71. F. Eisenberg, Jr., *J. Biol. Chem.* **242**, 1375 (1967); L. M. Hallcher and W. R. Sherman, *ibid.* **255**, 10896 (1980).
72. R. M. C. Dawson and N. G. Clarke, *Biochem. J.* **127**, 113 (1972); *ibid.* **134**, 59 (1973).
73. T. S. Ross and P. W. Majerus, *J. Biol. Chem.* **261**, 11119 (1986).
74. R. F. Irvine, A. J. Letcher, D. J. Lander, C. P. Downes, *Biochem. J.* **223**, 237 (1984); R. F. Irvine, E. E. Anggard, A. J. Letcher, C. P. Downes, *ibid.* **229**, 505 (1985).
75. I. R. Batty, S. R. Nahorski, R. F. Irvine, *ibid.* **232**, 211 (1985).
76. R. F. Irvine, A. J. Letcher, J. P. Heslop, M. J. Berridge, *Nature (London)* **320**, 631 (1986); C. A. Hansen, S. Mah, J. R. Williamson, *J. Biol. Chem.* **261**, 8100 (1986); P. W. Majerus *et al.*, unpublished observations.
77. G. M. Burgess, J. S. McKinney, R. F. Irvine, S. W. Putney, *Biochem. J.* **232**, 237 (1985); J. P. Heslop, R. F. Irvine, A. H. Tashjian, M. J. Berridge, *J. Exp. Biol.* **119**, 395 (1985); J. Turk, B. A. Wolf, M. L. McDaniel, *Biochem. J.* **237**, 259 (1986).
78. C. P. Downes, M. C. Mussat, R. H. Michell, *Biochem. J.* **203**, 169 (1982); M. J. Berridge *et al.*, *ibid.* **212**, 473 (1983); T. Sasaguri, M. Hirata, H. Kuriyama, *ibid.* **231**, 497 (1985); M. A. Seyfred, L. E. Farrell, W. W. Wells, *J. Biol. Chem.* **259**, 13204 (1984); R. S. Rana, M. C. Sekar, L. E. Hokin, M. J. MacDonald, *ibid.* **261**, 5237 (1986); S. K. Joseph and R. J. Williams, *Fed. Eur. Biochem. Soc. Lett.* **180**, 150 (1985); D. J. Storey, S. B. Shears, C. J. Kirk, R. H. Michell, *Nature (London)* **312**, 374 (1984); C. Erneux, A. Delvaux, C. Moreau, J. E. Dumont, *Biochem. Biophys. Res. Commun.* **134**, 351 (1986).
79. T. M. Connolly, T. E. Bross, P. W. Majerus, *J. Biol. Chem.* **260**, 7868 (1985).
80. T. M. Connolly, V. S. Bansal, R. F. Irvine, P. W. Majerus, *ibid.*, in press.
81. T. M. Connolly, W. J. Lawing, Jr., P. W. Majerus, *Cell* **46**, 951 (1986).
82. D. E. MacIntyre, A. McNicol, A. H. Drummond, *Fed. Eur. Biochem. Soc. Lett.* **180**, 160 (1985); S. E. Rittenhouse and J. P. Sasson, *J. Biol. Chem.* **260**, 8657 (1985); G. B. Zavoico, S. P. Halenda, R. I. Sha'afi, M. B. Feinstein, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 3859 (1985); S. P. Watson and E. G. Lapetina, *ibid.*, p. 2623; L. M. Molina y Vedia and E. G. Lapetina, *J. Biol. Chem.* **261**, 10493 (1986).
83. R. M. Lyons, N. Stanford, P. W. Majerus, *J. Clin. Invest.* **56**, 924 (1975); the actual molecular weight of the 40K protein is probably about 45,000 [R. Haslam and J. A. Lynham, *Biochem. Biophys. Res. Commun.* 77, 714 (1977)].
84. J. L. Daniel, H. Holmsen, R. S. Adelstein, *Thromb. Haemostasis* **38**, 984 (1977).
85. Y. Kawahara *et al.*, *Biochem. Biophys. Res. Commun.* **97**, 309 (1980); K. Sano, Y. Takai, J. Yamanishi, Y. Nishizuka, *J. Biol. Chem.* **258**, 2010 (1983).
86. A. J. R. Habenicht, *J. Biol. Chem.* **256**, 12329 (1981); I. G. Macara, *ibid.* **261**, 9321 (1986); L. J. Pike and A. Eakes, *ibid.*, in press.
87. H. Diringer and R. R. Friis, *Cancer Res.* **37**, 2979 (1977).
88. Y. Sugimoto, M. Whitman, L. C. Cantley, R. L. Erikson, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 2117 (1984); I. G. Macara, G. V. Marinetti, P. C. Balduzzi, *ibid.*, p. 2728.
89. S. Sugano and H. Hanafusa, *Mol. Cell Biol.* **5**, 2399 (1985); Y. Sugimoto and R. L. Erikson, *ibid.*, p. 3194 (1985); M. L. MacDonald, E. A. Keunzel, J. A. Glomset, E. W. Krebs, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 3993 (1985); M. J. Fry, A. Gebhardt, P. J. Parker, J. G. Foulkes, *EMBO J.* **4**, 3173 (1985).
90. S. Jackowski, C. W. Rettenmier, C. J. Sherr, C. V. Rock, *J. Biol. Chem.* **261**, 4978 (1986).
91. Supported by grants HLBI 14147 (Specialized Center for Research in Thrombosis), HL 16634, and Training Grant T32 HLBI 07088 from the National Institutes of Health; a NATO Research Fellowship; and a Fulbright Award (to H.D.). We thank L. J. Pike and J. E. Brown for their helpful suggestions concerning this article.

Research Articles

Structure of the DNA–Eco RI Endonuclease Recognition Complex at 3 Å Resolution

JUDITH A. MCCLARIN, CHRISTIN A. FREDERICK,* BI-CHENG WANG, PATRICIA GREENE, HERBERT W. BOYER, JOHN GRABLE, JOHN M. ROSENBERG†

The crystal structure of the complex between Eco RI endonuclease and the cognate oligonucleotide TCGC-GAATTCGCG provides a detailed example of the structural basis of sequence-specific DNA-protein interactions. The structure was determined, to 3 Å resolution, by the ISIR (iterative single isomorphous replacement) method with a platinum isomorphous derivative. The complex has twofold symmetry. Each subunit of the endonuclease is organized into an α/β domain consisting a five-stranded β sheet, α helices, and an extension, called the "arm," which wraps around the DNA. The large β sheet consists of antiparallel and parallel motifs that form the foundations for the loops and α helices responsible for DNA strand scission and sequence-specific recognition, respectively. The DNA cleavage site is located in a cleft that binds the DNA backbone in the vicinity of the scissile bond. Sequence specificity is mediated by 12 hydrogen bonds originating from α helical recognition modules. Arg^{200} forms two hydrogen bonds with guanine while Glu^{144} and Arg^{145} form four hydrogen bonds to adjacent adenine residues. These interactions discriminate the Eco RI hexanucleotide GAATTC from all other hexanucleotides because any base substitution would require rupture of at least one of these hydrogen bonds.

THE ABILITY OF A PROTEIN TO RECOGNIZE A SPECIFIC sequence of bases along a strand of double helical DNA lies at the heart of many fundamental biological processes. One of the most intriguing questions in molecular biology today is whether the details of these individual recognition mechanisms will form a small number of simple patterns that would lead to the development of a general recognition code.

This interest has stimulated crystallographic studies on many proteins that recognize specific sequences of DNA. The structures of four of these have been solved in the absence of DNA; these proteins are the Cro and CI repressors from coliphage λ, the *Escherichia coli* catabolite gene activator protein (CAP) and the tryptophan repressor (*1–6*). These four proteins share a common "helix-turn-helix motif" at the suggested DNA binding site, which has led to model building of the recognition complexes (*7–9*). In addition, the 7 Å structure of a co-crystalline complex between coliphage 434 repressor and a tetradecanucleotide containing its specific operator sequence supports the general features of these

J. A. McClarin, C. A. Frederick, J. Grable, and J. M. Rosenberg are in the Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA 15260; B.-C. Wang is in the Department of Crystallography, University of Pittsburgh, Pittsburgh, PA 15260; P. Greene and H. W. Boyer are in the Department of Biochemistry and Biophysics, University of California at San Francisco, San Francisco, CA 94143.

*Present address: Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139.
†To whom correspondence should be addressed.

models (*10*). These structures suggest that all five proteins are examples of one class of DNA recognition proteins.

The highly specific recognition of the double-stranded sequence d(GAATTC) by Eco RI endonuclease offers compelling advantages as a model system for investigating DNA recognition. It is a small (31,065 daltons) protein (276 amino acids) of known sequence (*11, 12*). The protein forms highly stable catalytically active dimers in solution and will form tetramers at higher protein concentrations (*13, 14*). The enzyme hydrolyzes the phosphodiester bond between the guanylic and adenylic acid residues resulting in a 5′-phosphate. The reaction proceeds with inversion of configuration at the reactive phosphorus (*15*), implying that there is an odd number of chemical events during the hydrolysis. The simplest interpretation of this observation is that the enzyme does not form a covalent intermediate with the DNA. Although Eco RI endonuclease requires Mg^{2+} for phosphodiester bond hydrolysis, it binds specifically to its cognate hexanucleotide in the absence of Mg^{2+} with a dissociation constant on the order of $10^{-11}M^{-1}$ (*16–19*).

In addition to the specific interaction of Eco RI endonuclease with the canonical sequence, the enzyme also binds DNA in a nonspecific manner that does not result in hydrolysis of the DNA (*16, 20, 21*). It has been postulated that the nonspecific complex enhances the rate of formation of the specific complex by facilitated diffusion along the DNA (*19, 22–24*).

Both the Eco RI endonuclease and the Eco RI methylase recognize the same hexanucleotide; however, the latter methylates the central adenine residues of both strands at the exocyclic N-6 amino group. When either one or both groups are methylated, the endonuclease no longer cleaves the DNA. Thus, Eco RI endonuclease not only discriminates between its hexanucleotide and all other hexanucleotides, it also discriminates between different methylation states of the same hexanucleotide.

A full understanding of sequence specificity requires cocrystals of DNA and protein that diffract to high resolution so that side chains can be visualized. We have obtained cocrystals of Eco RI endonuclease and the dodeca- and tridecanucleotides CGCGAATTCGCG and TCGCGAATTCGCG (Eco RI site underlined) (*25*). Results of the initial 3 Å electron density map of our tridecamer-endonuclease complex have been reported previously (*26*). Here we report the structure of the Eco RI endonuclease recognition complex including DNA-protein interactions that are involved in sequence specificity.

Structure determination. Crystallization conditions and methods of data collection were reported previously (*25, 26*). Hydrolysis of the DNA was prevented by omitting the required cofactor, Mg^{2+}, from the crystallization medium, and substituting EDTA. Platinum and mercury heavy atom derivatives were prepared as described (*26*).

A multiple isomorphous replacement (MIR) electron density map was calculated from these two derivatives to 5 Å resolution. The MIR phases were used to calculate a Pt–native difference Fourier map, which revealed the presence of a single minor heavy atom site. Wang's iterative single isomorphous replacement (ISIR) method (*27*) was independently applied to the platinum and mercury data. The general features, such as the solvent regions and the molecular outline, were similar in all three electron density maps; however the MIR and Hg-ISIR maps contained significant amounts of noise while the Pt-ISIR map was clear. We suspect that the noise in both cases is caused by a problem with the mercury derivative.

Statistics for the platinum derivative indicated a slight nonisomorphism at high resolution (*28*). We felt that the platinum phase information was dubious beyond 3.5 Å and therefore did not utilize it further.

The accuracy of the data and the absence of nonisomorphism are crucial to the success of the ISIR procedure. This can be seen by considering that the ISIR procedure resolves the phase ambiguity initially present in the SIR phases. If, however, both the probable phases are seriously in error for a significant fraction of the data, then the ISIR procedure most likely will converge to a false minimum. This is in contradistinction with the MIR case, where a preponderance of valid phase information can tend to overpower inaccuracies. For ISIR, the initial phase information should be carefully selected to ensure that it is accurate.

The positions and occupancies of the major and minor platinum sites were refined as follows. We calculated for each reflection within 5 Å resolution the absolute value of the difference between the native and derivative structure factors. We used all the differences for the centric data as well as the acentric reflections with larger differences, specifically the largest 40 percent of the acentric data. These data were used as input to a conventional full matrix least-squares refinement calculation with the platinum positions and occupancies being treated as variables (*29*). Statistics from the refinement are shown in Table 1.

The ISIR procedure was then used to resolve the phase ambiguity in the platinum SIR data to 3.5 Å; it was used again to extend the data to 3.2 Å and then 3.0 Å resolution, as reported earlier (*26*). The average figure of merit at the beginning of the process was 0.33 for those 4033 reflections which had both the native and the derivative information, and at the end of the process it was 0.79 for all 5880 observed reflections, including those 1847 reflections for which the derivative information had been rejected (Table 2). Although our earlier electron density map (*26*) based on these phases was very clear in most places and allowed us to trace the entire DNA double helix and much of the polypeptide backbone, especially in the areas with direct contact to the DNA, there were a few regions where the electron density was not easily interpretable.

A part of the data was missing from the original data sets, and we suspected that the absence of this information was interfering with

Table 1. Lattice parameters and heavy atom refinement statistics. V_M, is the Matthews' coefficient in cubic angstroms per dalton. Occupancy is the occupancy of heavy atom site. $\langle m \rangle$ is the mean figure of merit.

Lattice parameters
Unit cell: $a = b = 118.4$ Å; $c = 49.7$ Å; $\gamma = 120°$
Space group: P321
Asymmetric unit: one protein subunit and one DNA strand
Solvent content: 58 percent; $V_M = 2.8$

Heavy atom refinement statistics

Initial platinum sites

Occupancy	X	Y	Z	B
1.000	0.1258	0.5707	0.1109	25.0
0.383	0.4045	0.1329	0.3202	26.0

5.0 Å data: 1395 reflections; $R_C = 0.42$*; $\langle m \rangle = 0.41$
3.0 Å data: 5880 reflections; $R_C = 0.58$; $\langle m \rangle = 0.24$

QKREF refinement five cycles on all centric data + 40 percent largest acentric data within 5 Å. Initial $R = 0.33$; final $R = 0.30$.

Final platinum sites

Occupancy	X	Y	Z	B
0.996	0.1247	0.5695	0.1124	30.0
0.399	0.4066	0.1304	0.3138	35.0

5.0 Å data: 1395 reflections; $R_C = 0.41$; $\langle m \rangle = 0.42$
3.5 Å data: 4073 reflections; $R_C = 0.49$; $\langle m \rangle = 0.33$

*$R_C = \Sigma||F_{PH} \pm F_P| - F_{H(calc)}|/\Sigma|F_{PH} - F_P|$, where F_P and F_{PH} are the native and Pt-derivative structure factors, respectively and $F_{H(calc)}$ is the calculated contribution of the heavy atoms.

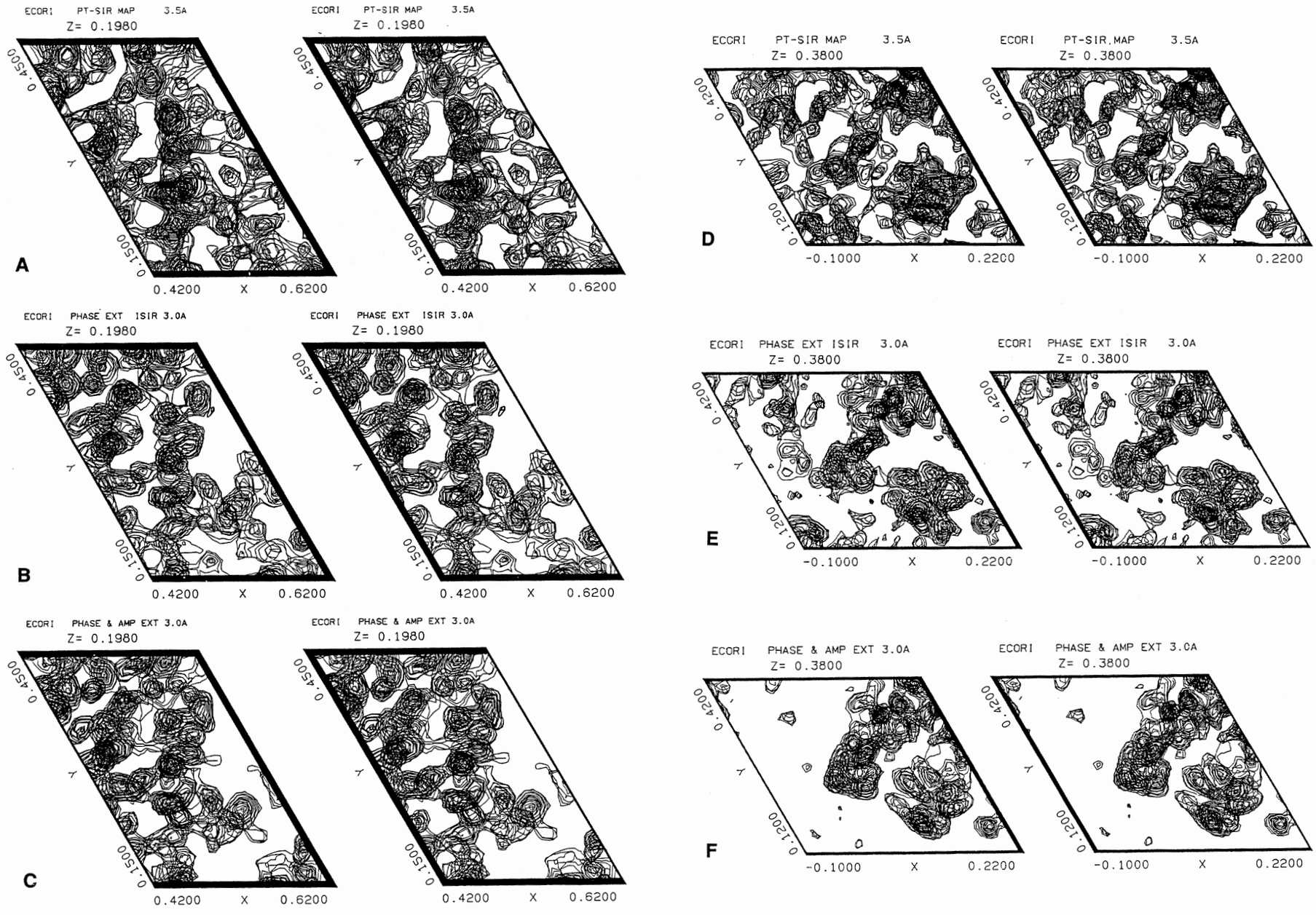


Fig. 1. (A) A stereo plot of the Pt-SIR electron density of a representative α helix. (B) The same view as in (A) in the initial Pt-ISIR electron density map. This is the α helix shown in figure 1 of Frederick *et al.* (*26*). (C) The same view as in (A) in the final Pt-ISIR electron density map. (D) The β hairpin section of the "arm" in the Pt-SIR electron density map of the DNA–Eco RI endonuclease complex. (E) The same view as in (D) in the initial Pt-ISIR electron density map that was the basis of our previous report (*26*). (F) The same view as in (D) in the final Pt-ISIR electron density map, which is the basis of this article.

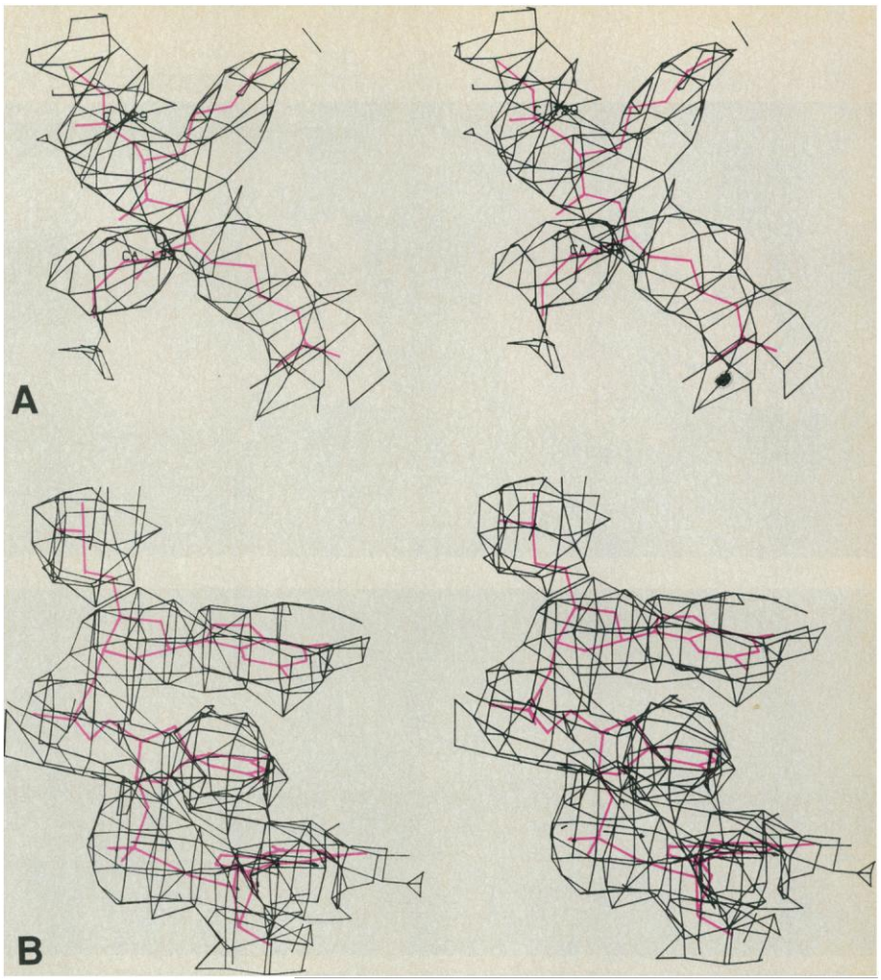


Fig. 2. (A) Residues 129 to 132, Gly-Lys-Arg-Gly, as currently fit. (B) DNA residues guanine 2, cytosine 3 and guanine 4, as currently fit. These pictures were drawn with the program FRODO.

the ISIR procedure. Three factors led to the absence of data: First, a few reflections at very low resolution were obscured by the beam stop of our Arndt-Wonacott camera. Second, a few reflections were saturated even on the third film of our film packs and were deleted from the data sets by the computer programs we used to process our film data (*30, 31*). Third, these programs also deleted a significant proportion of our weakly observed data because they were deemed statistically unreliable.

Efforts were then made to estimate the missing amplitudes and phases and to incorporate these estimates in the electron density calculations. The procedure was initiated for reflections within a 5 Å resolution limit (all the observed data to 3.0 Å were used during this process). Structure factor amplitudes and phases were estimated for the missing reflections by Fourier inversion of the modified electron density map. These estimates were used in subsequent iterations of the electron density calculations. After four iterations, a new solvent mask was calculated from both the 5880 originally observed reflections and the 293 estimates generated to this point. This entire procedure was similarly repeated in three additional stages to estimate the missing reflections to 4.0 Å, then to 3.5 Å and finally to 3.0 Å (Table 2). This process produced 2394 estimated structure factor amplitudes and phases.

At this stage, an electron density map was calculated from all the observed and estimated reflections (8274 in total). The map showed considerable improvement over the original; however, it still showed small ripples around some of the threefold axes. These were removed by a final set of iterations (filters 10 and 11 of Table 2) in which the solvent mask was calculated with a 10 Å radius in the masking function instead of the usual 5.1 Å radius. The electron density based on the final phases improved clarity (Fig. 1) and was used for the final chain tracing and fitting of the chemical sequence of the enzyme as described below.

We compared the electron density maps that preceded and followed both of the extension steps. In both cases the extensions reduced noise and improved the clarity of the maps while maintaining the basic features that were present in the initial 3.5 Å map. These features included the DNA (the phosphate positions were obvious features in all the maps), as well as several prominent α helices and strands of β sheet (Fig. 1). The improvements in detail were most noticeable in the problematic regions. These include the β hairpin which forms part of the "arm" (see below) and the region surrounding one of the threefold symmetry axes, which is very densely packed with protein. Some of the loops connecting secondary structure elements and some of the side chains were also clarified. These improvements enabled us to distinguish possibilities that had been ambiguous before the extension.

The final electron density map was displayed on plexiglass sheets. The DNA and protein secondary structure elements were very clear. More than two-thirds of the amino acid side chains were visible, and main chain density was visible for all but four amino acid residues. The missing residues were in the immediate vicinity of the major heavy atom site, and it appears likely that their movement is associated with the small non-isomorphism noted previously. Almost all of the side chains for tryptophan, phenylalanine, and tyrosine residues were clearly recognizable. Many basic residues, especially arginines, which were located at the DNA-protein inter-

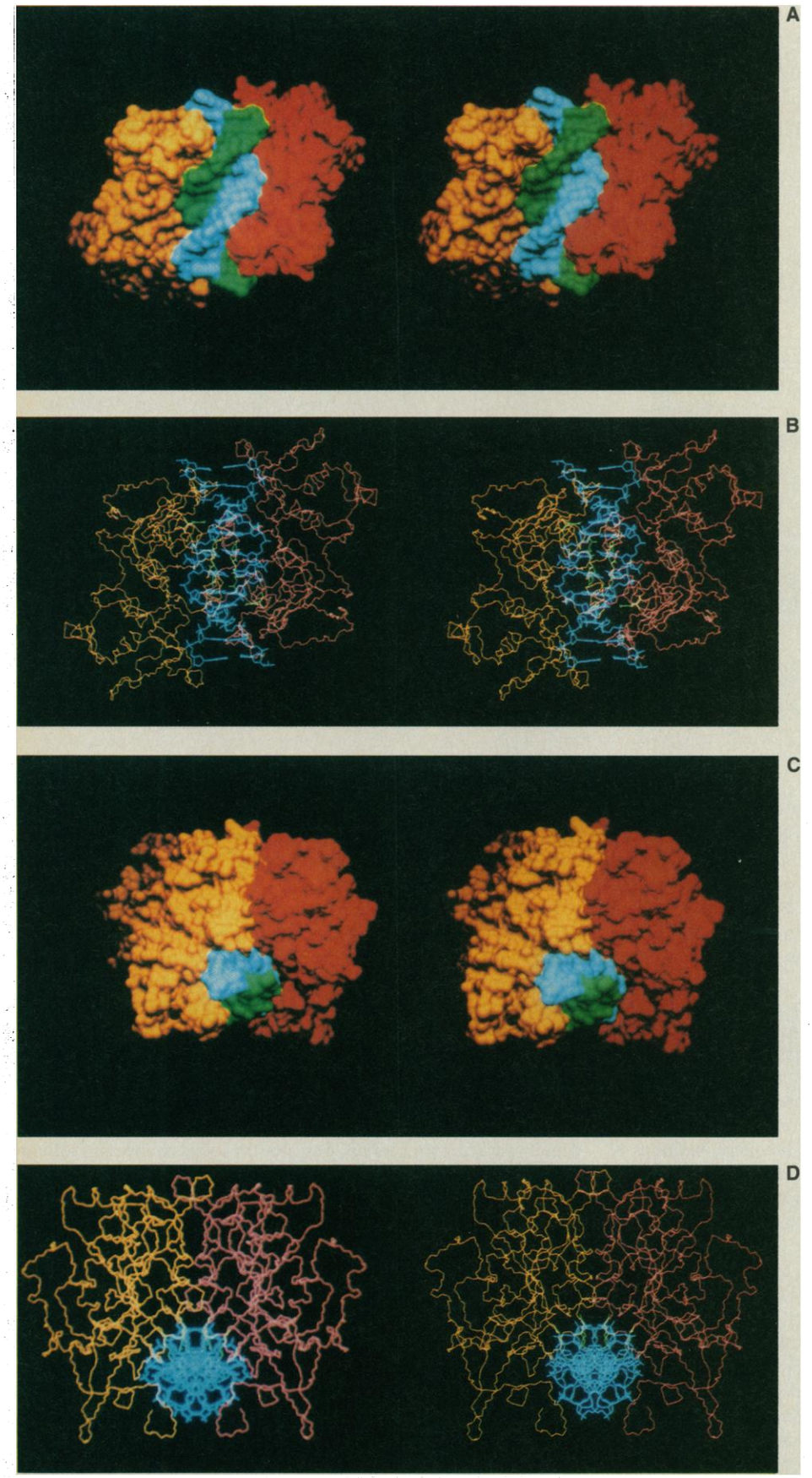


Fig. 3. Stereo drawings of the solvent-accessible surface of the Eco RI endonuclease–DNA complex are shown along with stereo drawings of the main chain atoms of the protein, the nonhydrogen atoms of the DNA and the amino acid side chains which participate in sequence-specific hydrogen bonding (Glu^{144}, Arg^{145}, and Arg^{200}). (A and B). The "front" view of the complex, which is a projection down a crystallographic twofold axis. (C and D). The "top" view of the complex, rotated 90° from that in (A) so as to view the structure down the *c*-axis which is also a view looking approximately down the average DNA helical axis. The raster computer graphics images of the solvent-accessible surfaces of the molecule were calculated with the programs AMS, and RAMS developed by Connolly *et al.* (*73, 74*), modified for use with the Evans and Sutherland PS340 raster graphics system (*75*).

face were also easily identifiable. Most of the poorly visualized side chains were located at the protein-solvent interface. Both the DNA-protein interface and the protein subunit-subunit interface were well ordered and provided useful constraints when we assigned the known amino acid sequence to the electron density map. These amino acid assignments were made via inspection of the electron density map, aided by model building, distance measurements, and the known stereochemistry of proteins. This process led to a tracing of the polypeptide chain through the protein-DNA complex.

Coordinates for an α carbon atom and for either a β carbon atom or a terminal side chain atom for larger amino acids were taken from the ISIR map on plexiglass sheets and used to generate atomic coordinates for the entire molecule with the program FRODO (*32, 33*). Electron density fitting continued with FRODO on an Evans and Sutherland PS340 computer graphics system. The coordinates were regularized to approximately ideal geometry alternately with improving the fit to the electron density. At present, the model has been fit to all of the electron density features noted above (Fig. 2). Refinement of the model should provide further accuracy.

General features of the complex. Both subunits of the enzyme form a globular structure with the DNA embedded in one side (Fig. 3). The complex as a whole is approximately 50 Å across. The major groove of the DNA is in intimate contact with the protein while the minor groove is clearly exposed to solvent. The complex has twofold symmetry, as expected from the symmetry of the recognition sequence. The molecular symmetry has also been incorporated into the crystal lattice. The protein has two projecting features, termed arms, that wrap around the DNA.

The DNA-protein complexes are packed within the crystalline lattice so that the DNA forms a continuous rod parallel to the *c*-axis. The unpaired 5′ thymine residues at each end of the double helix appear to be stacked on each other, leading to a continuous series of stacked bases across a crystallographic twofold axis. The oligonucleotide is actually somewhat larger than the DNA-binding face of the protein. However, its length closely matches the net width of the protein dimer, which tapers slightly at the DNA interface (Fig. 3). There is, therefore, a solvent gap at the binding face between the ends of the oligonucleotide and the protein dimer to which it is bound. The DNA-DNA interaction comprised a significant fraction of the net intermolecular interactions along the *c*-axis. This observation supports the concept that stability in DNA-protein cocrystals requires compatibility between the DNA-DNA, protein-protein, and protein-DNA contacts especially in the direction of the average DNA helix axis. Similar end-to-end packing of DNA was a salient feature of the 434 repressor-operator cocrystals (*34*). These results, combined with the recent success of Jordan *et al.* (*35*) in obtaining cocrystals of phage λ C1 repressor and operator by varying DNA length suggest that the length and terminal sequence of the cognate oligonucleotide should be treated as a critical variable in future attempts to form sequence specific DNA-protein cocrystals.

Three major areas of protein-protein interaction together with the DNA-DNA interaction, form the crystalline lattice. First, there is the subunit-subunit interface within the dimeric complex which contains the determinates of dimer formation. Second, there is the region around a threefold symmetry axis, where three dimers are tightly packed. Third is a smaller region of limited protein-protein interactions along the direction of the *c*-axis. These involve contacts between loops at the molecular surface of the protein dimer.

The DNA retains most of the structural features of the well-known double helix. In particular, Watson-Crick base pairing is maintained throughout the 12 paired bases. (The 5′ thymidylate residues do not participate in base pairing although they do have important base stacking interactions.) However, the DNA is kinked in the recognition complex, by which we mean that it departs

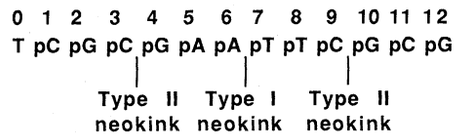


Fig. 4. The sequence of the tridecameric oligonucleotide used to make the DNA-protein complex. Also shown is the location of the kinks and the base numbering scheme, which was chosen to be consistent with the numbering system used by Dickerson and co-workers for the dodecamer (*45–47*); thus a given residue, for example, guanine 2, refers to the same residue in both the dodecamer and Eco RI complex.

significantly from the B conformation according to certain criteria (see below). These kinks appear to be stabilized by the binding of the protein. Our previous report (*26*) was primarily based on the location of the phosphate peaks, which are very prominent features of the initial electron density map. The electron density corresponding to the deoxyribose and base moieties showed significant improvement in the final ISIR electron density map and it is clear that these groups are also displaced from the positions they would occupy normally. Each kink distorts approximately two base pairs and the centers of the kinks are separated by three base pairs (Figs. 4 and 5).

The type I neokink. The most striking departure from B-DNA is centered on the crystallographic and molecular twofold axis, between adenine 6 and thymine 7 (Fig. 4). We refer to this feature as the "type I neokink." It represents a net rotation of the upper half of both strands of the DNA relative to the entire lower half of the double helix so as to unwind the DNA. The unwinding can be seen in the relative positions of phosphorus atoms 6 and 7, which show very little relative rotation about the average helix axis (they are at the center of Fig. 5). The unwinding is approximately 25° and would propagate through the DNA as a long-range effect on the net winding of the double helix. Kim and co-workers have measured the

Table 2. ISIR refinement statistics. Filter is the (sequential) number of the calculated solvent mask. Cycles is the number of cycles of solvent flattening and Fourier inversion with the use of the current solvent mask. Res. is the resolution (in Å) for the calculation; for phase refinement and extension this is resolution limit for all the data in the calculation while for amplitude extension it is the resolution limit for the generation of estimates for the unobserved reflections (all the observed data to 3.0 Å were used for the calculation). N_P, is the number of "paired" reflections for which both native and derivative data were available. N_{UP}, is the number of "unpaired" reflections for which only native data were available. N_G, is the number of estimates generated for unobserved data. Shift, is the mean phase shift from the SIR "best" phase. $\langle m \rangle$, is the mean figure of merit for all reflections based on the current ISIR phase probability distribution.

Filter	Cycles (No.)	Res. (Å)	N_P	N_{UP}	N_G	Shift	$\langle m \rangle$	R*
			Phase refinement					
1	4	3.5	4033	229		47.0	0.75	0.25
2	4	3.5	4033	229		47.2	0.75	0.23
3	8	3.5	4033	229		50.4	0.78	0.21
			Phase extension					
4	6	3.2	4033	1228		59.4	0.78	0.22
5	6	3.0	4033	1847		63.3	0.79	0.21
			Amplitude extension					
6	4	5.0	4033	1847	293		0.76	0.21
7	4	4.0	4033	1847	576		0.78	0.20
8	4	3.5	4033	1847	992		0.84	0.16
9	4	3.0	4033	1847	2394		0.87	0.15
10	6	3.0	4033	1847	2394	63.6	0.87	0.16
11	6	3.0	4033	1847	2394	63.8	0.87	0.15

*$R = \Sigma|F_{obs} - F_{calc}|/\Sigma F_{obs}$ where F_{obs} are the observed (native) structure factors and F_{calc} are the structure factors obtained from Fourier inversion.

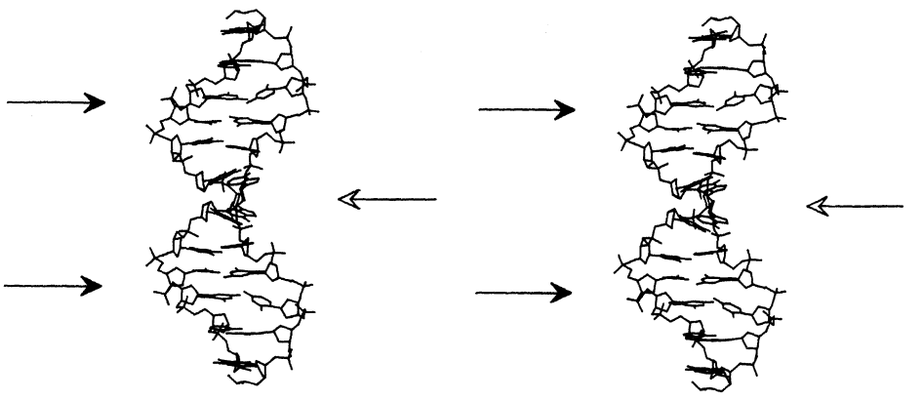


Fig. 5. A stereo figure of the DNA indicating the type I and type II neokinks. The single arrow on the right points to the center of the type I neokink. The twofold symmetry axis passes through both the arrow and the center of the type I neokink; hence the type I neokink has this symmetry. The two arrows on the left point to the centers of the type II neokinks; they are identical because of the twofold symmetry.

unwinding of DNA in solution when Eco RI endonuclease binds DNA in the absence of Mg^{2+}; they obtained an identical value (*36*).

The principal effect of the unwinding is that the major groove becomes wider. The phosphate-phosphate distances across the major groove are increased by approximately 3.5 Å. Interestingly, the base pairs do not significantly increase their interplanar separation although the base-base stacking contacts are clearly changed (Fig. 5). Thus, they type I neokink represents an effective mechanism for increasing the separation of the backbones of DNA strands without increasing the separation of the bases. The difference arises because a helix is a screw. Breaking the screw symmetry at one point of a helix and twisting one part with respect the other will alter the separation between the "threads" across the break. The increased backbone separation is essential because otherwise the recognition α-helices would not fit between and therefore could not approach closely enough to interact with the bases (see below). This consequence of the type I neokink suggests that it may be a general mechanism for facilitating access by proteins to the major groove of DNA. If so, similar DNA structures should be seen in some other recognition complexes.

There are also significant displacements of the A·T base pairs on either side of the kink center. These base displacements are critical to the recognition mechanism because they align adjacent adenine residues (5 and 6) within the recognition site (Fig. 4). These two purines are both involved in "bridging" interactions with amino acid side chains. These recognition interactions could not occur without the realignment because the N-6 moieties bridged by Glu^{144} and the N-7 moieties bridged by Arg^{145} would be too far apart if the DNA were in the B conformation.

Both the base pair realignment and the increased backbone separation are manifestations of a localized reduction in the twist of DNA; hence one could probably not exist without the other. However, the unwinding between adjacent phosphates appears to be localized at residues that are different from those where unwinding appears between adjacent base pairs. The phosphate unwinding is concentrated at the middle of the DNA (between phosphates of residues 6 and 7), whereas the base unwinding is displaced toward the adenines (residues 5 and 6); thus producing the realignment discussed above. Thymines 7 and 8, which are paired to displaced adenines, are also displaced.

The realignment of the base pairs reveals another aspect of the type I neokink that may be of general significance: namely, that it creates sites for multiple hydrogen bonds which are absent in B-DNA. Indeed, the idea that Eco RI endonuclease creates some of the detailed features on the surface of the DNA, which it then recognizes, is provocative and unexpected.

The type II neokink. The other localized departure from B-DNA, which we tentatively designate the type II neokink, is also highlighted in Fig. 5. The twofold symmetry of the recognition complex generates a duplicate of this feature (Fig. 4). The distortions are centered at phosphate moieties of guanine 4 and guanine 10. The backbone associated with nucleotides on either side of these phosphates is in an unusual conformation. For example, the distance between the phosphorus atoms associated with residues 4 (G) and 5 (A) is 7.3 Å, which is longer than expected for B-DNA. The distorted segment spans the scissile bond. Similarly, the phosphorus-phosphorus distance between residues 9 (C) and 10 (G) is 7.4 Å. The base pair immediately adjacent to the Eco RI hexanucleotide, that is, that involving cytosine 3 and the symmetry-related equivalent of guanine 10, is clearly anomalous. Its propeller twist appears exaggerated, and the pyrimidine is at an unusual angle in the electron density map (Fig. 2).

We determined the helical properties for the segments of DNA between the neokinks and between the type II neokink and the end of the DNA (*37*), using the coordinates we fit to the ISIR electron density map. We obtained results similar to the corresponding determinations based on the preliminary DNA model (*26*): The bend angle of the type II neokink is between 20° and 40°. However, the interpretation of this result is clouded because these calculations include nucleotides that are not in an exact helical conformation.

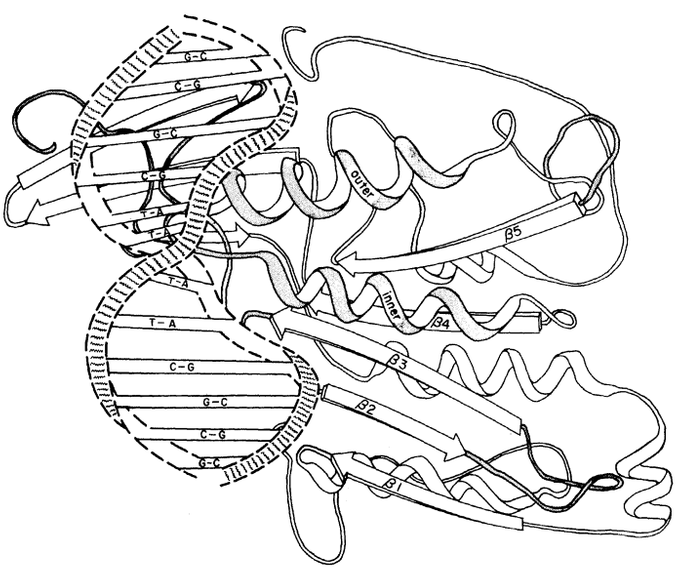


Fig. 6. Schematic backbone drawing of one subunit of (dimeric) Eco RI endonuclease and both strands of the DNA in the complex. The arrows represent β strands, the coils represent α helices, and the ribbons represent the DNA backbone. The helices in the foreground of the diagram are the inner and outer recognition helices. They connect the third β strand to the fourth and the fourth β strand to the fifth. The two helices also form the central interface interface with the other subunit. The amino terminus of the polypeptide chain is in the arm near the DNA.

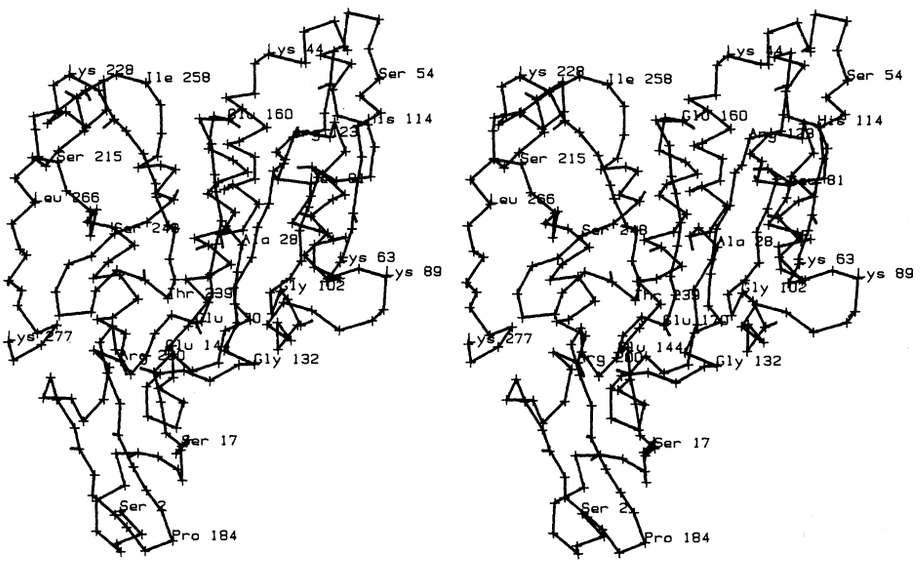


Fig. 7. A stereo drawing of the α carbon trace of one subunit of Eco RI endonuclease.

Highly refined coordinates (which are not yet available) are required to properly choose which to include in the calculations. Consequently, values for the bend angle of the type II neokink should be considered provisional. Unwinding can be more readily assessed by examining phosphorus positions in projection down the average helix axis, and the type II neokink does not introduce a major change in the net winding of the DNA.

General features of neokinks. We have based our usage of the term "kink" on some of the ideas originally introduced by Crick and Klug (*38*). Our concept invokes two criteria. (i) An abrupt, highly localized disruption of the overall double helical symmetry (screw or diad) and (ii) structural effects that propagate through the DNA over long distances. Since DNA is a relatively stiff rod, the simplest way to introduce long-range structural effects is to either bend or twist the double helix. Twisting of DNA can be readily detected in solution (*36, 39, 40*) as can bending (*41–44*), and these effects may be of functional significance. The term "kink" therefore refers to an abrupt disruption of the double helical symmetry, which includes a sharp bend, or a highly localized underwinding or overwinding of the DNA, or both.

We suspect that more "kinks" will be observed in crystal structures and that many of these kinks will combine both bending and twisting at the same locus. Indeed, close inspection of the type I neokink suggests that it could introduce a hinge into DNA; that is, the kink reported here might represent one member of a family of related structures with similar unwinding but different bending angles. A single term serves to focus attention on the critical features of "kinking"; namely, localized changes that generate long-range structural effects.

The prefix "neo-" in the term neokink indicates that the departure from B-DNA is induced by an external agent (the protein) and is not seen when DNA is studied in isolation. The oligonucleotide used in these cocrystals is virtually identical to that studied by Dickerson and colleagues (*45–47*). The structures they report do not contain dramatic kinks, such as the type I neokink. This suggests that the protein provides energy to drive the DNA into conformations that would otherwise be unfavorable and thus exist only transiently.

Other proteins distort DNA when they form complexes with it. Richmond *et al.* observed that the DNA in their 7 Å nucleosome structure contained "sharp bends" or possible kinks (or both) (*48*) which are not likely to be present in naked DNA; these could be neobends or neokinks, depending on the abruptness of the transition. Similarly, Anderson *et al.* reported that their 7 Å electron density map suggested that the 434 repressor introduced small perturbations into the structure of its operator (*10*). In vitro data suggest that the araC protein bends DNA to form functional complexes (*49, 50*); similarly, λ repressor can bend DNA molecules containing altered spacings between operator sites (*51*). Thus, protein-induced alterations of DNA structure appear to be common, and we suspect that additional neo-conformations will be observed as three-dimensional structural information becomes available on other DNA-protein complexes.

Structural organization of the protein. Each Eco RI endonuclease subunit is organized into a single domain consisting of a five-stranded β sheet surrounded on both sides by α helices (see Fig. 6). The domain is therefore of the well-known α/β architecture (*52*). Four of the five strands in the β sheet are parallel; however the location of the single antiparallel strand makes it possible to divide the sheet conceptually into parallel and antiparallel three-stranded motifs. The parallel motif (β3, β4, and β5 of Fig. 6) is the foundation for the direct contacts between the protein and DNA bases as well as subunit-subunit interaction, and the antiparallel motif (β1, β2, and β3 of Fig. 6) is the foundation for the site of DNA strand scission. We have also noted that the parallel motif is very similar to one-half of the well-known nucleotide binding domain (*53*), which is a six-stranded parallel β sheet, constructed out of two topologically identical three-stranded motifs.

The course of the polypeptide chain reveals the principal features of the protein structure (Figs. 6 and 7). The amino terminal section of the chain (residues 2 through 17) forms part of the "arm," which wraps around the DNA. The polypeptide chain passes along the surface of the molecule (residues 18 to 28). It then forms a long α helix on the surface of the molecule (residues 29 though 43), which is followed by a loop into the first strand of the β sheet (residues 44 through 56). The β sheet is formed sequentially starting from the outside of the antiparallel motif. The next loop (residues 63 through 102) connects the first and second β strands; it also contains another α helix situated on the surface of the molecule. The loop between the second and third antiparallel β strands (residues 110 through 122) projects somewhat into the solvent and is involved in the limited protein-protein interactions noted along the *c*-axis. The third β strand is a common element of both the antiparallel and parallel motifs. The overlap between the two motifs provides a means of structural interaction between the two regions of the enzyme which are responsible for DNA recognition and DNA strand scission activity.

The parallel motif is formed sequentially from the middle of the β sheet to the fifth strand at the edge of the sheet (residues 123 through 228). The α helices found at the inter-subunit interface are the crossover helices (*52*) of the parallel motif. The α helix connecting the third β strand to the fourth (residues 146 through 158) is called the "inner α helix" because it is part of an "inner recognition module," which is described below. Similarly, the α helix connecting the fourth β strand to the fifth (residues 201 through 209) is called the "outer α helix" (Fig. 6). After exiting the fifth β strand at residue 229, the polypeptide chains forms an extended loop around the surface of the complex, placing the carboxyl terminus in the proximity of the DNA backbone.

All of the major α helices in the protein are aligned so that their amino terminal ends are pointing in the general direction of the DNA. This orients the α helix dipoles so that they interact favorably with the electrostatic field generated by the negatively charged phosphates on the DNA backbone, thereby contributing to the net stability of the complex. Because of the alignment of the peptide bonds, an α helix has a net dipole moment, which can be approximated by placing one-half of a positive virtual charge at the amino terminus of the helix and one-half of a negative virtual charge at the carboxyl terminus (*54*). The inner and outer α helices from each subunit are oriented so that their amino terminal ends project into the major groove of the DNA. The amino acid side chains that interact with the DNA bases are located at the ends of these helices or in residues that immediately precede the helix.

Subunit-subunit interactions are primarily mediated by amino acid residues located in the parallel motif. The subunit-subunit interface can be subdivided into two general regions: a central portion, which is inaccessible to solvent, and a surface portion, which is solvent accessible. The central portion of the extensive interface includes interactions between residues within the two crossover α helices, that is, the inner and outer α helices. The NH_2-terminus of β-strand 5 is also part of the central interface. These interfacial residues have hydrophobic and neural polar side chains. The surface portion of the interface includes many salt links between subunits situated around the exterior edge of the interface. These are formed by charged residues in the turn preceding β-strand 5, residues in the carboxyl-terminal surface loop, and two residues from the surface of the antiparallel motif (all other residues in the interface are in the parallel motif). Charged residues at the subunit-subunit interface are also involved in the DNA-protein interface.

Eco RI endonuclease has arms that wrap around the DNA. The "arm" is an extension of the α/β domain (Fig. 7), which wraps around the DNA partially encircling it, thereby clamping it into place on the surface of the enzyme. Because of the twofold symmetry of the complex there are two arms, each of which interacts with the DNA directly across the double-stranded helix from a scissile bond. The arms contact the DNA at the type II neokinks and may be causative elements in the formation of these DNA structures. Each arm is composed of the amino terminus of the protein and a β hairpin sequentially located between the fourth and fifth strands of the large β sheet (residues 176 through 192) (*55*). Part of the amino terminal portion of the polypeptide chain (residues 17 through 20) adds a third β strand to the β hairpin, thereby forming a three-stranded antiparallel β sheet, which is the structural foundation of the arm. Thus, there are two β sheets in each Eco RI endonuclease subunit: the large five-stranded sheet described above and the smaller three-stranded sheet described here.

The first 14 amino acid residues of the polypeptide chain form an irregular structure, which is sandwiched in between the smaller β sheet and the DNA. The sandwiched region of the arm mediates several nonspecific DNA-protein contacts. Additional DNA backbone contacts are located in the short segment of polypeptide chain that connects the β hairpin with the outer α helix, which follows it in the primary sequence.

The endonuclease arms are conceptually similar to the arms that Pabo *et al.* propose to be part of the DNA–λ CI repressor complex (*56, 57*). In both cases, amino terminal sections of the polypeptide chain are involved. However, Pabo *et al.* suggest that the λ repressor arms contribute sequence specific contacts between the protein and DNA bases, while Eco RI endonuclease arms interact with the DNA backbone. The λ arms are very extended elements of polypeptide chain which could only be stabilized by association with DNA. In contrast, the Eco RI endonuclease arms are more substantial elements of structure, which could be intrinsically stable.

Jen-Jacobson *et al.* have obtained evidence that strongly suggests that the nonspecific contacts between the DNA and the amino terminal 14 residues within the arm are required for catalytic activity (*58*). This evidence consists of modified endonucleases produced by selective proteolytic removal of portions of the amino terminus from the DNA-endonuclease complex. Many of the resulting proteolytic derivatives retain sequence-specific DNA binding but lack strand scission capability. The data suggest that without critical contacts between the DNA and the arm, either the DNA backbone in the vicinity of the scissile bond may not be held in the correct orientation within the catalytic site or the type II neokinks are not correctly formed, or both.

DNA binding must be associated with a conformational change of the protein because the arms encircle the DNA to such an extent that it is unlikely that DNA could enter the active site in the absence of some movement. There are four general possibilities. (i) The arms may have two stable structures, one in the presence and one in the absence of DNA. We favor this possibility because the amino terminal 14 residues of the arms (which are sandwiched between the β hairpin and the DNA) appear to be rather loosely associated with the β hairpin, suggesting that these residues fold against the DNA when it is present and refold in a tighter association with the protein when DNA is absent. (ii) Part of the arms (probably that consisting of the amino terminal 14 residues) may undergo an order-disorder transition in which they are disordered in the absence of DNA and condense on it during complexation. (iii) The arms may be relatively rigid structures that are attached to the main part of the molecule by flexible hinges. (iv) The dimeric endonuclease could undergo a quaternary conformational change in which each subunit moves with respect to the other subunit. These possibilities are not all mutually exclusive and the actual changes could involve a combination of several of these factors. We have grown crystals of the protein in the absence of DNA, but that structure determination has not yet been completed.

Catalytic clefts in the enzyme. The two DNA backbone segments that face toward the major portion of the endonuclease are buried in clefts in the protein. These segments include the scissile bonds. Both DNA backbone segments and the corresponding clefts are identical because of the twofold symmetry of the complex. The carboxyl edge of the antiparallel segment of the β sheet forms the base of the cleft which binds phosphates 3, 4, and 5 (Fig. 4). (The scissile bond is at the fifth phosphate.) One side of the catalytic cleft is formed by the loops which interconnect the β strands in the antiparallel motif and that connect β-strand 3 to the inner α helix. The scissile bond is facing this side of the cleft. The other side of the cleft is formed by the inner and outer α helices from the other subunit. The cleft surface contains many basic amino acid residues that interact electrostatically with the DNA phosphates, contributing to the binding energy.

It has been shown for some time that Mg^{2+} can be added to preformed Eco RI endonuclease–DNA complexes in solution, which are then activated for cleavage (*13*), that is, the order of addition can

be first DNA and then Mg^{2+}. We therefore diffused Mg^{2+} into the cocrystals and found that the hydrolytic reaction was carried out in the crystalline state (*59*). This demonstrates the catalytic competence of the crystalline DNA-protein complex. The Mg^{2+}-treated crystals survive the structural transitions and they still diffract x-rays. The structure of the enzyme-product complex is not yet known.

The active site for DNA strand cleavage is not fully assembled in our structure. There is a solvent channel, with DNA backbone on one side and protein on the other, ending at the scissile bond. It is through this solvent channel that magnesium probably enters the active site. We presume that the structure in this region rearranges after magnesium is bound, forming a functional active site. In other words, in the absence of Mg^{2+}, the complex is analogous to an inactive zymogen that is activated by a structural isomerization triggered by the cation. This temporal order is probably important in the function of the endonuclease (see below).

DNA backbone–protein interactions. There appear to be interactions between the protein and the backbone of the DNA from residues 2 through 9 (Fig. 4). Phosphate moieties from residues 3, 4, and 7 are buried in the protein and are inaccessible to solvent. Phosphate and deoxyribose residues 3, 4, and 5 on each strand line the sides of the recognition hexanucleotide major groove, which is expanded by the type I neokink. These phosphates are bound within the catalytic clefts in the protein. Electrostatic interactions are also formed between the arms of the protein and phosphates from residues 8 and 9.

The two symmetrically related clefts, one in each subunit, are approximately 3.5 Å farther apart than the normal separation between the DNA backbones across the major groove of B-DNA. The increased separation, coupled with the basic residues within the clefts, probably produces an electrostatic field, which would tend to drive the DNA backbones apart. This could be a major factor promoting the formation of the type I neokink.

Phosphates 3 and 4, which flank the type II neokink, are not only buried in the cleft but interact with several basic amino acid residues. This strong interaction could be associated with a requirement to precisely position the scissile bond in the active site of the enzyme.

Ethylation interference experiments showed the largest effects at phosphate moieties of residues 3, 4, and 7 (*60*), which match the phosphates that are buried in the protein and protected from solvent. The next largest ethylation effect is observed for the reactive phosphate at the fifth position. Small effects are noted for the sixth phosphate, which is probably forming interactions to the protein even though it is partially exposed to the solvent. (We suspect that a stronger ethylation interference would have been observed at lower protein concentrations where the equilibrium is sensitive to smaller reductions in the protein-DNA association constant.)

The oligonucleotide used in our cocrystal is long enough to include all of the major contacts that form between long DNA substrates and the endonuclease. The association constant measured for the dodecamer CGCGAATTCGCG is within experimental error of that measured for plasmid DNA (*12, 14, 19, 61*). The unusually large Michaelis constant for an octanucleotide substrate as compared with dodecameric or larger substrates (*14, 62*) suggests that interactions between the enzyme and the flanking regions of the DNA backbone make significant contributions to the net stability of the complex.

The loop connecting β-strand 3 with the inner α helix (residues 131–143) appears to have a pivotal role in facilitating structural communication between regions of the complex. Part of this loop is involved in the formation of the cleavage site, as indicated above. However, its three-dimensional neighbors include many vital components of the complex. This loop is simultaneously adjacent to the arm, close to the amino acid residues involved in the direct recognition; the region around residue 140 is also packed against phosphate 7, which is the center of the type I neokink. The structure of these residues could be directly influenced by (i) the formation of direct hydrogen bonds to bases, (ii) formation of the type I neokink, (iii) the conformation of the arm, and (iv) the conformation of the cleavage site. Therefore the 131–143 loop could transmit structural information between these sites thereby serving to facilitate a temporal ordering of events within the overall catalytic cycle.

The recognition mechanism. The DNA-protein interface can be viewed in two portions: An extensive interface between the protein and the backbone of the DNA (already discussed) and a protein-base interface that partially covers the major groove of the recognition hexanucleotide (GAATTC). The minor groove is open to the solvent. The protein-backbone interface spans more nucleotide residues than the protein-base interface; that is, the protein interacts with the phosphate and deoxyribose moieties from nucleotides adjacent to the canonical hexanucleotide.

Hydrogen bonds between amino acid side chains (Glu^{144}, Arg^{145}, and Arg^{200}) and the purine bases of the canonical hexanucleotide constitute the direct, sequence-specific DNA-protein interactions in the complex. The bases and amino acid side chains must be precisely positioned relative to each other so that the interactions between them can generate the correct specificity. The α helical motifs form critical structural elements that facilitate the establishment of that spatial juxtaposition. Different recognition α helices provide the structural foundation for the interaction with different sections of the canonical hexanucleotide GAATTC. An inner module recognized the inner tetranucleotide AATT, and two symmetry-related outer modules recognize the outer G · C base pairs.

The interaction involving the outer module is relatively simple (Figs. 8 and 9): The guanidinium moiety of Arg^{200} forms two hydrogen bonds with the guanine base in an interaction designated

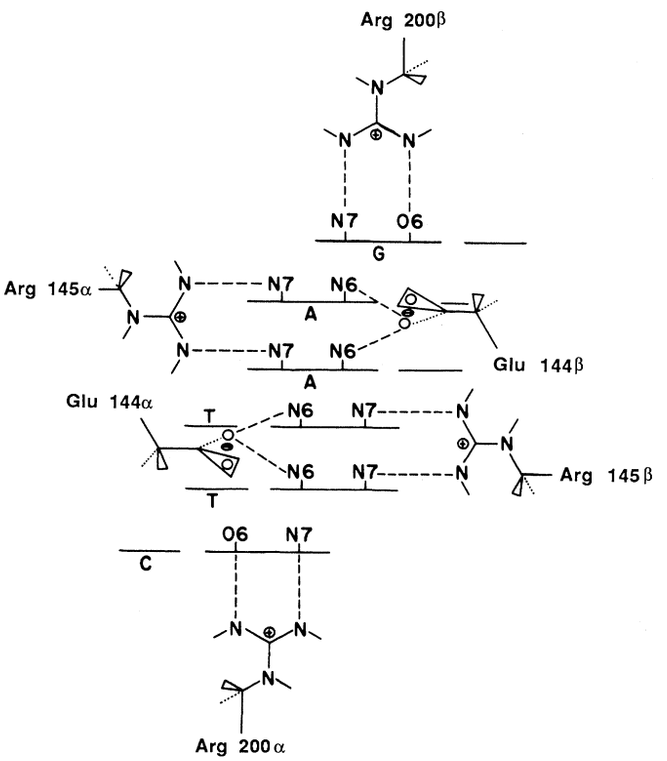
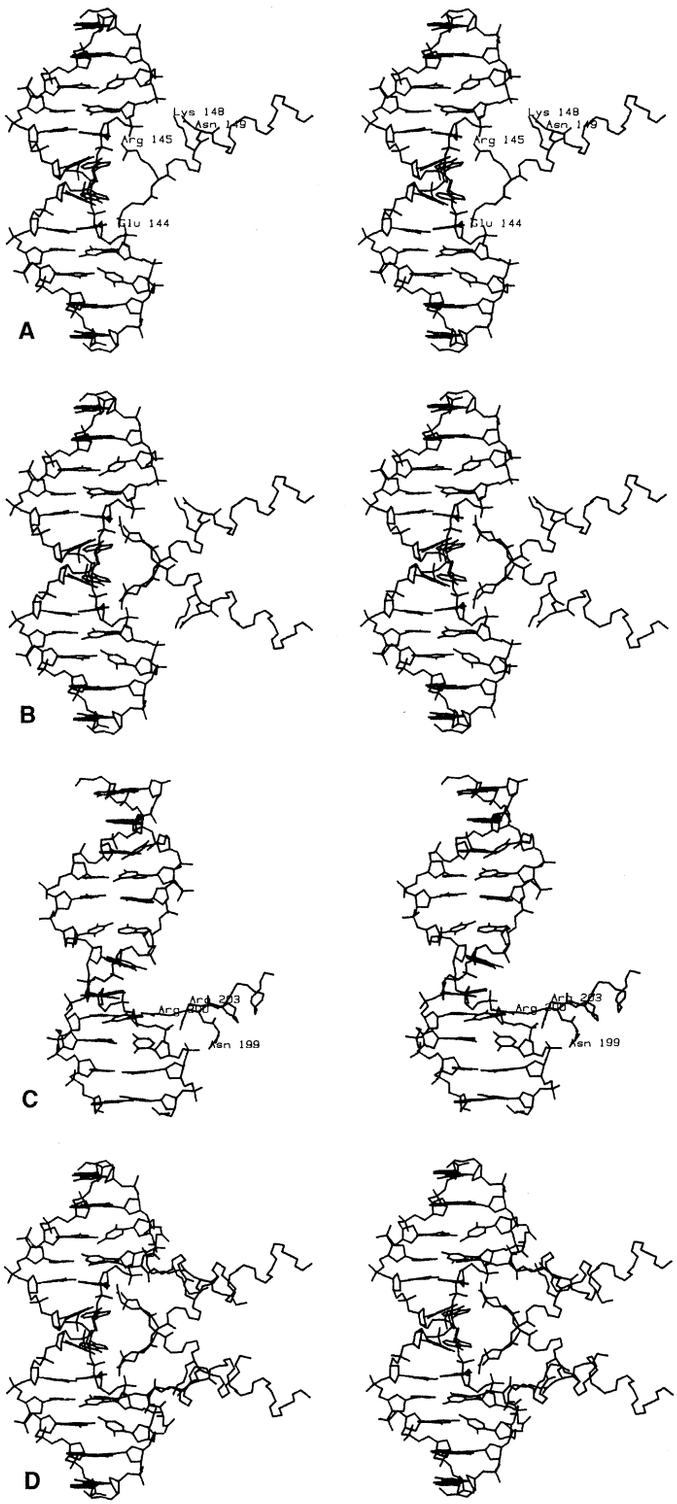


Fig. 8. A schematic representation of the recognition interactions and the 12 hydrogen bonds that determine the specificity of Eco RI endonuclease. Here, α and β refer to the two identical subunits of the enzyme. The positions of the bases and amino acid side chains have been shifted from the current model as shown in Fig. 9 in the interests of clarity.

Arg::G. One hydrogen bond is donated by Arg^{200} to the guanine N-7 atom and another is donated to the O-6 atom (see Fig. 10 for the numbering system). The Arg::G interaction was predicted by Seeman, Rosenberg, and Rich (*63*).

The inner module forms a more complicated set of interactions in that pairs of amino acid side chains interact with pairs of adjacent adenine residues. Each pair of adjacent adenines interacts with one amino acid from each subunit, (Fig. 8). These residues are Glu^{144} and Arg^{145} and the interaction is designated Glu-Arg::AA.



The side chain of Glu^{144} receives two hydrogen bonds from the adenine N-6 amino groups. In our current model, both hydrogen bonds are to the same carboxyl oxygen atom. (Each carboxyl oxygen atom can receive two hydrogen bonds.) The second oxygen atom may be interacting with residues Arg^{200} or Arg^{203} (or both) of the outer module via water bridges. Arg^{145} donates two hydrogen bonds to the adenine N-7 atoms. Thus, recognition in the inner tetranucleotide is based on "bridging" interactions in which amino acid side chains interact with two adjacent bases.

The recognition α helices illustrate a principle of "positioning," that refers to all the structural factors responsible for the precise three-dimensional juxtaposition of the correct elements of the protein and the DNA. For example, Arg^{145} recognizes features of the DNA that are different from those recognized by Arg^{200}, in part because the two amino acid side chains are positioned differently with respect to the DNA. The physical locations of the inner and outer α helices are the most important determinants of the positions of these amino acids.

Another central principle is one of discrimination. The central function of any sequence specific protein is its ability to discriminate its cognate sequence from the vast excess of noncognate DNA sequences in which it is embedded. Hence, any putative structural model of a sequence-specific interaction between DNA and protein must provide a satisfactory answer to the question of what happens when noncognate bases are present in the binding site of the protein. Position and discrimination are central issues in the following discussion of the major groove DNA-protein interface of Eco RI endonuclease.

The inner α helix is one of two single α helices that form recognition motifs. It is oriented so that its amino terminal end points toward the major groove of the DNA. The inner α helix makes an angle of approximately 60° with the average DNA helix axis (Fig. 9A). The polypeptide chain turns sharply at the end of the α helix so that residues at the amino terminal end of the helix and those in the bend are in close proximity to the DNA. The amino terminus of the inner α helix is also adjacent to the molecular twofold axis, and therefore it is in close proximity to the amino terminus of the symmetry-related helix from the other subunit. This symmetric pair of helices together form the inner module (Fig. 9B).

Both of the outer modules consist of a single α helix, namely the outer α helix, which connects the fourth and fifth strands of the large β sheet (Fig. 9C). The inner and outer α helices are positioned somewhat differently with respect to the DNA; the helix axis of the inner α helix almost intersects the average DNA helix axis while the axis of the outer α helix passes well to the outside (Fig. 9D). This positional difference is important because arginine side chains from the two helices have different recognition roles. They determine different specificities because they are positioned differently with respect to the DNA bases.

Fig. 9. Stereo drawings showing the recognition α helices and modules. (A) The "inner" α helix, which is part of the inner recognition module. The inner helix is also a crossover helix, connecting the third and fourth strands of the β sheet. Glu^{144} interacts with adenine residues in the lower half of the DNA and Arg^{145} interacts with adenine residues in the upper half. Lys^{148} and Asn^{149} interact with the phosphate moiety from guanine 4. (B) The inner recognition module, consisting of the inner α helices from both subunits. The inner module determines the specificity in the inner tetranucleotide; AATT. (C) The "outer" α helix, which is also one of the two identical outer recognition modules. The outer helix connects the fourth and fifth β strands. Arg^{200} interacts with guanine. In the other views shown here, the twofold symmetry axis is in the plane of the drawing; however, this view has been rotated approximately 20° for clarity. Asn^{199} interacts with the phosphate moiety from cytosine 3′ (C3 on the opposite strand), while Arg^{203} interacts with phosphate moieties from cytosine 3′ and guanine 4′. (D) The four-helix bundle consisting of the inner and outer α helices from both subunits.

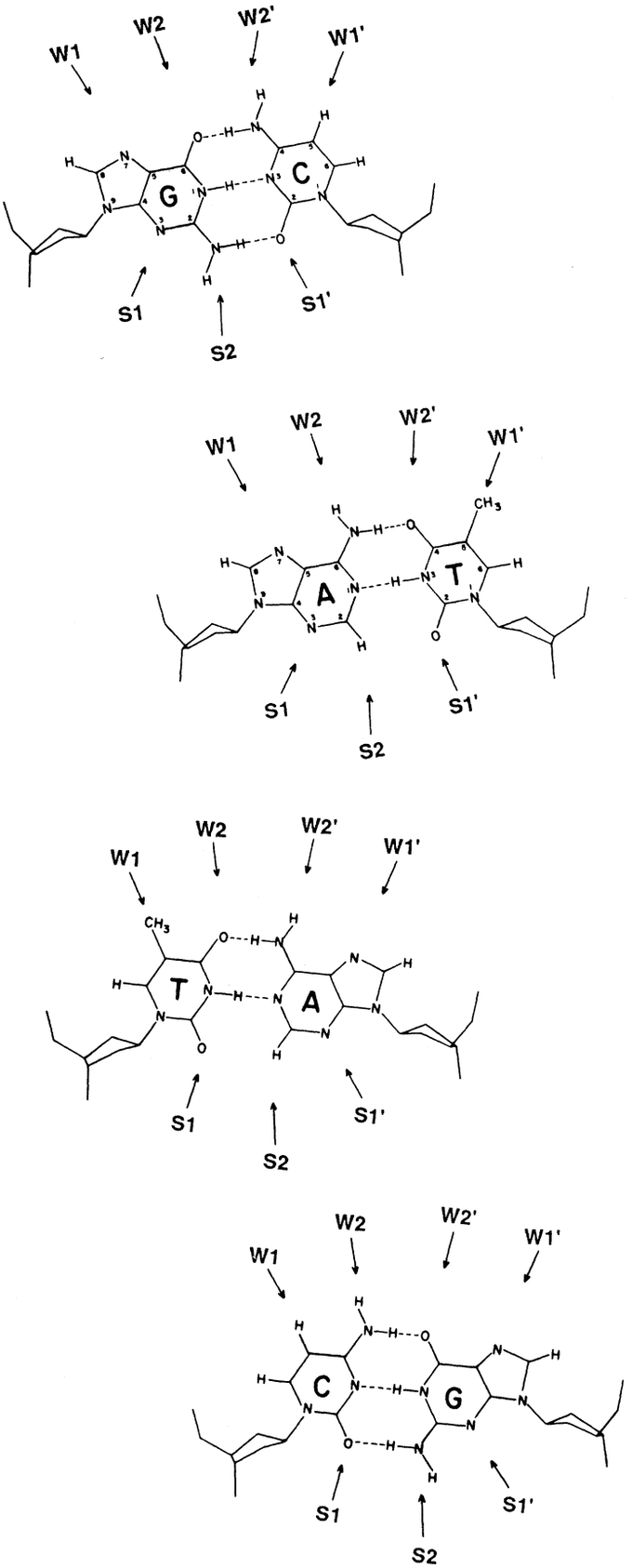


Fig. 10. The four base pairs showing the numbering scheme and the specificity sites based on those proposed by Seeman, Rosenberg, and Rich (*63*). Sites W2 and W3 of Seeman, Rosenberg, and Rich, which are 1 Å apart, have been merged in this treatment and are shown here as W2. Similarly, W2′ and W3′ of Seeman, Rosenberg, and Rich are combined into W2′ here (see text).

Eco RI endonuclease therefore repeatedly utilizes a simple structural motif, the α helix, to interact with DNA. These structural motifs are parts of larger topological motifs which are also repeated; both are part of β-α-β units. Thus, simple repeated motifs form the structural foundation of the recognition interactions.

All four helices form a parallel helix bundle (Fig. 9D), which is stabilized by interactions between the side chains of the individual helices. This parallel α-helical bundle is significantly different from the common four-helical motif referred to as an "up and down" or antiparallel helix bundle by Richardson (*52*). The antiparallel bundle has been observed in proteins such as the tobacco mosaic virus coat protein and myohemerythrin. The antiparallel architecture produces an internally favorable interaction between the electric dipoles associated with each α helix. By contrast, there is an internal energetic penalty associated with the parallel arrangement of the helices in Eco RI endonuclease. However the parallel arrangement produces an electrostatic field that facilitates the DNA-protein interaction. Another point of comparison is that the antiparallel bundle is formed by a contiguous stretch of polypeptide chain with turns connecting the helices. It is therefore a stable domain that can and does constitute the bulk of a globular protein. The parallel helix bundle must, of necessity, be part of a larger structural unit since additional elements of secondary structure are needed to connect the ends of the helices.

The inner and outer α helices determine the positions of key amino acid side chains with respect to the bases; the placement of the α helices with respect to the DNA is determined in part by side-chain interactions between the helix bundle and the DNA backbone (Fig. 9). The α helices are also packed against the β sheets of their respective subunits, thus firmly fixing the location of the entire four-helix bundle with respect to the protein as a whole. Thus, all the interactions between the protein and the DNA backbone indirectly serve to locate the helix bundle with respect to the DNA.

The Eco RI endonuclease recognition motif is clearly different from that of the helix-turn-helix DNA-binding proteins, which include the Cro and λ CI repressors, 434 repressor, CAP, and the tryptophan repressor (*1–6, 10*). The difference may be due to the high specificity demanded of a restriction enzyme or to the highly concentrated nature of the Eco RI recognition hexanucleotide. However, the difference shows that the helix-turn-helix motif is not a universal DNA recognition element.

The available data show that the unitary α helix is a general recognition motif that is found in Eco RI endonuclease and the five binding proteins. Within the binding proteins, the amino acid side chains that interact directly with the bases are assigned to only one helix of the helix-turn-helix motif, specifically the second α helix. This assignment was noted in the early speculative models (*7–9*) as well as the low resolution 434 repressor-DNA structure (*10*), and it has been supported by genetic data from "helix-swap" experiments, in which mutations are introduced into the part of the gene that codes for the second α helix of a particular repressor (*64, 65*). The resulting repressors recognize altered operator sequences; therefore only one of the two helices is actually discriminating between different base sequences. In other words, in the five binding proteins, the actual recognition motif is a unitary α helix.

In all of the models, as well as the low resolution repressor-DNA cocrystal, the first α helix within helix-turn-helix has the role of forming the foundation for amino acid side chains that interact with the DNA backbone. The first α helix thereby helps to fix the position of the recognition helix with respect to the cognate bases. Here too, it seems that there are not enough interactions between the recognition helix itself and the DNA backbone to position the recognition helix with sufficient precision. The two-helix motif has been found in proteins (or domains) that are much smaller than Eco

RI endonuclease and that do not have a large number of points on their surfaces that interact with the DNA backbone. It is likely that the helix-turn-helix motif represents an efficient way to combine a recognition helix with additional DNA backbone interactions so as to precisely position the recognition helix in a small protein. The structural conservation within the known examples of the two-helix motif probably reflects a particularly firm relative positioning of two α helices with respect to each other, which would be necessary for this function.

Twelve hydrogen bonds provide sequence specificity. There are two protein-base hydrogen bonds associated with each of the two Arg::G interactions of the two outer modules and eight hydrogen bonds associated with the inner module; four from each of the two Glu-Arg::AA interactions. This gives a total of 12 hydrogen bonds between protein and bases.

It is vital to determine whether or not the 12 hydrogen bonds discriminate between the Eco RI site and all other possible hexanucleotides. The following discussion shows that they do because substitution of any noncognate base pair would rupture one or more hydrogen bonds. A secondary purpose of that discussion is to develop a systematic method for analyzing the sequence specificity of particular protein-DNA interactions, which is based on the ideas of Seeman *et al.* (*63*). They showed that there were four principal interaction sites on the major groove side of a base pair and three sites on the minor groove side (*66*). These specificity sites (Fig. 10) make it possible to specify a template for the DNA-protein interface that can be used to analyze the match between the protein and alternative DNA sequences.

The templates can be specified and analyzed systematically for all possible combinations of base pairs (*67*) (Tables 3 to 5). Consider the Arg::G interaction. Guanine has hydrogen bond acceptors in W1 and W2, which are matched by corresponding hydrogen bond donors on Arg^{200}. No other base pair has hydrogen bond acceptors in both W1 and W2 (Table 3). The analysis for the Glu-Arg::AA interaction is similar. Both adenine residues have hydrogen bond acceptors in W1 and hydrogen bond donors in W2; these are matched by the donors and acceptors on Arg^{145} and Glu^{144}, respectively. None of the other bases match this pattern (Table 3).

Any attempt to provide a recognition mechanism for Eco RI endonuclease must also account for the fact that the extraordinarily high cleavage specificity under physiological conditions can be relaxed by simple buffer conditions. In the altered conditions, Eco RI endonuclease recognizes many nucleotide sequences that differ from the canonical site, GAATTC, at one or more base pairs (*68, 69*). The altered buffer conditions include elevated *p*H (8 to 9.5), substituting Mn^{2+} for Mg^{2+}, low ionic strength, and the addition of organic compounds such as glycerol or ethylene glycol. The modified sequences, termed Eco RI* sites, are cleaved at variable rates which can be summarized by the simple hierarchical rules: G >> A > T >> C at the first position (that is, GAATTC is cleaved much faster than AAATTC, which is cleaved slightly faster than TAATTC, which in turn is cleaved much faster than CAATTC). Similarly the hierarchy at the second and third positions is A >> [G, C] >> T (*70*).

The 12 hydrogen bonds are also consistent with these Eco RI* hierarchies if it is assumed that the protein adjusts its structure in order to maintain as many of the protein-base hydrogen bonds as possible. For example, if adenine were substituted for guanine at the first position, at least one hydrogen bond would be ruptured (the one in W2), as we have seen. Similarly, thymine could form, at most, one hydrogen bond with Arg^{200} because thymine has a methyl group in W1 (and an acceptor in W2). Thus, the Eco RI* sequences AAATTC and TAATTC could form, at most, 11 protein-base hydrogen bonds (one at the first position and two at each of the succeeding five canonical positions). Cytosine could not form any hydrogen bonds with Arg^{200} because it does not have any hydrogen bond acceptors in the major groove. Thus, the sequence CAATTC could form only ten sequence-specific hydrogen bonds. If the bases in the first position of these hexanucleotides are ordered by the total number of protein-base hydrogen bonds, we obtain the sequence G, (A, T), C: the observed Eco RI* hierarchy.

The observed Eco RI* hierarchy is also obtained at the position of the second base; GGATTC has 11 possible protein-base hydrogen bonds with the loss of the hydrogen bond in W2 where a hydrogen

Table 3. Base-pair specificity sites. The contents of the specificity sites on common base pairs are listed. Methylated bases are included as well as uracil for RNA: ^{me}A, N^6-methyladenine; ^{me}C, 5-methlycytosine. W1, W2, and so forth are the sites shown in Fig. 10. A, a hydrogen bond acceptor, is present on the indicated base pair at the indicated site; D, a hydrogen bond donor, is present; M, a methyl group, is present; H, a hydrogen atom (C-H), is present.

Base pair	Contents of site						
	W1	W2	W2′	W1′	S1	S2	S1′
A · T	A	D	A	M	A	H	A
G · C	A	A	D	H	A	D	A
C · G	H	D	A	A	A	D	A
T · A	M	A	D	A	A	H	A
^{me}A · T	A	M	A	M	A	H	A
T · ^{me}A	M	A	M	A	A	H	A
G · ^{me}C	A	A	D	M	A	D	A
^{me}C · G	M	D	A	A	A	D	A
A · U	A	D	A	H	A	H	A
U · A	H	A	D	A	A	H	A

Table 4. Base pairs recognized by single interactions. As can be seen in Table 3, any given site is generally occupied by the same functional group on more than one base pair. Thus, as noted by Seeman, Rosenberg, and Rich (*63*), a single protein-base interaction would lead to recognition of a degenerate set of base pairs. The base pairs recognized by such singular interactions are listed.

Site	Occupied by*	Symbol†	Base pairs recognized
W1	D_p	Pu	A · T, G · T, (^{me}A · T, G · ^{me}C, A · U)
W1	V_o	Me	T · A, (^{me}C · G, T · ^{me}A)
W1	V_i	CU	C · G, (U · A)
W2	D_p	Gt	G · C, T · A, (G · ^{me}C, T · ^{me}A, U · A)
W2	A_p	Ac	A · T, C · G, (^{me}C · G, A · U)
W2	V_o	MA	(^{me}A · T)
W2′	D_p	Ac′	A · T, C · G, (^{me}A · T, ^{me}C · G, A · U)
W2′	A_p	Gt′	G · C, T · A, (G · ^{me}C, U · A)
W2′	V_o	MT	(T · ^{me}A)
W1′	D_p	Py	T · A, C · G, (T · ^{me}A, ^{me}C · G, U · A)
W1′	V_o	AM	A · T, (^{me}A · T, G · ^{me}C)
W1′	V_i	GV	G · C (A · U)
S1	D_p	N	All base pairs
S2	V_i	At	A · T, T · A, (^{me}A · T, T · ^{me}A, A · U, U · A)
S2′	A_p	Gc	G · C, C · G, (G · ^{me}C, ^{me}C · G)
S1′	D_p	N	All base pairs

*The column indicates the functional group on a hypothetical protein that interacts with the base pair in the indicated specificity site: D_p indicates that a hydrogen bond donor is present on the protein, which would be paired to an acceptor on the base; A_p indicates that a hydrogen bond acceptor is present on the protein; V_o indicates the presence of an an "outer" van der Waals contact to a methyl group on the DNA; V_i indicates the presence of an an "inner" van der Waals contact to a C-H hydrogen on a base. Here, "outer" and "inner" refer to the distance between the protein side chain and the base pair. A hydrophobic amino acid side chain in W1 would code for T or ^{me}C if it were positioned just far enough from the base pair to contact the methyl group. However, if it were positioned closer to the base pair, it would code for C by contacting C^5-H. †The symbol in this column refers either to the site-interaction combination or to the degenerate set of base pairs recognized. For example, Pu refers to a hydrogen bond donor on a protein in W1 which, in effect, codes for purines; Py, purine.

bond donor on adenine is "converted" to a receptor on guanine. GCATTC also has 11 possible protein-base hydrogen bonds because of the loss of one in W1 where an acceptor N-7 on adenine is "replaced" with the hydrogen atom on the C-5 of cytosine. GTATTC has ten possible hydrogen bonds because no hydrogen bonds can be formed between the protein and the thymine at the second position. This count gives the sequence A, (G, C), T, the observed Eco RI* hierarchy. The identical result is obtained at the third position. The fourth, fifth, and sixth positions follow from the symmetry of the Eco RI site.

Thus all the Eco RI* hierarchies can be correlated with the maximum number of possible hydrogen bonds between the enzyme and the particular Eco RI* sequence. The idea that the number of protein-base hydrogen bonds determines the Eco RI* hierarchies was first suggested by Rosenberg and Greene (*70*), who correctly identified the major groove contacts subsequently observed in the electron density map. The hydrogen bonds counted in this article do not include numerous hydrogen bonds between the protein and the DNA backbone because they should not be affected by base substitutions.

The amino acid residues that interact directly with the bases are arranged in space so that there are alternating positive and negative charges. The four amino acid residues in the inner recognition module are located around the central twofold crystallographic axis (Fig. 9). The residues from the outer recognition modules, Arg^{200} Arg^{203}, are above and below the twofold axis. When the DNA phosphates are included in the charge distribution, there is alternation of charges over the entire complementary binding site forming a very stable array of electrostatic charges.

The negative charges associated with the carboxyl groups of Glu^{144} (from both subunits) are "keystones" of the electrostatic array. It is likely that a significant displacement of either or both of these side chains would lead to disruption of the entire recognition interface. The Glu^{144} interacts with the central adenine bases at the site where the Eco RI methylase modifies the DNA; that is, at the exocyclic N-6 amino group. Methylation of either or both of these sites would inevitably displace one or both Glu^{144} side chains. If the endonuclease were to bind to a methylated Eco RI site, then the direct hydrogen bonds from N-6 would be lost, and the electrostatic character of the interface would be destabilized. Thus, the Eco RI endonuclease recognition interface seems highly poised to discriminate between the modified and unmodified hexanucleotides.

The spatial alternation of electrostatic charge suggests that the protein-DNA binding energy is a nonadditive function of the number of hydrogen bonds between the protein and the DNA bases; this means that each "correct" interaction should facilitate the formation of additional interactions of DNA and protein via the electrostatic forces. However, an "incorrect" structure due to the presence of a noncognate base in the enzyme's recognition site, would not facilitate and may even inhibit formation of additional protein-base interactions. These electrostatic interactions therefore constitute a form of cooperativity that would serve to sharpen the discrimination between the canonical hexanucleotide and all the incorrect sequences. We refer to this phenomenon as cooperative enhancement of specificity. Cooperative enhancement is also suggested by binding data with oligonucleotide substrates which show that the binding free energy is not a linear sum over the available hydrogen bonding sites (*58*). Nonadditivity has also been observed in the interaction between the lac repressor and operator (*71*), which could represent a second example of cooperative enhancement.

Conformational change and specificity. From a mechanistic viewpoint, the difficult theoretical problem is not to "explain" the Eco RI* activity; rather it is to understand the physical basis of the highly precise canonical specificity that occurs at physiological conditions. The hierarchical spectrum of Eco RI* sites is just what we should expect from a simple energy analysis of a recognition mechanism based solely on hydrogen bonds. Loss of a single hydrogen bond would be expected to reduce the interaction energy by 1 to 4 kcal/mol. The energy would probably be reduced further by an additional term due to cooperative enhancement. The resulting reduction in association constant or catalytic rate constant would be about two to four orders of magnitude. In other words, a recognition mechanism based solely on binding and hydrogen bonds would predict an "error rate" that is comparable to the misreading associated with Eco RI* activity. However, under physiological conditions, there is no detectable activity at Eco RI* sites.

The amino acid side chains that interact with the specific bases do not participate directly in the cleavage reaction and vice versa because the recognition and cleavage sites are physically separate. It is not an accident of the crystallization procedure that the cleavage site is not assembled in the structure reported here. Rather, our current working hypothesis is that this structure represents a functional intermediate in the catalytic pathway. Specifically, we propose that the recognition and cleavage sites are formed in an obligate temporal order that includes an isomerization from an "inactive" form to an active form of the sequence-specific complex. The structure reported here is the specifically bound inactive conformer. Furthermore, we suggest that there is physical coupling between the recognition and cleavage sites. As a result, the enzyme retains the inactive conformation under physiological conditions until all the sequence-specific DNA-protein interactions have formed. The transition is a form of allostery since the recognition and cleavage sites are spatially separate. We refer to the sequence dependent, allosteric isomerization from the inactive to the active form as "allosteric activation." Part of the free energy obtained from binding Mg^{2+} may be used to augment the sequence specificity of

Table 5. Combinations giving unambiguous base recognition. Unambiguous recognition of base pairs requires at least two protein-base interactions, which would be pairings of the interactions listed in Table 4. If all physically possible combinations of such pairings are examined, they fall into three categories: Those that cannot be satisfied by any base pair, those that are still not unique, and those that unambiguously specify a single base pair, which are shown here. The methylation states of adenine and cytosine are differentiated by some combinations, while others are insensitive to this modification. These are differentiated in the table. The combinations actually observed in Eco RI endonuclease are shown in bold face type. It should be noted that this table applies only to DNA.

Base pair	Combination	Base pair	Combination
A · T	**Pu + Ac**	T · A	**Gt′ + Py**
	Ac + AM		Me + Gt′
	Ac + At		At + Gt′
^{me}A · T	MA	T · ^{me}A	MT
A · T	Pu + Ac′	T · A	Gt + Py
or	Pu + At	or	At + Py
^{me}A · T	Ac′ + AM	T · ^{me}A	Me + Gt
	Ac′ + At		At + Gt
	AM + At		At + Me
G · C	GV	C · G	CU
G · ^{me}C	Gt + AM	^{me}C · G	Me + Ac′
	Gt′ + AM		Me + Ac
	Gc + AM		Me + Gc
G · C	**Pu + Gt**	C · G	**Ac′ + Py**
or	Pu + Gt′	or	Ac + Py
G · ^{me}C	Pu + Gc	^{me}C · G	Gc + Py
	Gt′ + Gc		Gc + Ac
	Gt + Gc		Gc + Ac′

allosteric activation. Relatively subtle effects could dramatically alter the equilibrium between the inactive and active states. It is not unreasonable to argue that Eco RI* buffer conditions alter that equilibrium toward the active form even when one or two hydrogen bonds have not formed correctly.

Modrich and co-workers have independently arrived at the allosteric activation hypothesis in order to account for two observations they have recently made (*72*). (i) Their kinetic data show that during the normal catalytic cycle, Eco RI endonuclease is bound to nonspecific DNA (which is not hydrolyzed) for a much larger fraction of time than it is bound specifically to cognate DNA (which is hydrolyzed). The data show that the total lifetime of all bound states is not the determinant of the cleavage rate; and they show that there are multiple bound states, some of which are inactive for cleavage. (ii) A mutation that replaces Glu^{111} with Gly retains full DNA binding specificity, but shows no cleavage activity under physiological conditions. Under Eco RI* conditions, the mutant enzyme cleaves DNA at Eco RI sites (at a rate much slower than that of wild type). Modrich and co-workers suggest that the mutation interferes with an isomerization between an inactive and active form. Glu^{111} is not near the DNA and cannot directly participate in the formation of either the recognition or cleavage sites.

The very high sequence specificity in Eco RI endonuclease derives from a series of sequence-specific steps including DNA binding and allosteric activation. Errors are corrected at each step via dissociation of noncognate DNA-protein complexes, resulting in a very low rate of cleavage at noncognate sites. This analysis suggests that an enzyme that covalently modifies DNA is intrinsically capable of achieving a much higher level of sequence discrimination than is a simple binding protein. Thus, sequence-specific covalent modification of DNA may be important in higher organisms, which contain large quantities of DNA and which must precisely regulate crucial cellular events, such as those associated with development.

Recapitulation. The Eco RI endonuclease–DNA recognition complex consists of a distorted double helix and a protein dimer composed of identical subunits related by a twofold axis of rotational symmetry. The distortions of the DNA are induced by the binding of the protein. They are concentrated into separate features that are localized disruptions of the double helical symmetry. These disruptions appear to have structural consequences that propagate over long distances through the DNA via twisting and perhaps bending effects. They are therefore referred to as neokinks. The type I neokink spans the central twofold symmetry axis of the complex, and it introduces a net unwinding of 25° into the DNA. The unwinding increases the separation of the DNA backbones across the major groove thereby facilitating access by the protein to the base edges, which are at the floor of the groove. The type I neokink also realigns adjacent adenine residues within the central AATT tetranucleotide in order to create the detailed geometry necessary for amino acid side chains to bridge across these purines.

Each subunit is composed of a single principal domain with a central five-stranded wall of β sheet bracketed by α helices; that is, it is organized according to α/β architecture. Each domain also has an extension called an arm, which wraps around the DNA. The domain can be subdivided into topological motifs that have identifiable functional roles. The three-stranded parallel motif is associated with sequence recognition and the subunit interface. The three-stranded antiparallel motif is associated with phosphodiester bond cleavage. The two segments overlap to form the five-stranded β sheet.

The surface of the protein is involuted to form two symmetry-related clefts which bind segments of the DNA backbone including the scissile bond. The cocrystals were grown in absence of Mg^{2+} in order to prevent DNA cleavage, but they can be activated for strand scission by diffusing Mg^{2+} into the crystals. The structure reported in this article appears to represent a specifically bound, inactive conformer that isomerizes to a specifically bound, active enzyme upon addition of Mg^{2+}. We suggest that the isomerization plays the important functional role of enhancing the specificity of Eco RI endonuclease by allosteric activation. The protein-base interactions at the sequence recognition site have a strong allosteric effect on the equilibrium between the inactive and active forms so that the active form is favored only when the cognate sequence is bound (under physiological conditions). The allosteric activation model accounts for the relaxation of specificity under Eco RI* conditions by invoking a solvent-mediated shift of the conformational equilibrium toward the active form even when Eco RI* sites are bound to the protein.

Sequence specificity is mediated by 12 hydrogen bonds between the protein and bases within the Eco RI hexanucleotide. These interactions depend on both the relative positioning as well as the identity of the bases and amino acid side chains at the DNA-protein interface. Unitary α helices position the key amino acid residues with respect to the DNA. These α helices are organized into modules with a spatial division of labor across the recognition site. The outer G·C base pairs are recognized by identical, symmetry-related outer modules. Each outer module consists of a single α helix. The inner tetranucleotide, AATT, is recognized by an inner module which consists of two symmetry-related α helices, one from each subunit. Amino acid side chains from the modules establish the relative position of the α helices to form a four-helix bundle. Additional amino acid side chains position the bundle with respect to the DNA by interacting with the DNA backbone and by anchoring the recognition bundle within secondary structure of the complex.

Bidentate hydrogen bonds between Arg^{200} and guanine (Arg::G) determine the base specificity of the outer module. Substitution of any base other than guanine would lead to rupture of at least one of these hydrogen bonds. The inner module also utilizes bidentate hydrogen bonds, but in a bridging tetrad arrangement with Glu^{144} and Arg^{145} forming four hydrogen bonds to adjacent adenine residues (Glu-Arg::AA). Substitution of any other base for either adenine residue would also result in rupture of at least one hydrogen bond. No hydrogen bonds are formed with the pyrimidine residues; however, they are recognized by hydrogen bonds to the purines on the complementary strand. The 12 hydrogen bonds therefore occur only in the canonical Eco RI hexanucleotide. These interactions are also consistent with the spectrum of Eco RI* cleavage rates because the observed hierarchies of cleavage rates can be predicted simply by counting the maximal number of hydrogen bonds possible between the protein and relevant Eco RI* sites.

The recognition interactions are stabilized by interactions between amino acid side chains, including electrostatic interactions between oppositely charged pairs: Glu^{144}-Arg^{145} and Glu^{144}-Arg^{200}. These interactions suggest that the DNA-protein interaction energy is not a simple additive sum over the individual interactions; that is, the system utilizes cooperative enhancement to sharpen the discrimination between cognate and noncognate sites. Formation of some correct protein-base interactions facilitates formation of additional correct interactions, whereas incorrect interactions with noncognate bases have an inhibitory effect. Glu^{144} side chains from both subunits are centrally located in the electrostatic array. Methylation of either N-6 amino group by Eco RI methylase would rupture a hydrogen bond and displace one of these negative charges. The charge displacement should perturb the entire recognition interface, thereby sharpening the discrimination between the modified and unmodified Eco RI sites.

REFERENCES AND NOTES

1. W. F. Anderson, D. H. Ohlendorf, Y. Takeda, B. W. Matthews, *Nature (London)* **290**, 754 (1981).
2. W. F. Anderson, Y. Takeda, D. H. Ohlendorf, B. W. Matthews, *J. Mol. Biol.* **159**, 745 (1982).
3. C. O. Pabo and M. Lewis, *Nature (London)* **298**, 443 (1982).
4. D. B. McKay and T. A. Steitz, *ibid.* **290**, 744 (1981).
5. T. A. Steitz, D. H. Ohlendorf, D. B. McKay, W. F. Anderson, B. W. Matthews, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 3097 (1982).
6. R. W. Schevitz, Z. Otwinowski, A. Joachimiak, C. L. Lawson, P. B. Sigler, *Nature (London)* **317**, 782 (1985).
7. M. Lewis *et al.*, *Cold Spring Harbor Symp. Quant. Biol.* **42**, 440 (1983).
8. D. H. Ohlendorf, W. F. Anderson, R. G. Fisher, Y. Takeda, B. W. Matthews, *Nature (London)* **298**, 718 (1982).
9. R. T. Sauer, R. R. Yocum, R. F. Doolittle, M. Lewis, C. O. Pabo, *ibid.*, p. 447.
10. J. E. Anderson, M. Ptashne, S. C. Harrison, *ibid.* **316**, 596 (1985).
11. P. J. Greene, M. Gupta, H. W. Boyer, W. E. Brown, J. M. Rosenberg, *J. Biol. Chem.* **256**, 2143 (1981).
12. A. K. Newman, R. A. Rubin, S.-H. Kim, P. Modrich, *ibid.*, p. 2131.
13. P. Modrich and D. Zabel, *ibid.* **251**, 5866 (1976).
14. L. Jen-Jacobson *et al.*, *ibid.* **258**, 14638 (1983).
15. B. A. Connolly, F. Eckstein, A. Pingoud, *ibid.* **259**, 10760 (1984).
16. P. Modrich, *Q. Rev. Biophys.* **12**, 315 (1979).
17. S. E. Halford and N. P. Johnson, *Biochem. J.* **191**, 593 (1980).
18. J. M. Rosenberg, H. W. Boyer, P. J. Greene, in *Gene Amplification and Analysis: Volume 1, Restriction Endonucleases*, J. G. Chirikjian, Ed. (Elsevier/North-Holland, Amsterdam, 1981), p. 131.
19. W. E. Jack, R. A. Rubin, A. Newman, P. Modrich, *ibid.*, p. 165.
20. J. L. Woodhead and A. D. B. Malcolm, *Nucleic Acids Res.* **8**, 389 (1980).
21. P. Modrich, *CRC Crit. Rev. Biochem.* **13**, 287 (1982).
22. B. J. Terry, W. E. Jack, R. A. Rubin, P. Modrich, *J. Biol. Chem.* **258**, 9820 (1983).
23. H.-J. Ehbrecht, A. Pingoud, C. Urbanke, G. Maass, C. Gualerzi, *ibid.* **260**, 6160 (1985).
24. B. J. Terry, W. E. Jack, P. Modrich, in preparation.
25. J. Grable *et al.*, *J. Biomol. Struct. Dyn.* **1**, 1149 (1984).
26. C. A. Frederick *et al.*, *Nature (London)* **309**, 327 (1984).
27. B.-C. Wang, *Methods Enzymol.*, in press.
28. The source of the slight nonisomorphism was obvious once the structure was solved. The Pt was bonded to the sulfur atoms of two methionyl residues, which fortuitously were in close proximity on the surface of the protein. However, the (mean) positions of the two sulfur atoms in the native structure were not precisely those required by the geometry of the bridging reaction which therefore appears to have required a small structural adjustment in a short segment of the polypeptide chain.
29. W. Furey, Program QKREF, unpublished data.
30. M. G. Rossmann, *J. Appl. Crystallogr.* **12**, 225 (1979).
31. M. G. Rossmann, A. G. W. Leslie, S. S. Abdel-Meguid, T. Tsukihara, *ibid.*, p. 570.
32. T. A. Jones, *ibid.* **11**, 268 (1978).
33. ———, in *Computational Crystallography*, D. Sayre, Ed. (Clarendon, Oxford, 1982), p. 303.
34. J. Anderson, M. Ptashne, S. C. Harrison, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 1307 (1984).
35. S. R. Jordan, C. O. Pabo, A. K. Vershon, R. T. Sauer, *J. Mol. Biol.* **185**, 445 (1985).
36. R. Kim, P. Modrich, S.-H. Kim, *Nucleic Acids Res.* **12**, 7285 (1984).
37. J. M. Rosenberg, N. C. Seeman, R. O. Day, A. Rich, *Biochem. Biophys. Res. Commun.* **69**, 979 (1976).
38. F. H. C. Crick and A. Klug, *Nature (London)* **255**, 530 (1975).
39. J. C. Wang, *J. Mol. Biol.* **43**, 25 (1969).
40. W. R. Bauer, F. H. C. Crick, J. H. White, *Sci. Am.* **243**, 118 (July 1980).
41. P. J. Hagerman, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 4632 (1984).
42. H. Wu and D. M. Crothers, *Nature (London)* **308**, 509 (1984).
43. H.-S. Koo, H.-M. Wu, D. M. Crothers, *ibid.* **320**, 501 (1986).
44. L. Ulanovsky, M. Bodner, E. N. Trifonov, M. Choder, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 862 (1986).
45. R. E. Dickerson and H. R. Drew, *J. Mol. Biol.* **149**, 761 (1981).
46. R. E. Dickerson, *ibid.* **166**, 419 (1983).
47. ——— and H. R. Drew, *Proc. Natl. Acad. Sci. U.S.A.* **78**, 7318 (1981).
48. T. J. Richmond, J. T. Finch, B. Rushton, D. Rhodes, A. Klug, *Nature (London)* **311**, 532 (1985).
49. T. M. Dunn, S. Hahn, S. Ogden, R. F. Schleif, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 5017 (1984).
50. K. Martin, L. Huo, R. F. Schleif, *ibid.* **83**, 3654 (1986).
51. A. Hochschild and M. Ptashne, *Cell* **44**, 681 (1986).
52. J. S. Richardson, *Adv. Protein Chem.* **34**, 167 (1981).
53. M. G. Rossmann, A. Liljas, C-I. Branden, L. J. Banaszak, in *The Enzymes*, P. O. Boyer, Ed. (Academic Press, New York, ed. **3**, 1975), p. 61.
54. W. G. S. Hol, *Prog. Biophys. Mol. Biol.* **45**, 149 (1985).
55. The exclusion of the β hairpin from the primary topology of the domain derives from the convention that protuberances of this type can be excluded from the assignment of the basic topological elements of a domain (*52*) since they could represent the evolutionary consequences of an insertion of DNA into an ancestral gene at a point which coded for a loop at the protein surface.
56. C. O. Pabo, W. Krovatin, A. Jeffrey, R. T. Sauer, *Nature (London)* **298**, 441 (1982).
57. J. L. Eliason, M. A. Weiss, M. Ptashne, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 2339 (1985).
58. L. Jen-Jacobson, D. Lesser, M. Kurpiewski, *Cell* **45**, 619 (1986).
59. J. Picone, Y. Kim, J. A. McClarin, P. Greene, J. M. Rosenberg, in preparation.
60. A-L. Lu, W. E. Jack, P. Modrich, *J. Biol. Chem.* **256**, 13200 (1981).
61. J. R. Lillehaug, R. K. Kleppe, K. Kleppe, *Biochemistry* **15**, 1858 (1976).
62. P. J. Greene *et al.*, *J. Mol. Biol.* **99**, 237 (1975).
63. N. C. Seeman, J. M. Rosenberg, A. Rich, *Proc. Natl. Acad. Sci. U.S.A.* **73**, 804 (1976).
64. R. P. Wharton, E. L. Brown, M. Ptashne, *Cell* **38**, 361 (1984).
65. R. H. Ebright, P. Cossart, B. Gicquel-Sanzey, J. Beckwith, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 7274 (1984).
66. Seeman, Rosenberg, and Rich (*63*) actually defined six major groove sites; however some of them were not functionally separate sites. In their notation, W2 and W3 were within 1 Å of each other and really functioned more or less as a single site. The same was true for W2′ and W3′. We have combined these into the functionally unique sites W2 and W2′.
67. J. M. Rosenberg, unpublished data.
68. B. Polisky *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **72**, 3310 (1975).
69. J. L. Woodhead, N. Bhave, A. D. B. Malcolm, *Eur. J. Biochem.* **115**, 293 (1981).
70. J. M. Rosenberg and P. J. Greene, *DNA* **1**, 117 (1982).
71. M. C. Mossing and M. T. Record, Jr., *J. Mol. Biol.* **186**, 295 (1985).
72. P. Modrich, personal communication.
73. M. L. Connolly, *Science* **221**, 709 (1983).
74. ———, *J. Amer. Chem. Soc.* **107**, 1118 (1985).
75. W. Provost, J. A. McClarin, J. M. Rosenberg, *Solvent Accessible Surfaces on the PS340, Laboratory Manual.*

76. We thank Paul Modrich for sharing unpublished results, William Provost for his assistance with PS340 raster graphics, and Oliver Bashor for technical assistance. This work was supported by NIH grant GM25671 (J.M.R.). Additional support was derived from BRSG grant RR07084, GM33506 (H.W.B.) and GM25729 (P.G.). The coordinates will be deposited at Brookhaven when refinement has been completed. At that time a full set could also be obtained from J.M.R.

19 August 1986; accepted 8 October 1986

"ATTENTION—I'M HAVING AN ENDOCRINOLOGIST IN TOMORROW TO LOOK AT ALL OF YOU."