Hin Recombinase Bound to DNA: The Origin of Specificity in Major and Minor Groove Interactions

Jin-An Feng, Reid C. Johnson, Richard E. Dickerson*

The structure of the 52–amino acid DNA-binding domain of the prokaryotic Hin recombinase, complexed with a DNA recombination half-site, has been solved by x-ray crystallography at 2.3 angstrom resolution. The Hin domain consists of a three–α-helix bundle, with the carboxyl-terminal helix inserted into the major groove of DNA, and two flanking extended polypeptide chains that contact bases in the minor groove. The overall structure displays features resembling both a prototypical bacterial helix-turn-helix and the eukaryotic homeodomain, and in many respects is an intermediate between these two DNA-binding motifs. In addition, a new structural motif is seen: the six–amino acid carboxyl-terminal peptide of the Hin domain runs along the minor groove at the edge of the recombination site, with the peptide backbone facing the floor of the groove and side chains extending away toward the exterior. The x-ray structure provides an almost complete explanation for DNA mutant binding studies in the Hin system and for DNA specificity observed in the Hin-related family of DNA invertases.

The Hin recombinase catalyzes a DNA inversion reaction in the *Salmonella* chromosome (*1*, *2*). This site-specific recombination reaction controls the alternate expression of two flagellin genes by reversibly switching the orientation of a promoter. During the process of inverting the 1-kb segment of DNA, Hin proteins bind to left and right recombination sites (*hixL* and *hixR*, respectively) located at the boundaries of the invertible DNA segment. The *hixL* and *hixR* sites with their bound Hin protein then form a synaptic complex with a third *cis*-acting site, a recombinational enhancer, which itself is bound by two dimers of the 98–amino acid Fis protein. Formation of this invertasome complex (*3–5*) aligns the two recombination sites correctly and activates the Hin protein to initiate the exchange of DNA strands, leading to inversion of the intervening DNA.

Hin belongs to a family of bacterial DNA invertases or recombinases that includes Gin from phage Mu, Cin from phage P1, and Pin from the *e*14 prophage of *Escherichia coli*. In addition to sharing 66 to 80 percent sequence identity between pairs of sequences, this family of proteins can substitute functionally for one another in each biological system (*1*). These DNA invertases most likely constitute an evolutionary family not unlike the c-type cytochromes. The availability of DNA sequence information from the binding sites of all four systems makes the present study of Hin-DNA binding especially informative in elucidating principles of protein-DNA recognition.

Hin binds to each *hixL* and *hixR* recombination site as a dimer. The final 52 amino acids of the two chains (Fig. 1A) are involved in binding to a 26-bp recombination site (Fig. 1B) built from two 12-bp imperfect inverted repeats separated by a 2-bp core region where DNA strand exchange occurs (*6–8*). The amino-terminal 138–amino acid "catalytic" domain is positioned in part over the core nucleotides. Although the monomeric 52–amino acid peptide by itself can bind to a recombination half-site with low-to-moderate affinity (dissociation constant $K_d \approx 10^{-7}$), cooperative interactions between the amino-terminal domains of two intact Hin molecules are required for high-affinity binding ($K_d \approx 10^{-9}$) (*8–10*).

A large body of footprinting, mutation, and chemical derivatization data has indicated features of Hin-DNA interaction which are distinctive to prokaryotic DNA-binding proteins (*8–13*). Specific binding requires both major groove interactions involving a helix-turn-helix (HTH) α-helix motif and minor groove interactions involving the sequence Gly^{139}-Arg^{140}-Pro^{141}-Arg^{142}. If residues 139 and 140 are deleted from the carboxyl-terminal domain, for example, then sequence-specific binding to DNA is abolished. As Fig. 1A shows, these

A Amino Acid Sequences of DNA-Binding Domains of Enteric Invertases

```
α-Helix:                1---------1     2------2     3------3
  Hin     GRPRAITKH.EQEQISRLLEK.GHP.RQQLAIIF.GIG.VSTLYRYF.PASSIKKRMN
  Gin     GRPPKLTKA.EQEQAGRLLAQ.GIP.RKQVALIY.DVA.LSTLYKKH.PAKRAHIENDDRIN
  Cin     GRRPKYQEE.TQQQMRRLLEK.GIP.RKQVAIIY.DVA.VSTLYKKF.PASSFQS
  Pin     GRRPKLTPE.QQAQAGRLIAA.GTP.RQKVAIIY.DVG.VSTLYKRF.PAGDK
          |         |           |           |         |
          139       148         162         173       181
```

B *hixL* binding site for Hin protein:

```
 -13        -8   * * *    -1+1            +8 * *   +13
5'-T-T-C-T-T-G-A-A-A-A-C-C-A-A-G-G-T-T-T-T-T-G-A-T-A-A-3'
3'-A-A-G-A-A-C-T-T-T-T-G-G-T-T-C-C-A-A-A-A-A-C-T-A-T-T-5'
                                  * * *
```

C Synthetic *hixL* half-site for crystallography:

```
                 2  3  4  5  6  7  8  9 10 11 12 13 14 15
Strand 1:     5'-T--G--T--T--T--T--T*-G--A--T--A--A--G--A
Strand 2:           C--A--A--A--A--A--C--T--A--T--T--C*-T--A-5'
                   29 28 27 26 25 24 23 22 21 20 19 18 17 16
```

Fig. 1. Amino acid and DNA base sequences in the Hin recombinase family (*1*, *34*). (**A**) Amino acid sequence of the DNA-binding domain (the 52 carboxyl-terminal residues) of the Hin protein and of corresponding regions in Gin, Cin, and Pin. Residues in boldface are identical in all four sequences or at least in three of the four. α Helices 1 to 3 as located in our Hin structure analysis are marked above the Hin sequence. For crystallography, this Hin fragment was synthesized manually or on an ABI 430A synthesizer by the solid-phase method as described previously (*8*). (**B**) Base sequence of the left DNA inversion site, *hixL*. Numbering is to either direction from the center of the inverted repeat. Asterisks mark purine bases that are protected from methylation by the binding of Hin. (**C**) The 14-bp synthetic *hixL* half-site as cocrystallized with the Hin 52-mer. Strand 1 of the duplex is numbered 2 through 15 to match the right half of the entire *hixL* site in (B). Bases in strand 2 are numbered separately. For ease of reference, note that base *n* in strand 1 is paired with base (32 − *n*) in strand 2. Base pairs will be referenced by strand 1 numbers alone: that is, "base pair 9" is understood to signify base pair G9·C23. Phosphates always are numbered according to the base that follows: Phosphate P9 occurs between T8 and G9, whereas across the helix on the other strand, phosphate P24 lies between C23 and A24. Hence phosphate *n* on strand 1 lies opposite phosphate (33 − *n*) on strand 2. Asterisks mark bases that were iodinated for purposes of multiple isomorphous replacement phase analysis.

The authors are with the Molecular Biology Institute and Department of Biological Chemistry, University of California, Los Angeles, CA 90024.

*To whom correspondence should be addressed at the Molecular Biology Institute.

two residues also are invariant among all four of the DNA invertases.

By comparison with pairwise amino acid sequence identities of 66 to 80 percent in the entire protein sequences, the carboxyl-terminal peptides shown in Fig. 1A have somewhat fewer identities, 49 to 62 percent, but still are obviously homologous proteins. The Hin interactions resemble those in the binding of DNA by the homeodomain in eukaryotes. Indeed, as noted by Affolter *et al.* (*14*), the amino acid sequence of the Hin binding domain can be aligned with that of the homeodomain of the *Drosophila* engrailed protein with a 27 percent sequence identity, sufficient to suggest a class resemblance.

We report here the 2.3 Å resolution crystal structure analysis of the 52–amino acid carboxyl-terminal DNA-binding domain of Hin complexed with a 14-bp DNA oligomer containing a half *hixL* binding site (Fig. 1C). The Hin peptide forms a three–α-helix core with an extended chain at each end. The core interacts with the major groove of DNA, whereas the flanking amino- and carboxyl-terminal chains extend along two regions of the minor groove. The carboxyl-terminal eight-residue tail of the Hin peptide crosses the phosphodiester backbone and is inserted in the minor groove in a manner that has not heretofore been encountered in DNA-protein complexes. The crystal structure provides a virtually complete explanation of base specificity experiments in solution, including mutant studies.

Structure of the complex. The structure of the Hin-DNA complex was solved by multiple isomorphous replacement with the use of three iodine derivatives (Table 1). A typical section of the final ($2F_o - F_c$) map is shown in Fig. 2. One 52–amino acid Hin domain binds to each 14-bp DNA half-site (Fig. 3). DNA helices consisting of the 13 complete base pairs are stacked end-to-end in the crystal, as in figure 2 of (*15*). The unpaired base T2 on strand 1 (Fig. 1C) swings up to make a Hoogsteen-like interaction with base pair 3, G3·C29. At the other end, unpaired base A16 on strand 2 is not defined in the electron density map and presumably is disordered. The *a* axis of the crystal, the direction of stacking of two DNA helices, has a length at room temperature of 86.4 Å = 26 × 3.32 Å. The 12–base pair steps along the helix produce a total rotation of 407° (average 33.9° per step), so that the nonbonded interhelix junction between base pair 15 (A15·T17) of one helix and base pair 3 (G3·C29) of the next requires a reverse twist of 360° − 407° = −47°.

The DNA half-site is a standard B-DNA helix, with the usual local variation in helix parameters (*16, 17*). The helix is relatively straight and not curved around the protein domain as in CAP, *trp* repressor, 434 repressor, and Met repressor (*18–22*), and has been proposed for the Fis-DNA complex (*23, 24*). However, an unusually large amount of DNA surface area is contacted by Hin. Upon binding Hin peptide, the DNA half-site monomer loses 1816 Å^2 of static solvent-accessible surface area.

The DNA half-site contains a short run of five AT base pairs (numbers 4 through 8) that could be regarded as a segment of A-tract DNA. Three frequent characteristics of A-tract DNA are a straight and unbent helix axis, narrow minor groove, and large propeller twist, large enough for the formation of bifurcated hydrogen bonds within the major groove between adjacent base pairs (*25–27*). However, in the Hin-DNA complex the minor groove maintains a uniform width of approximately 6.5 to 8.5 Å (minimal P-P atom separation across the groove, less 5.8 Å for two phosphate group radii), rather than the 3.5 to 4.5 Å typical of most A-tracts. Propeller twist is large all along the Hin-DNA complex, averaging

Table 1. Summary of crystallographic analysis. Crystals of Hin domain/DNA complex were grown by vapor diffusion against 15 to 18 percent PEG1500 as described elsewhere (*15*). The structure was determined initially to 3.2 Å resolution by multiple isomorphous replacement (MIR). Heavy atom parameters were refined and MIR phases calculated with the program HEAVY (*44*). The initial MIR map generated after solvent flattening (*45*) revealed clear density for B-form DNA and for most of the protein backbone density. This map was improved further by refining heavy atom parameters against solvent-flattened phases (*46*). After two additional cycles of phasing, solvent flattening, and heavy atom parameter refinement, the final MIR map, with mean phasing figure of merit of 0.55 for data between 20.0 and 3.2 Å, was used to build a model of the complex. However, it still was difficult to fit the amino acid sequence into many regions of the map. Only after phases were extended and modified to 2.8 Å by the method of Zhang and Main (*47*) did the map show clear density for side chains of some "marker" residues. At that point, all residues could be fitted unambiguously with the exception of the final eight carboxyl-terminal amino acids. Conventional positional refinement then was carried out to 2.8 Å with X-PLOR (*48, 49*). To refine the model further against a new 2.3 Å data set collected at −150°C, rigid-body refinements were carried out in successive steps to 2.5 Å. After positional refinement and simulated annealing, the ($2F_o - F_c$) map was of sufficient quality to allow the last eight residues to be built into the minor groove of the DNA, following a clear and continuous density. Electron density for residues Ser^{183} and Ser^{184}, however, remained poorly defined. Refinement was extended to 2.3 Å in four cycles of simulated annealing with X-PLOR prior to tightly restrained B-factor refinement. At the present stage of refinement, the agreement of the atomic model to crystallographic data is R = 0.228 for 8.0 to 2.3 Å resolution data. Coordinates have been deposited with the Brookhaven Protein Data Bank and are available for immediate distribution.

Parameter	Native	Native (−150°C)	IdU^8	IdC^{18}	IdU^8 + IdC^{18}
		*Unit cell dimensions**			
a (Å)	86.4	84.9	85.9	86.0	86.6
b (Å)	84.7	81.4	82.6	82.6	83.5
c (Å)	47.4	44.0	47.0	46.4	47.2
		Data collection statistics			
Resolution (Å)	2.8	2.3	3.2	3.2	3.5
Measured reflections	11673	17839	10387	7578	7578
Unique reflections	4177	5645	2703	2501	2020
Reflections > 2σ(F)	2653	5346			
Completeness (%)	92.6	80.1	92.1	87.5	87.0
R_{sym}† (%)	5.7	5.6	7.5	7.8	8.4
Mean isomorphous difference‡ (%)			18.4	16.9	18.3
		Phasing statistics			
Resolution (Å)			20–3.2	20–3.2	20–3.5
Cullis R factor			0.59	0.58	0.54
Phasing power§			1.80	1.74	2.42
		Refinement			
Resolution (Å)		8.0–2.3			
R factor‖		0.228			
Reflections with $F > 2\sigma$		5346			
Total number of atoms		978			
Water molecules	16				
Rms deviations					
Bond lengths (Å)		0.024			
Bond angles (deg.)	3.97				

*Space group $C222_1$ with one Hin-DNA complex per asymmetric unit. †$R_{sym} = \Sigma(|I - \langle I\rangle|)/\Sigma\langle I\rangle$, where I is the observed intensity and $\langle I\rangle$ is the averaged intensity obtained from multiple observations of symmetry-related reflections. ‡The mean isomorphous difference is $\Sigma||F_{PH}| - |F_P||/\Sigma|F_P|$, where $|F_P|$ and $|F_{PH}|$ are structure factor amplitudes of protein and heavy atom derivative of the protein, respectively. §Phasing power is the mean amplitude of heavy atom structure factors, F_H, divided by E, the root-mean-square lack-of-closure error. The mean figure of merit, 20.0 to 3.2 Å, is 0.55. ‖Crystallographic R factor = $\Sigma(|F_o - F_c|)/\Sigma(F_o)$ where F_o and F_c are the observed and calculated structure magnitudes, respectively.

−16°, but is not systematically larger in the A-A-A-A-A region than elsewhere. Hence the five AT base pairs do not constitute a classical A-tract structure, perhaps because of binding to Hin.

Overall protein folding. The 52–amino acid DNA-binding domain of Hin consists of a compact bundle of three α-helices, with extended amino-terminal arm and carboxyl-terminal tail (Fig. 3). α-Helix 1 (Glu^{148} to Lys^{158}) lies parallel to the axis of the DNA, α-helix 2 (Arg^{162} to Phe^{169}) is nearly antiparallel to helix 1 with an angle of −25° between helix axes, and α-helix 3 (Val^{173} to Phe^{180}) is inserted in the major DNA groove parallel to the base pairs (not to the floor of the groove itself). The HTH motif formed by helices 2 and 3 is similar to those found in other prokaryotic regulatory DNA-binding proteins. The Hin HTH region can be superimposed on equivalent regions of Fis (*23*, *24*) and λ repressor (*28*) with root-mean-square deviations in Cα atomic positions of only 0.61 and 0.76 Å, respectively.

All three α helices are amphipathic, with hydrophobic residues packed tightly against one another in a hydrophobic core (Fig. 3A). Ile^{152} and Leu^{156} of helix 1 interdigitate with Leu^{165} and Phe^{169} of helix 2. Val^{173}, Leu^{176}, and Phe^{180} of helix 3 also point into the hydrophobic core, which is delineated by the blue polypeptide backbone regions in Fig. 3A. At the bottom in that view, Ile^{144} on the amino-terminal arm closes the hydrophobic pocket. These hydrophobic interactions appear to be the main forces stabilizing the folding of the Hin protein. They also are strongly conserved among the other DNA invertases, Gin, Cin, and Pin (Fig. 1A), supporting the inference that all four of these proteins are folded in the same way. Hydrophobic interactions are supplemented by hydrogen bonds between side chains: Arg^{162} (invariant among the four invertases), at the beginning of helix 2, is hydrogen-bonded to main chain carbonyl oxygens of Phe^{180} (the final residue of helix 3) and Pro^{181}. Most charged side chains of the protein are either in contact with DNA or exposed to the solvent.

Major groove protein-DNA interactions. α-Helix 3 is the DNA recognition helix for the Hin protein; helices 1 and 2 are too far from the DNA to permit direct interactions. Only Gln^{163} at the amino terminus of helix 2 makes an indirect DNA contact through a hydrogen bond to residue Tyr^{177} (invertase invariant) in helix 3, which in turn contacts phosphate P19 (Fig. 3B). Five interactions between helix 3 and DNA backbone phosphates position the recognition helix properly, and two–amino acid side chains, Ser^{174} and Arg^{178}, make specific bonds to the edges of base pairs G9·C23 and A10·T22, as detailed below. It is significant that four of the eight amino acids in helix 3 are completely invariant among the four DNA invertases, and another three are semi-invariant, with the same residue in three sequences and a closely related one in the fourth.

The five nonspecific interactions with DNA backbone phosphates are depicted in Fig. 3B. The side chain of Tyr^{177} reaches up to phosphate P19 on one edge of the major groove, whereas Tyr^{179} on the other side of the α helix reaches down to phosphate P8 directly across the groove on the other wall. One of the terminal $-NH_2$ groups of the Arg^{178} side chain donates a hydrogen bond to the remaining oxygen of phosphate P8. The side chain of Thr^{175} and the main chain amide of Gly^{172} anchor helix 3 even further by donating hydrogen bonds to phosphate P9. In contrast to other HTH DNA-binding proteins, all of these nonspecific anchoring contacts to DNA phosphates are made by residues of helix 3; the "three-point contact" made by residues Gly^{172}, Thr^{175}, Tyr^{177}, and Tyr^{179} efficiently braces helix 3 against the opening of the major groove of DNA in a position to make specific recognition interactions.

Specific base sequence recognition uses only two Hin side chains, Ser^{174} and Arg^{178}, and in part involves the mediation of water molecules (Fig. 4) in a manner that has been proposed for the *trp* repressor (*29*). The side chain of Ser^{174} donates a hydrogen bond to the N-7 atom of A10. One turn of α helix away from this position, the terminal $-NH_2$ of Arg^{178} that is not involved with P8 donates a similar hydrogen bond to the N-7 of G9. The Arg^{178} ε-imino nitrogen donates another hydrogen bond to bound water molecule 1, which in turn donates a bond to the O-4 of T22, essentially "reading" the fact that this base pair is indeed A10·T22 and not a G·C pair. The other proton of this water molecule bonds to water molecule 2, which receives another hydrogen bond from the N-6 amine of A21 (recognizing this as an A·T pair and not G·C), and donates hydrogen bonds to N-7 of the same base and to the main chain carbonyl oxygen of Ser^{174}. Ser^{174} is invariant among all four DNA invertases. Arg^{178} is substituted only by Lys, and when this happens, two basic side chains always appear adjacent at positions 178 and 179 (Fig. 1A), meaning that some of the hydrogen bonds of the Hin structure could well be preserved. Both of the bound waters have the tetrahedral geometry expected for water molecules, donating two hydrogen bonds and accepting two others. The fourth bond to water 1, not shown explicitly in Fig. 4B, must be to another water molecule not well localized in the electron-density map.

All of these specific interactions are drawn in Fig. 4, A and B. Together they recognize base pairs 9, 10, and 11 of the half recombination site. Indeed, the binding of Hin is particularly sensitive to alterations of base pairs 9 and 10 (*13*). Dimethyl sulfate modification of the N-7 position of G9 inhibits Hin binding (*8*, *10*). Methylation at the N-6 position of A10 by the deoxyadenosine methylase of *Salmonella* decreases binding affinity (*13*). Hin binding

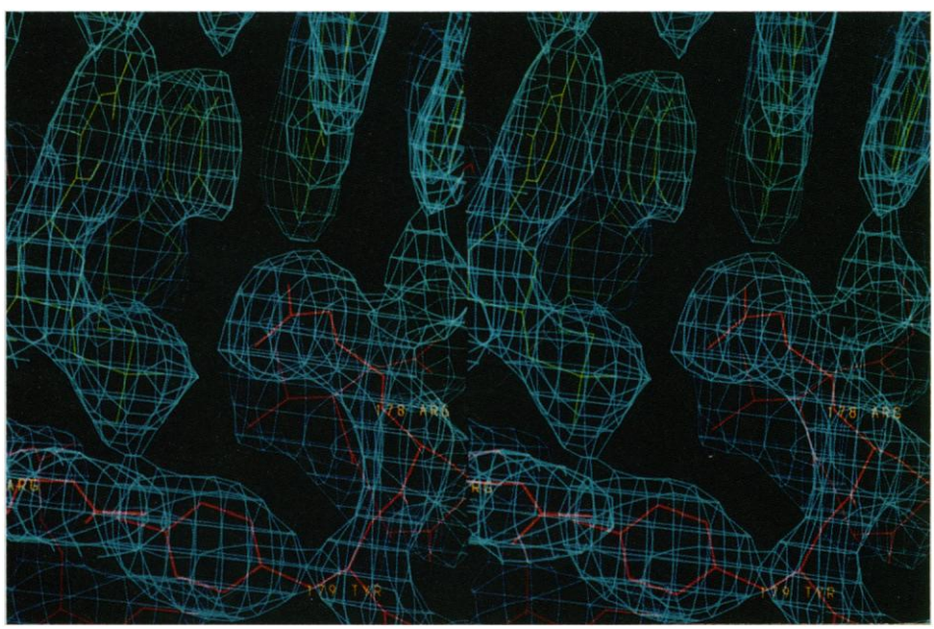


Fig. 2. Stereo view of the ($2F_o - F_c$) electron density map at 2.3 Å resolution (blue contours) showing portions of DNA base pairs 8 to 11 (top) and the region of helix 3 around Arg^{178} and Tyr^{179} (marked). Contour level is 1σ. The protein framework is in red, and DNA is in green.

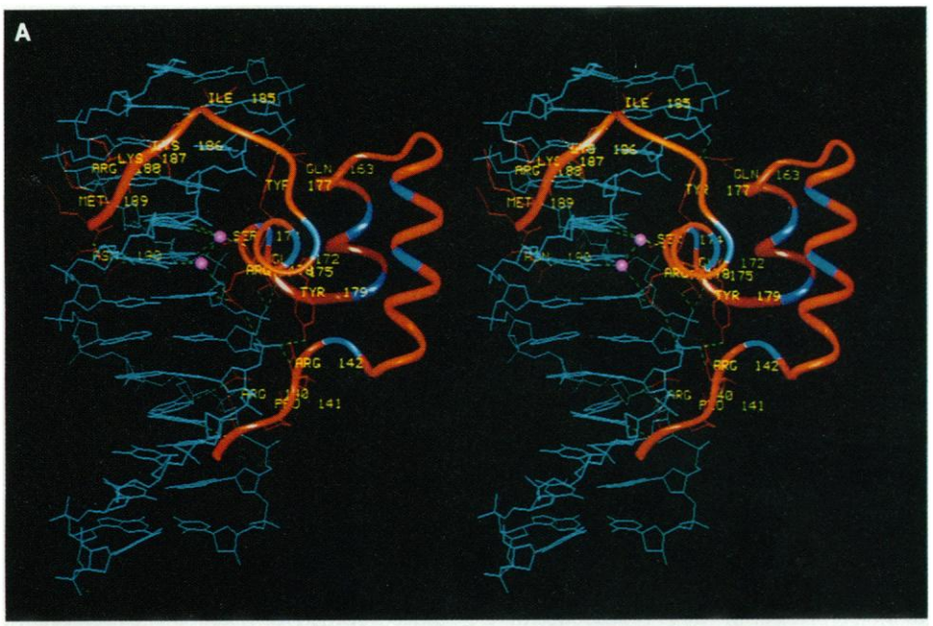
also is strongly and adversely affected by mutations at these sites, being inhibited by substitution of C at positions 9 and 11 or either C or T at position 10.

Minor groove protein-DNA interactions—the amino-terminal arm. Genetic and biochemical studies have demonstrated that contacts made by the amino-terminal arm of the Hin DNA-binding domain are at least as critical to DNA recognition as are those of helix 3 (*10*, *12*, *13*). Indeed, merely deleting Gly^{139} and Arg^{140} from the Hin DNA-binding peptide is sufficient to abolish specificity of binding to *hixL* (*12*). These residues are invariant in all of the DNA invertases.

The amino-terminal arm, Gly^{139}-His^{147}, adopts an extended conformation (Fig. 5). Clear electron density (Fig. 5A) allows Gly^{139} and Arg^{140} to be located unambiguously within the minor groove. The ε-imine of the Arg^{140} side chain donates a hydrogen bond to N-3 of A26. The unusually high 26° propeller twist of base pair 6 (T6·A26) permits a second hydrogen bond from the main chain amide of Arg^{140} to the O-2 of T6. If a G·C pair were to be substituted, this latter bond would become impossible, and the N-2 amine of guanine would push the Arg^{140} side chain away. Although the neighboring A·T base pairs are less propeller-twisted, the ability of an A·T pair to adopt such a large propeller contributes to the recognition process (*25*–*27*). Gly^{139} rests in close van der Waals contact with base pair 5; the main chain Cα atom of that residue is only 3.4 Å from the O-2 of T5 and 4.1 Å from the C-2 of A27. Introduction of an amine group at that locus, as in guanine, would push the Hin polypeptide chain up and away from the floor of the minor groove at that point. Each base pair substitution of A·T by G·C at positions 5 and 6 abolishes the binding affinity of Hin (*13*). Indeed, A·T base pairs at positions 5 and 6 are universally present in all of the recombination sites of various enteric inversion systems: *hixL* and *hixR*; *gixL* and *gixR*; *cixL* and *cixR*; and *pixL* and *pixR* (*1*) (Fig. 6). Biochemical footprinting experiments also show that both intact Hin and the Hin peptide protect adenines 5 and 6 from methylation (*10*).

Pro^{141} arches across one wall of the minor groove, and the ε-imino of Arg^{142} is hydrogen-bonded to phosphate P8, an interaction that may be important in directing the amino-terminal arm into the minor groove of the DNA. Hin, Gin, Cin, and Pin all have Pro at position 141 or 142, followed immediately by a basic Arg or Lys. A significant role may be played by Ile^{144}, which interacts with the hydrophobic core formed by packing helices 1 to 3 against one another. Ile^{144} may restrict movement of the amino-terminal arm, thus bringing Arg^{142} into proximity to phosphate P8. In other DNA invertases, position 144 is always a bulky hydrophobic side chain, either Leu or Tyr (Fig. 1A).

Minor groove protein-DNA interactions—the carboxyl-terminal tail. The carboxyl-terminal tail of the Hin polypeptide crosses the phosphodiester backbone at the



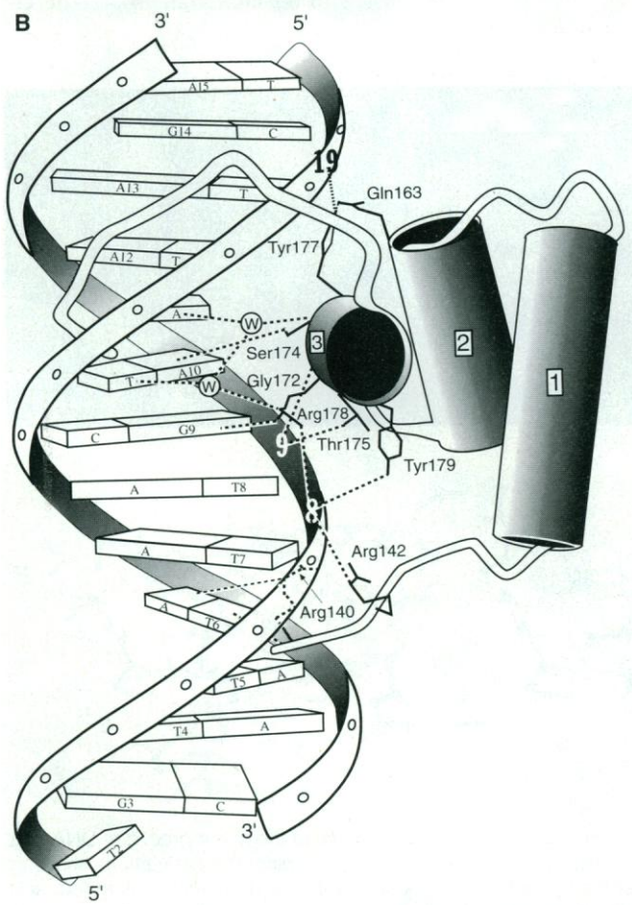


Fig. 3. (**A**) Stereo diagram of the Hin-DNA complex. DNA is in blue stick bonds. The path of the Hin polypeptide chain is shown as a flexible tubing: orange in general, but blue for hydrophobic residues. Side chains that contact the DNA are drawn in orange and labeled in yellow. Note the amino-terminal arm within the minor groove at lower right (Gly^{139}-Arg^{142}), the carboxyl-terminal tail in the minor groove at upper left (Ile^{185}-Asn^{190}), and helix 3 nested in the major groove. Pink spheres are two bound water molecules involved in protein-DNA contacts within the recognition site. Hydrogen bonds are in green. (**B**) Schematic drawing of the complex, in the same view as the stereo. Helices are numbered 1 to 3, and key side chains that interact with the DNA are identified. Open dots along backbone ribbons locate C1′ atoms. Phosphates 8, 9, and 19 are indicated specifically on the ribbons. Base pairs 9 to 11 are shown in stereo close-up in Fig. 4A, base pairs 4 to 8 in Fig. 5B, and pairs 9 to 15 in Fig. 7.

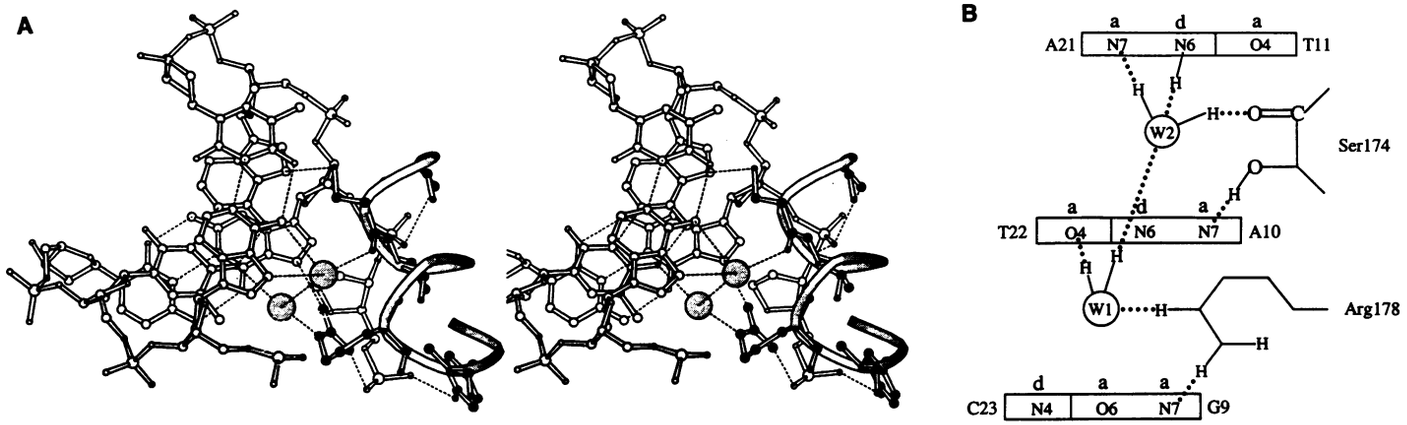


Fig. 4. (**A**) Stereo close-up of the interaction between DNA and helix 3 in the major groove, viewed approximately down the DNA helix axis. Base pair T11·A21 is nearest the viewer, with A10·T22 beneath, and G9·C23 farthest away. Strand 1 backbone continues to lower right through P9 and P8. Helix 3 is drawn as a smooth curve, with specific depiction, from top to bottom, of residues Gly^{172}, Ser^{174}, Thr^{175}, Arg^{178}, and Tyr^{179}. Two shaded spheres are bound water molecules. (**B**) Schematic of specific base pair recognition involving Ser^{174}, Arg^{178}, and two bound water molecules. Along base edges, "a" marks a hydrogen bond acceptor (ring N-7 or carbonyl O-4 or O-6) and "d" marks a hydrogen bond donor (N-4 or N-6 amine group).

outer edge of the recombination site and then follows the minor groove back toward the center of the 13-bp DNA helix (Figs. 3 and 7). The final six residues of the Hin polypeptide, Ile^{185}-Lys^{186}-Lys^{187}-Arg^{188}-Met^{189}-Asn^{190}, adopt an extended conformation and lie within the minor groove, but the side chains themselves make no contacts with the floor of the groove. Instead, they point outward, with the polypeptide backbone resting against base edges. At the point where the final six–amino acid residues dip into the minor groove, the main chain CO of Ile^{185} hydrogen bonds to the N-2 of G14. The main chain NH of Lys^{187} bonds to the O-2 of T20, and a little farther along, the main chain amide of Asn^{190} interacts with the O-2 of T22 while the side chain of the carboxyl-terminal residue bonds to the N-3 of A10. These interactions may be responsible for the large propeller twist and roll angles of base pairs 10 and 12 and the ~16° bend of DNA toward the major groove. Consistent with the interactions just discussed, the N-3 of A10 is partially protected from dimethyl sulfate attack by Hin binding (*10*). In addition, Mack *et al.* (*11*) have noted that a Hin peptide lacking the last eight residues, when modified with EDTA·Fe, cleaved DNA with reduced efficiency as compared with a peptide containing the complete carboxyl terminus. This portion of the chain is variable among the DNA invertases: whereas Hin has six final amino acids, Gin has ten, Cin has three, and Pin has a lone Lys (Fig. 1A). This carboxyl-terminal tail is presumably supportive but not essential.

The hydrogen-bonded extension of the last six residues along the floor of the minor groove recalls AT-specific binding of minor groove drugs such as netropsin and distamycin (*30*, *31*). Such binding involves an element of base specificity: if any of the base pairs 10 to 13 were G·C rather than A·T, then the tail of the Hin peptide would be pushed away from the floor of the groove. In another context, it has been proposed (*32*, *33*) that an extended polypeptide containing repeats of SPKK (*34*) sequence may interact with DNA minor groove in a similar fashion to netropsin, with main chain amide nitrogen forming hydrogen bonds with base pairs in the minor groove. Our structure seems to provide a concrete example of such a model.

The molecular basis of specificity. Two aspects of the Hin system make it especially conducive to an understanding of the ef-

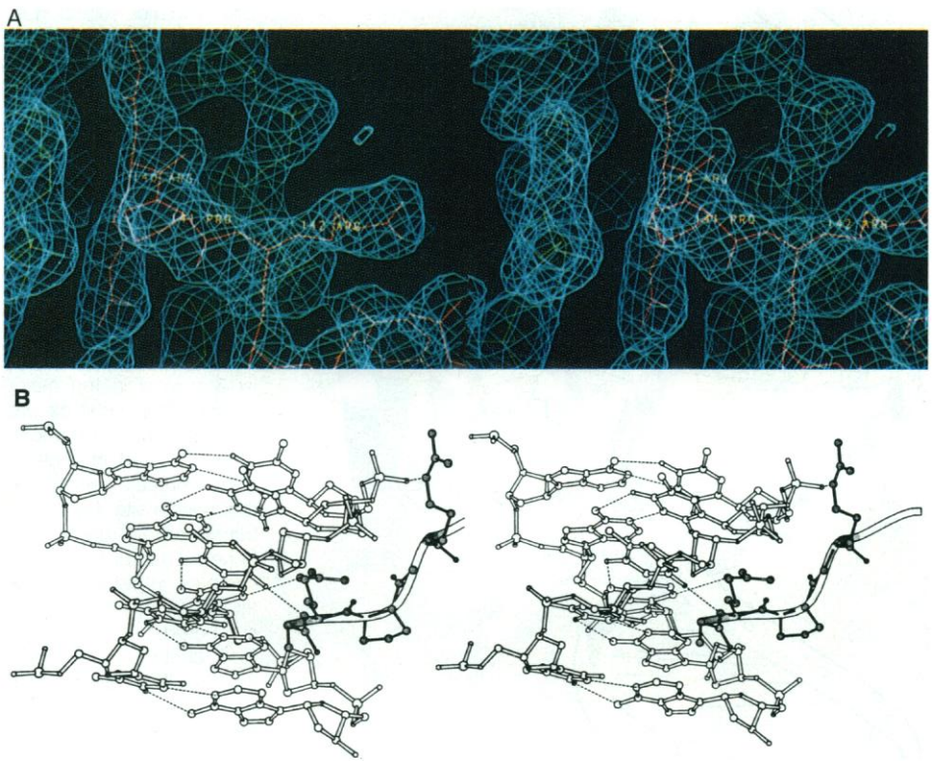


Fig. 5. Stereo views of the amino-terminal arm of the Hin peptide, in the minor groove of DNA. (**A**) The refined ($2F_o - F_c$) electron density map (blue framework) with minor groove vertical, showing Arg^{140}-Pro^{141}-Arg^{142} looping over the phosphate backbone toward the right. Protein is in red, and DNA is in green. (**B**) View along the minor groove, from the top in (A), showing the entire "A-tract" region, from T4·A28 at bottom through T8·A24 at top. This view is approximately that of the lower half of Fig. 3B. The ribbon extending from center toward upper right is the Hin peptide region Gly^{139}-Arg^{140}-Pro^{141}-Arg^{142}, with side chains drawn explicitly.

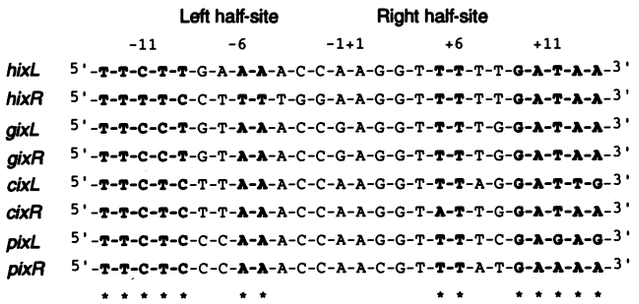


Fig. 6. Base sequences in enteric bacterial inversion sites: the left and right inversion sites from the Hin inversion system of *Salmonella*, the Gin inversion elements from bacteriophage Mu, Cin from phage P1, and Pin from the *e*14 prophage of *Escherichia coli* (*1*). Only one chain from each complex is shown; the other is complementary as in Fig. 1B. Each of the eight sites is built from two roughly symmetrical half-sites. Asterisks at bottom indicate positions that are especially important in site recognition by invertases.

fects of sequence on specificity. The first is the availability of binding data on 39 different base substitution mutations generated by Hughes *et al.* (*13*). They constructed a symmetrized *hix*C sequence in which the left half is given the complementary sequence to the right half shared by both *hix*L and *hix*R and established that this symmetrized *hix*C binds Hin fully as well as the wild-type *hix*L and *hix*R. (It is the *hix*C sequence that we used for crystallographic analysis.) They then constructed an exhaustive set of symmetrical mutants in the two halves of *hix*C, varying each of the 13 positions among all three of the other bases. Hence we have complete information about the strength of Hin binding with every possible single-base change in the optimal *hix*C sequence. The second favorable aspect is the existence of four homologous DNA inversion systems: Hin, Gin, Cin, and Pin. Taken together, these provide 8 complete recombination sites or 16 binding half-sites (Fig. 6). How far can our x-ray analysis of the Hin-DNA complex account for this wealth of data, and provide a molecular basis for DNA-protein specificity?

The Hin-DNA crystal structure shows that all three components of the Hin peptide, the amino-terminal arm, the HTH region, and the carboxyl-terminal tail, contribute to base sequence recognition. The phosphate backbone contacts of helix 3 help position Hin on the DNA, but one could easily imagine that Hin could slide along the DNA in a nonspecific manner until it encounters the correct local base sequence.

Interactions of Hin residues Ser^{174} and Arg^{178}, both direct and through intermediate water molecules (Fig. 4B) place restrictions on base pairs 9, 10, and 11. Some latitude in base sequence is possible if different arrangements of hydrogen bond donors and acceptors are permitted. Those rearrangements that are possible without losing the total number of hydrogen bonds are shown in Fig. 8. Base 9 is restricted to being a purine (G or A) by virtue of the hydrogen bond donated to ring atom N-7. In complete agreement with this model, of the 16 half-sites shown in Fig. 6 and listed in Table 2, G occurs 13 times at position 9 and A occurs 3 times. No pyrimidine is ever found at that locus. Replacement of G at position 9 by A or T (and C at position −9 by T or A) in the mutant studies of Hughes *et al.* (*13*) is acceptable, but C at position 9 reduces binding significantly. Our crystal structure shows why: G, A, or T at position 9 offer a hydrogen bond acceptor to Arg^{178}, whereas C offers a N-4 amine donor instead, and hence is disfavored. Even T is less favorable, because it positions the hydrogen bond acceptor differently and partially blocks it with its own C-5 methyl group.

Position 10 also must be a purine for the same reason as 9. The choice in 14 of the

Table 2. Frequency of occurrence of bases at key positions in bacterial inversion sites. Sequences are read in a 5′-to-3′ direction from the center of the inversion site as shown in Fig. 6. Hence, for left half-sites the other chain, not shown in Fig. 6, is tabulated. Allowed substitution data derive from the frequency of lysogenization in a P22 challenge phage assay at 100 μM isopropyl-β-D-thiogalactopyranoside (IPTG) concentration, table 1 of (*13*).

Position	Natural *hix, gix, cix,* and *pix* sites				Acceptable mutations in symmetrized *hixC* site (*13*)	DNA groove
	G	A	T	C		
5		2	14		Not G, not C	Minor
6		1	15		Not G, not C	Minor
9	13	3			Not C	Major
10	2	14			Not T, not C	Major
11	8	2	6		Not C	Major
12		15	1		Not C	Minor
13	2	14			All equivalent	Minor

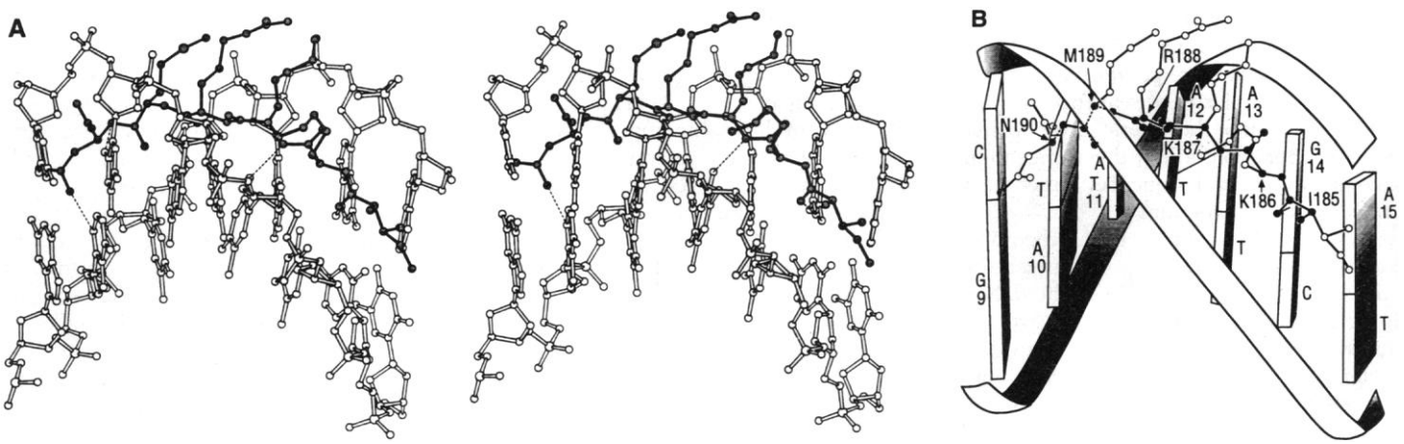


Fig. 7. Linear extension of carboxyl-terminal residues 185 to 190 down the minor groove. (**A**) Stereo pair representation. DNA base pair G9·C23 is at left, and A15·T17 at right. Amino acid side chains are identified in (**B**), which is a sketch from the same orientation. The main polypeptide chain atoms are in black dots; side chains are in open circles.

16 half-site sequences is A10, resulting in T22 at the other end of the base pair, with a hydrogen bond–accepting O-4 atom (Fig. 8A). However, G10·C22 is an acceptable minority choice in two sequences, and in that case the N-4 amine of C22 must donate a hydrogen bond to water 1 (Fig. 8B). Water 1 then would donate a hydrogen bond to another water molecule not shown here.

Base pair 11 is more variable than might have been expected, and for an interesting reason. Water molecule 2 in Fig. 8A accepts a hydrogen bond from the N-6 of A21 and donates a bond to N-7, but all that is required for a hydrogen balance is that this water molecule should donate one hydrogen and accept another. The two base atoms could just as easily be thymine O-4 and adenine N-6 as in Fig. 8A or cytosine N-4 and guanine O-6 as in Fig. 8B. The only combination not permitted would be guanine at position 21, with dual acceptors N-7 and O-6. For in that case, water 2 would not have enough protons to form the bond with the main chain carbonyl of Ser^{174}. In other words, the requirement on the strand 1 side of base pair 11 is "not-C," and indeed this requirement is borne out in Table 2 both by the invertase family sequence comparisons and mutational substitutions.

The direction of the hydrogen bond between water molecules—water 1 donating to water 2—actually is firmly established. As Fig. 8C shows, if water 2 were to donate a bond both to water 1 and to the Ser^{174} main chain carbonyl, then the two positions on base pair 11 would have to be adjacent donors, and no base pair shows this behavior. The full pattern of hydrogen bonds can be maintained only by arrangements as in Fig. 8, A and B.

A·T base pairs are favored at positions 12 and 13 because of a netropsin-like extension of the carboxyl-terminal six Hin residues down the floor of the minor groove (Fig. 7). Table 2 shows that A·T pairs indeed are overwhelmingly favored at these two loci, although mutant studies show position 13 to be more permissive. It could be that the similarity of sequences at this point among all of the enteric recombinases is a matter of evolutionary divergence from a common ancestral sequence, rather than convergence on function.

At the other end of the recognition domain, positions 4, 5, and 6 universally are A·T base pairs in all of the DNA inversion sites, and G·C pairs are strongly disfavored at positions 5 and 6 in the mutant studies. The x-ray structure shows the reason: Hin residues Gly^{139} and Arg^{140} are so intimately linked to the floor of the minor groove that there simply is no room for the C-2 amine group of guanine. As noted above, Gly^{139} and Arg^{140} are absolutely essential for sequence-specific binding of Hin to its DNA site.

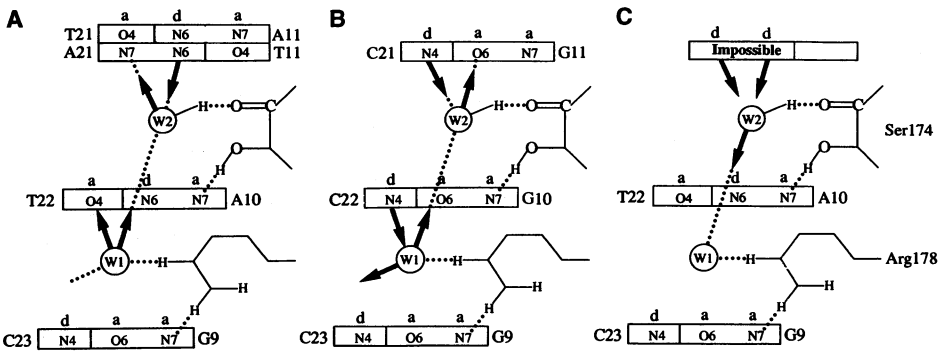


Fig. 8. The molecular basis for specificity, in the form of possible patterns of hydrogen bond donors and acceptors. (**A**) The pattern actually observed in the crystal structure. Base pair 11 could be reversed end-for-end without altering the pattern of hydrogen bond acceptor-donor-acceptor along its edge. (**B**) Hydrogen bond donors and acceptors could be reversed at base pair 11, allowing G11·C21. However, base 21 could not be G, with dual acceptors without breaking the bond between water 2 and the Ser^{174} carbonyl. Hence C11·G21 is strongly disfavored. The acceptor at the central base pair can independently be replaced by a donor, permitting G10·C22 in place of A10·T22, but base 10 is still required to be a purine. (**C**) The hydrogen bond connecting the two water molecules cannot be reversed, for then, if the donation to Ser^{174} carbonyl is maintained, base pair 11 would have to contribute two adjacent donors, which is chemically impossible for any of the base pairs.

Fig. 9. Optimal base sequence for enteric inversion half-sites. Numbers mark the distance out from the center of the site as in Fig. 6. A/T = an A·T base pair in either orientation. N = any base.

5	6	7	8	9	10	11	12	13
(A/T)–	(A/T)–	–N–	–N–	–(purine)–	–A–	–(not-C)–	–(A/T)–	–(weak A)

In summary, the recognition element of the enteric inversion half-sites appears to involve two A·T base pairs (5 and 6) recognized by amino acid residues Gly^{139} and Arg^{140} in the minor groove, two nonspecific base pairs (7 and 8), and then a five base sequence (9 to 13) recognized by helix 3 and the carboxyl-terminal tail in major and minor grooves, respectively. The optimal binding sequence is shown in Fig. 9. The Hin dimer has ~100-fold higher binding affinity to the full recombination site than does the Hin peptide binding to a half-site, and hence cooperative interactions by the Hin dimer may also contribute to recognition.

Hin-DNA versus other HTH DNA-binding complexes. The HTH motif occurs frequently in DNA-binding proteins, including prokaryotic regulators (*18, 20, 28, 35*), eukaryotic homeodomains (*36–38*), eukaryotic transcription factors such as HNF-3γ (*39*), the Oct-1 POU-specific domain, POUs (*40*), the third tandem repeat of the cMyb protein (*41*), and the globular domain of histone H5, GH5 (*42*). Complexes of these proteins with DNA show two principal variants that are represented in Fig. 10 by the λ repressor and the engrailed homeodomain. In all cases, recognition helix 3 fits into the major groove and helix 2 runs across the width of the groove. In homeodomain structures, helix 1 lies essentially antiparallel to helix 2, which also spans the width of the major groove. The residues preceding helix 1 are positioned to interact with the minor groove, and recognition helix 3 is oriented along the floor of the major groove. This pattern is persistent; the cluster of three helices is virtually superimposable from one homeodomain complex to the next. In contrast, prokaryotic regulatory proteins (*18, 20, 28, 35*) have similar dispositions of helices 2 and 3, but with the exception of *trp* repressor, helix 1 is swung out and away from the DNA (Fig. 10A). Helix 3, on the other hand, tends to lie parallel to the edges of base pairs in the prokaryotic regulators, rather than along the floor of the groove as in homeodomains.

The present Hin-DNA complex is intermediate between these two structures. Recognition helix 3 is parallel to base pair edges rather than to the groove itself, as with prokaryotic repressors, but helix 1 is nearly antiparallel to 2, with its amino terminus extending toward the minor groove where a nonhelical chain continuation contributes interactions essential to specific recognition. The alignment of helix 3 parallel to base pairs in Hin and repressors is functional: comparison of the structures of 434 repressor-operator, 434 *Cro*-operator, λ repressor-operator, and CAP-DNA complexes, shows that amino acids at positions 1, 2, and 6 of the helix make specific contacts with bases in the major groove in each case

Fig. 10. Comparative interactions of a three-helix unit with the DNA double helix in the λ repressor–operator complex (**left**), the Hin-DNA complex (**center**), and the engrailed homeodomain complex (**right**). In each case, the carboxyl-terminal helix of the three is inserted into the major groove.

(*43*). Similarly, Hin uses positions 2 and 6 (Ser^{174} and Arg^{178}) for its primary recognition process.

The disposition of helix 1 with respect to the DNA is not completely identical in Hin and homeodomains; in the latter, helix 1 crosses perpendicular to the walls of the major groove, whereas Hin has helix 1 aligned parallel with the overall DNA axis. The two loops connecting helices are shorter in Hin than in the homeodomain proteins and more like those of prokaryotic repressors. In addition, the α helices themselves are shorter than their homeodomain equivalents, especially helix 3, which in both yeast α2 and *Drosophila* engrailed homeodomains are 17 amino acids long.

There also is a remarkable difference in minor groove binding by the sequence Arg-Pro-Arg in the amino-terminal arm of Hin and in the engrailed homeodomain. In Hin, Arg^{140} makes specific base contacts, whereas Arg^{142} anchors the two-pronged fork by binding to the phosphate backbone of DNA. In the engrailed homeodomain, Arg^3-Pro^4-Arg^5 inserts both Arg side chains into the minor groove and makes specific base contacts, and it is the adjacent Thr^6 that interacts with phosphate backbone. Thus, the same three–amino acid recognition module can interact with the minor groove in two profoundly different ways to generate contacts essential for protein-DNA binding.

In another significant difference, the amino-terminal arms of homeodomain proteins run along the minor groove in the direction of the HTH unit itself, whereas the amino-terminal arm of the Hin protein runs in an opposite direction (Figs. 3A and 9), toward what would be the center of the recombination site. Model-building of an intact recombinase bound to a complete *hix* site suggests that the positions cleaved by Hin during recombination are located on the opposite face of *hix* from the DNA-binding domains. It is likely that residues preceding the amino-terminal arm within an intact Hin protomer may continue running along the minor groove around the DNA and link with the catalytic domain that presumably is positioned over the cleavage site.

REFERENCES AND NOTES

1. A. C. Glasgow, K. T. Hughes, M. I. Simon, in *Mobile DNA*, D. E. Berg and M. M. Howe, Eds. (American Society for Microbiology, Washington, DC, 1989), pp. 637–659.
2. R. C. Johnson, *Curr. Opin. Genet. Dev.* 1, 404 (1991).
3. K. A. Heichman and R. C. Johnson, *Science* **249**, 511 (1990).
4. K. A. Heichman, I. P. G. Moskowitz, R. C. Johnson, *Genes Dev.* **5**, 1622 (1991).
5. I. P. G. Moskowitz, K. A. Heichman, R. C. Johnson, *ibid.*, p. 1635.
6. R. C. Johnson and M. I. Simon, *Cell* **41**, 781 (1985).
7. R. C. Johnson and M. F. Bruist, *EMBO J.* **8**, 1581 (1989).
8. M. F. Bruist, S. J. Horvath, L. E. Hood, T. A. Steitz, M. I. Simon, *Science* **235**, 777 (1987).
9. J. P. Sluka, S. J. Horvath, M. F. Bruist, M. I. Simon, P. B. Dervan, *ibid.* **238**, 1129 (1987).
10. A. C. Glasgow, M. F. Bruist, M. I. Simon, *J. Biol. Chem.* **264**, 10072 (1989).
11. D. P. Mack *et al.*, *Biochem. J.* **29**, 6561 (1990).
12. J. P. Sluka, S. J. Horvath, A. C. Glasgow, M. I. Simon, P. B. Dervan, *ibid.*, p. 6551.
13. K. T. Hughes, P. C. W. Gaines, J. E. Karlinsey, R. Vinayak, M. I. Simon, *EMBO J.* 11, 2695 (1992).
14. M. Affolter *et al.*, *Cell* **64**, 879 (1991).
15. J.-A. Feng *et al.*, *J. Mol. Biol.* **232**, 982 (1993).
16. G. G. Privé, K. Yanagi, R. E. Dickerson, *ibid.* **217**, 177 (1991).
17. K. Grzeskowiak, K. Yanagi, G. G. Privé, R. E. Dickerson, *J. Biol. Chem.* **286**, 8861 (1991).
18. S. C. Schultz, G. C. Shields, T. A. Steitz, *Science* **253**, 1001 (1991).
19. Z. Otwinowski *et al.*, *Nature* **335**, 321 (1988).
20. A. K. Aggarwal, D. W. Rodgers, M. Drottar, M. Ptashne, S. C. Harrison, *Science* **242**, 899 (1988).
21. W. S. Somers and S. E. V. Phillips, *Nature* **359**, 387 (1992).
22. T. A. Steitz, *Q. Rev. Biophys.* **23**, 205 (1990).
23. H. S. Yuan *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 9558 (1991).
24. D. Kostrewa *et al.*, *Nature* **349**, 178 (1991).
25. H. R. Drew *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **78**, 2179 (1981).
26. H. C. M. Nelson, J. T. Finch, B. F. Luisi, A. Klug, *Nature* **330**, 221 (1987).
27. M. Coll, C. A. Frederick, A. H.-J. Wang, A. Rich, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 8385 (1987).
28. S. R. Jordan and C. O. Pabo, *Science* **242**, 893 (1988).
29. T. E. Haran, A. Joachimiak, P. B. Sigler, *EMBO J.* 11, 3021 (1992).
30. M. L. Kopka, C. Yoon, D. Goodsell, P. Pjura, R. E. Dickerson, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 1376 (1985).
31. ———, *J. Mol. Biol.* **183**, 553 (1985).
32. M. Suzuki, *EMBO J.* **8**, 797 (1989).
33. M. E. Churchill and A. A. Travers, *Trends Biochem. Sci.* **16**, 92 (1991).
34. Abbreviations for the amino acid residues are: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
35. A. Mondragon and S. C. Harrison, *J. Mol. Biol.* **219**, 321 (1991).
36. C. R. Kissinger, B. S. Liu, E. Martin-Blanco, T. B. Kornberg, C. O. Pabo, *Cell* **63**, 579 (1990).
37. C. Wolberger, A. K. Vershon, B. Liu, A. D. Johnson, C. O. Pabo, *ibid.* **67**, 517 (1991).
38. G. Otting *et al.*, *EMBO J.* **9**, 3085 (1990).
39. K. L. Clark, E. D. Halay, E. Lai, S. K. Burley, *Nature* **364**, 412 (1993).
40. N. Dekker *et al.*, *ibid.* **372**, 852 (1993); N. Assa-Munt, R. J. Mortishire-Smith, R. Aurora, W. Herr, P. E. Wright, *Cell* **73**, 193 (1993).
41. K. Ogata, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 6428 (1992).
42. V. Ramakrishnan, J. T. Finch, V. Graziano, P. L. Lee, R. M. Sweet, *Nature* **362**, 219 (1993).
43. S. C. Harrison and A. K. Aggarwal, *Annu. Rev. Biochem.* **59**, 933 (1990).
44. T. C. Terwilliger, S.-H. Kim, D. Eisenberg, *Acta Crystallogr.* **A43**, 1 (1987).
45. B.-C. Wang, *Methods Enzymol.* 115, 90 (1985).
46. M. A. Rould, J. J. Perona, T. A. Steitz, *Acta Crystallogr.* **A48**, 751 (1992).
47. K. Y. J. Zhang and P. Main, *ibid.* **A46**, 377 (1991).
48. A. T. Brünger, J. Kuriyan, M. Karplus, *Science* **235**, 458 (1987).
49. A. T. Brünger, *X-PLOR Manual*, version 3.1 (Yale Univ. Press, New Haven, CT, 1992).

50. We thank M. I. Simon and P. B. Dervan for providing Hin peptide, S. Finkel for help with DNA synthesis and purification, K. Zhang for many helpful discussions on phasing, and D. S. Goodsell for help with schematic drawings. Supported by NIH Program Project Grant GM-31299 (R.E.D.) and by GM-38509 (R.C.J.).

17 August 1993; accepted 15 December 1993