker, J., and Brown, W.E. (1980) Anal. Biochem. 103, 350-358
57. Bailey, J.L. (1967) In Techniques in Protein Chemistry, 2nd Edition, pp. 340-352, Elsevier, Amsterdam
58. Greene, P.J., Poonian, M.S., Nussbaum, A.L., Tobias, L., Garfin, D.E., Boyer, H.W., and Goodman, H.M. (1975) J. Mol. Biol. 99, 237-261
59. Hershfield, V., Boyer, H.W., Chow, L., and Helinski, D. (1976) J. Bacteriol. 126, 447-453