DNA RECOGNITION BY PROTEINS WITH THE HELIX-TURN-HELIX MOTIF

Stephen C. Harrison and Aneel K. Aggarwal

Department of Biochemistry and Molecular Biology, Howard Hughes Medical Institute, Harvard University, Cambridge, Massachusetts 02138

KEY WORDS: transcriptional regulation, DNA-binding proteins, X-ray crystallography, repressors

CONTENTS

INTRODUCTION ................................................................. 933
NOTIONS ABOUT RECOGNITION ........................................... 934
THE HTH MOTIF .............................................................. 936
STRUCTURES ................................................................. 937
  Phage 434 Repressor and Cro ........................................... 937
  Lambda Repressor and Cro ............................................. 949
  trp Repressor .......................................................... 954
  lac Repressor DNA-Binding Domain .................................. 958
  CAP ........................................................................... 959
  The Antp Homeodomain ................................................ 961
COMMON FEATURES ......................................................... 962
  Families of HTH Domains ................................................. 962
  Characteristics of Protein/DNA Contacts .............................. 963
  Anchoring the HTH to DNA ............................................. 963
RECOGNITION .................................................................. 965

INTRODUCTION

Proteins involved in transcriptional regulation recognize specific DNA sequences through the properties of discrete DNA-binding domains. These do-

1Current Address: Columbia University, 630 West 168th Street, New York, New York 10032.
mains are in general relatively small. Several kinds of structures have been recognized: The most thoroughly studied contain a helix-turn-helix (HTH) motif (1-4). HTH-containing domains vary in size from about 60 to 90 amino-acid residues. This review describes the properties of these domains as seen in a number of prokaryotic and eukaryotic regulatory proteins and analyzes the way in which they recognize cognate DNA-binding sites (for other reviews, see 4-7). Most of the work reviewed involves X-ray crystallographic studies of phage and bacterial repressors and activators, alone and in complex with DNA. The components of the phage 434 "genetic switch" are the most thoroughly studied (8-14), and we use those structures as a reference. The determinations by 2-D NMR of the structures of the lac repressor DNA-binding domain (15) and of a homeodomain (16, 17) confirm earlier proposals (2, 3, 18) that these proteins belong to the HTH class and demonstrate that complete structure analysis by 2-D NMR is feasible for small domains of this type.

The article is organized as follows. We briefly mention some notions about DNA recognition, which have been described more extensively elsewhere (4, 19-21). We next outline the features of the HTH motif as seen in known structures of proteins and protein/DNA complexes. Each group of structures is then described in turn, with emphasis on how the proteins recognize their specific binding sites. In each section, the DNA target sites and their characteristics are mentioned first, as well as experiments that determine "what is recognized," followed by a description of the structures themselves and an analysis of the extent to which the structures explain what is recognized. At the end of the review, we summarize common features of HTH motifs and of the way they contact DNA, and we discuss the extent to which the notions outlined at the beginning are useful for understanding the structural results.

NOTIONS ABOUT RECOGNITION

That proteins can recognize particular DNA sequences was experimentally shown by the isolation of the lac (22) and lambda (23) repressors and by the demonstration that their operators are distinct segments of DNA (24, 25). These proteins also bind to nonoperator sequences, but with roughly 10^5 lower affinity. The structural basis of such preferential binding is the subject of this review. We outline first some frequently discussed notions about the molecular details of DNA recognition.

1. Contacts to functional groups of bases in the major groove (19). The edges of base pairs expose hydrogen-bond donor and acceptor groups in the major and minor grooves. There is also a nonpolar surface in the major groove near C5 of pyrimidines, where a projecting methyl group distinguishes thymine from cytosine. The major-groove hydrogen-bond patterns are distinctive
for each of the four base pairs (26). The minor-groove hydrogen-bond patterns are not distinctive. Much attention has focused on how a "recognition matrix" of major-groove hydrogen-bonding sites might be complemented by appropriate groups on the protein (27, 28). The phrase "specific interactions" is therefore often loosely used to refer especially to protein/base-pair hydrogen bonds. It is clear, however, from the structures described here that although hydrogen bonds to base pairs are very important, they are not the only interactions determining specificity.

2. A protein side-chain/DNA base-pair code? Particular amino-acid residues can form patterns of hydrogen bonds that uniquely specify a particular base. For example, the bidentate hydrogen-bonding interaction that may take place between an appropriately oriented glutamine and adenine, with the side chain -NH2 donating a hydrogen bond to N7 and the side chain -CO accepting one from N6, cannot be formed with any other base (26). Likewise, arginine can donate hydrogen bonds to both N7 and O6 of guanine (26). Is there a unique way, or a small number of recurring ways, for recognizing a particular base pair or a short sequence of base pairs? Such a possibility would correspond to a (perhaps degenerate) "recognition code." The known structures rule out simple versions of this notion.

3. A "recognition helix." The protruding helix of the HTH motif is now known to lie in the major groove, with several of its side chains touching DNA bases. The notion that specificity might be largely or entirely a property of residues on this helix is the basis of so-called "helix swap" experiments—efforts to alter the specificity of a protein by exchanging the exposed residues on the second HTH helix to match those of another HTH protein (29, 30). Such an experiment should work if the two proteins bind DNA in very similar register and if the sole or principal specificity determinants are interactions from residues in the second HTH helix. The swap of this helix alone appears to work only in special cases.

4. DNA conformation and "indirect readout." B-DNA exhibits significant conformational variability (31–33). Crystal structures of double-stranded DNA oligomers show various correlations between the path of the sugar-phosphate backbone and the local base sequence. The relationship between the two appears to be established by conformational adjustments that optimize base stacking (31, 33). Energetically favorable stacking interactions can lead to significant noncoplanarity of base pairs, especially through propeller twist. In effect, a compromise is reached between optimal hydrogen bonding and maximal base overlap. GC base pairs generally have low propeller twist, since coplanarity is favored by the presence of three Watson-Crick hydrogen bonds. AT base pairs can have high or low propeller twist. Inter-basepair hydrogen bonds in the major or minor groove can stabilize high propeller twist in a sequence of base pairs (34–36). For example, a major-groove
inter-basepair hydrogen bond can occur whenever two adenines occur in sequence along a strand. The resulting diagonal pattern of bifurcated hydrogen bonding seen in tracts of AT base pairs appears to stiffen such sequences (34, 35). A systematic analysis of other potential inter-basepair hydrogen bonds is not yet available. Another correlation is between minor-groove width and base-pair composition: AT regions generally have a narrower minor groove than GC regions.

Most of the proteins described in this review bend DNA to some extent. A sequence dependence of DNA bendability is indicated by studies of nucleosome “phasing” (37, 38). A-T steps appear to favor bending toward the minor groove, whereas G-C steps favor bending toward the major groove (39).

The conformational properties of DNA thus depend on base sequence, and proteins could in principle recognize sequence by its consequences for the positions of atoms in the sugar-phosphate backbone (20). The phrase “indirect readout” has sometimes been used to refer to this mechanism (40).

THE HTH MOTIF

The conventional designation of the HTH (4) is a twenty-residue segment corresponding to particular residues in the first three HTH structures solved (41–43). The assignment is arbitrary, since one or more residues at the N-terminal end of the motif may contribute to the first helix, and several additional residues at the C-terminal end (e.g. in trp repressor and CAP) may contribute to the second. Moreover, designation of residues at the ends of helices as helical or nonhelical can be imprecise unless the structure is refined at very high resolution. But the conserved three-dimensional structure is well represented by the central twenty residues chosen as indicated, and we use a numbering scheme that refers to this module (Figure 1). The HTH is properly described as a “motif” because it is a recurring substructure seen in otherwise different, cooperatively folding domains and because it is not in general a stably folded unit on its own. Residues 1–7 form the first helix, and 12–20, the second, but by some criteria 11 is “helical,” as is the residue preceding 1 in 434 and lambda repressors. Residues at positions 4, 8, 10, 16, and 18 are usually hydrophobic: They define the interior of the two-helix elbow and require additional hydrophobic residues from elsewhere in the domain to complete an enclosed hydrophobic core. Residue 5 is usually Gly or Ala: Larger side chains disturb the geometry of the HTH, since they interfere with the peptide backbone at the N-terminus of the second helix. In trp repressor, residue 5 is a lysine, and the HTH elbow is indeed perturbed by its presence (44). Residue 9 is usually Gly, and its backbone conformation angles in several structures are reported to be in the region readily accessible only to glycine (12, 13, 45), but serine (4), cysteine (4, 18), and even glutamic acid
Figure 1  Amino-acid sequences of the HTH elements found in structures described in this review. The numbers at the beginning and end of each line are the residue numbers in the full protein sequence. Characteristic residues are shown in bold face. Note that the first HTH helix is in some cases extended at its N-terminus and the second HTH helix is in some cases extended at its C-terminus. For references, see sequence compilations in (4) and (122).

(46) are found at this position in some HTH proteins. The HTH is, of course, further defined by the absence of prolines at internal positions in the two helices. Nonetheless, the properties just outlined are insufficient to specify the motif uniquely, and some sequences in the data base that adequately obey this consensus do not represent HTH elements.

Because the two HTH helices are assigned different numbers or letters in different proteins, we refer to them here as the “first HTH helix” and “second HTH helix” respectively (Table 1). The second HTH helix is also frequently called the “recognition helix.” We avoid this name here because one of our major conclusions is that the second helix is not the only participant in what might most properly be regarded as recognition. Nonetheless, structures of HTH-protein/DNA complexes have borne out the model, first proposed on the basis of the lambda Cro and CAP structures (1), that the second HTH helix should lie in the major groove and contribute important base-pair contacts.

STRUCTURES

The known HTH-containing structures (September, 1989) are reviewed in the pages that follow.

Phage 434 Repressor and Cro

OPERATOR SPECIFICITIES The temperate bacteriophages related to lambda have two major operators, O_R and O_L, that determine the switch between lysogeny and lytic growth (47). Each operator has three sites, all of which bind both repressor and Cro (48–51). A critical property of the two proteins is that in O_R the repressor binds most tightly to O_R1 and least tightly to O_R3.
Table 1 Designations of the first and second HTH helices in the structures described in this review

<table>
<thead>
<tr>
<th>Protein</th>
<th>First HTH helix</th>
<th>Second HTH helix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lambda repressor</td>
<td>Helix 2</td>
<td>Helix 3</td>
</tr>
<tr>
<td>Lambda Cro</td>
<td></td>
<td></td>
</tr>
<tr>
<td>434 repressor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>434 Cro trp</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAP</td>
<td>E</td>
<td>F</td>
</tr>
<tr>
<td>lac repressor</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Antp homeodomain</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

while Cro has opposite relative affinities (47). The 434 operator sites (51) are usually designated as 14-base-pair units with consensus sequence 5'-ACAAxxxxxxxTTGT (Table 2). The only exception to this consensus is in OR3. Affinities of 434 repressor, its DNA-binding domain (designated R1-69), and 434 Cro have been measured for each of these sites (51), as well as for the symmetrical site labelled "14mer" in Table 2 and for a large series of systematic variants of the 14mer sequence (52; G. B. Koudelka et al, unpublished results). The effects of substitutions into the 14mer sequence may be summarized as follows: (i) Variations in the conserved 5'-ACAA at base pairs 1–4 strongly diminish repressor (and R1-69) binding. Cro binding is equally sensitive to changes at base pairs 1–3, but it is indifferent to a substitution of G for A at base pair 4. This property is clearly important for preferential binding of Cro to OR3, which has 5'-ACAG in one half-site. (ii) Both proteins respond similarly to variations in base pairs 5–7, but the effects are at least tenfold smaller than those of changes at base pairs 1–4. In general, AT or TA base pairs at positions 6 and 7 lead to tighter binding than GC or CG base pairs (52). Certain sequences, such as a row of five adenines from 6L to 5R, are particularly favorable (60). That is, the influence of substitution at individual positions near the center of the operator may correlate with the identity of neighboring base pairs. (iii) The effects of flanking base pairs -2 and -1 have not been studied as thoroughly, but there appears to be a modest preference for TA at -1. (iv) R1-69 exhibits essentially the same binding preferences as intact repressor, but the effects of central base-pair substitutions are slightly smaller (51, 52; G. B. Koudelka et al, unpublished results).
Table 2  Operator sites of bacteriophage 434 and normalized dissociation constants for R1-69 and 434 Cro<sup>a,b</sup>

<table>
<thead>
<tr>
<th>Site</th>
<th>Sequence</th>
<th>R1-69</th>
<th>Cro</th>
</tr>
</thead>
<tbody>
<tr>
<td>OR1</td>
<td>TACAGA</td>
<td>1</td>
<td>2.5</td>
</tr>
<tr>
<td>OR2</td>
<td>AAGATACATGTAT</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>OR3</td>
<td>CACACTGTTCTACTGTA</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>OL1</td>
<td>TACAGAGGAGATTGTATA</td>
<td>2.5</td>
<td>2</td>
</tr>
<tr>
<td>OL2</td>
<td>AATAATATTTATGATACT</td>
<td>2.5</td>
<td>3</td>
</tr>
<tr>
<td>OL3</td>
<td>AACATGGATTGTATCAACA</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>14mer</td>
<td>AATTATATATTACACAG</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

<sup>a</sup>Consensus ACAA element is shown in boldface.

<sup>b</sup>From Refs. 51, 52; G. B. Koudelka, unpublished results.
Correlations between operator specificity and amino-acid substitutions in the proteins further define the important interactions. For example, the altered protein known as repressor* has residues 32, 33, 34, 36, and 38 of 434 repressor (5, 6, 7, and 9 of the second HTH helix and a residue two positions after the helix) changed in a so-called “helix swap” to those found at the corresponding positions in 434 Cro (29). Repressor* has a significantly reduced affinity for OR1 and a slightly reduced affinity for OR3, which creates a more “Cro-like” preference. A more dramatic altered specificity is obtained when residues 27, 28, 29, 32, and 36 are changed to their P22 repressor homologs (30). The relative affinities of the hybrid molecule for the three sites in P22 OR also follow those of the normal P22 repressor. The results demonstrate that P22 and 434 repressors bind DNA in nearly identical ways and that the mutated residues can in this context account for many of the important contacts.

Attempts to obtain altered specificity through more restricted changes in amino-acid residues have been largely unsuccessful. For example, random mutagenesis of the first two residues in helix 3 of 434 Cro yields no altered proteins capable of specific binding to a set of twelve mutant 434 operators that were obtained by introducing into the 14mer sequence all possible single-position symmetrical changes at base pairs 1–4 (51). By contrast, a 434 repressor with alanine at residue 28 does bind specifically to the 14mer with TA instead of AT at position 1 (54). The new interaction appears to involve the thymine methyl group, since the Ala 28 species also has a weak specificity for (5MeC)G but none for UA. A corresponding altered specificity cannot, however, be generated by the homologous change in 434 Cro.

STRUCTURES OF R1-69 AND 434 CRO The two proteins R1-69 and 434 Cro are very similar, both in sequence and in folded structure (12, 13; R. P. Wharton and G. Lauer, unpublished results). Residues 3–71 of 434 Cro align with 1–69 of repressor, and a numbering scheme for 434 Cro beginning with −1 was therefore adopted in order to have identical numbers for homologous locations. There are five α-helices, of which the second and third form the HTH motif (see Figure 2a,b). The first and fourth helices pack against the HTH to enclose a hydrophobic core. The proteins are monomeric in solution, and they do not dimerize even at the high concentrations needed for crystallization (12, 13). They nonetheless have significant dimer contacts when bound to DNA (9–11, 14), and independent binding of a monomer to a half-site is never observed (51). There is no paradox in these observations: The free energy at a protein interface necessary to create a dimer in solution must be at least 15–20 kcal/mol in order to overcome loss of translational and rotational degrees of freedom (56), whereas the free energy of interaction needed for highly cooperative binding is significantly less. Dimer interactions
on DNA involve residues in helices 4 and 5 and in the helix 3-4 loop (9-11, 14).

RI-69/OPERATOR COMPLEXES  Crystal structures have been determined for RI-69 in complex with various synthetic DNA operators: a 14-base-pair fragment with the symmetrical 14mer sequence (9), and 20-base-pair fragments with the OR1 and OR3 sequences (11, 57) (Figure 3). The DNA fragments stack end-to-end in crystals, forming a pseudo-continuous DNA helix (58). All 20-base-pair fragments have a 5' "overhang," designed to provide a base pair between adjacent fragments in a stack (11, 59). Complexes with these fragments have been refined at 2.5-Å resolution, and work is in progress to extend the resolution to 2.1 Å.

The 14mer and OR1-containing 20mer have very similar conformations in complexes with RI-69 (9, 11). They are somewhat bent and significantly overwound at their centers. The same synthetic operators have also been studied bound to 434 Cro (10, 14). In those complexes, the two DNA fragments have very similar conformations to each other, but somewhat different from their conformations when bound to RI-69. Thus, the proteins appear to impose related but distinct conformations on operator DNA.

The conformation of OR1 in complex with RI-69 has been analyzed in detail (11). Interactions with repressor, described below, bend the DNA. The bending is not smooth: The operator is relatively straight at its center, and it bends symmetrically by approximately 12° at about three base steps out from the center. There is further bending near the periphery of the site on both sides. The minor groove is very narrow at the midpoint of the operator, and it opens up at the ends, opposite the positions where the major groove receives helix 3. The central compression is a consequence both of the overwinding and of a shift of phosphate 4' away from the major groove; the peripheral widening is a consequence of bending around the repressor. A pattern of bifurcated hydrogen bonds is permitted by the large propeller twist of several base pairs in the center of the OR1 fragment (11). This capacity to form bifurcated hydrogen bonds may compensate for distortions in Watson-Crick hydrogen bonding that are imposed on operator DNA by the bound repressor.

The repressor lies with helix 2 across the major groove, and helix 3 inserted into it roughly in the way proposed when the HTH was first discovered (Figure 3). Comparison of bound RI-69 with the free domain shows that no large-scale protein conformational changes occur on binding, but certain side chains that interact with DNA rearrange. For example, Gln 29 is extended in free RI-69 and folded back in the complex.

The OR1-repressor interface has matched molecular surfaces over an extended area (Figure 3). Most interactions between one RI-69 monomer and DNA occur within a single half-site. That is, each repressor subunit "sees"
one side of the operator. The protein anchors itself across the major groove with extensive contacts along two segments of the sugar-phosphate chain, one to either side of the groove. Hydrogen bonds from peptide -NH groups to nonesterified oxygens of DNA phosphates are remarkably prevalent in these contacts.

The detailed interactions between protein and sugar-phosphate backbone (Figure 4a) may be summarized as follows. We refer to the backbone of nucleotides -2 to 1 as "segment 1" and of nucleotides 4' to 6' as "segment 2."

1. In segment 1 (nucleotides -2 to 1), the DNA backbone lies against the amino terminus of helix 2, and five hydrogen bonds link the protein to the sugar-phosphate chain. The phosphate of nucleotide 1 accepts hydrogen bonds from the side chains of Gln 17 and Asn 36. The phosphate of nucleotide -1 accepts hydrogen bonds from the peptide -NH of Gln 17 and from the side chain of Arg 10; that of nucleotide -2, a hydrogen bond from Asn 16. (Arg 10 and Asn 16 are not shown in Figure 4a.) 2. In segment 2, the DNA backbone lies against the loop between helices 3 and 4. Three main-chain -NH groups of the protein form hydrogen bonds with phosphates 5' and 6'. The side chain of Arg 43 projects into the minor groove and probably stabilizes its compression. The main-chain -NH groups and the arginine side chains thus form a tight clamp across this DNA backbone segment. Parts of
Figure 3  The complex of RI-69 with a 20-base-pair DNA fragment (11). (left) Space-filling representation, emphasizing tight fit over an extensive van der Waals contact surface. (right) Schematic diagram showing α-helices as cylinders, DNA backbone as ribbon, and base pairs as rods. A few side chains are shown on one monomer. Numbers in the protein refer to the amino-acid sequence, and numbers on the DNA backbone refer to base-pair designation.

all residues between Lys 40 and Phe 44 contribute to van der Waals complementarity between the protein and the sugar-phosphate surfaces of nucleotides 5' and 6'.

There are direct contacts to conserved base pairs 1 to 4 in the major groove, and less extensive contacts to base pairs −1 and 5. Hydrogen bonds, between amino-acid side chains and DNA base pairs, and nonpolar interactions, such as the fit of a thymine methyl group into a suitable van der Waals pocket, appear to be equally important in determining specificity (Figure 4a). In general, a base pair may interact with more than one side chain, and any one side chain may contain two or even three base pairs.

Enumeration of base-pair contacts shows variety and complexity in the noncovalent interactions. At base pair 1 (AT), bidentate hydrogen bonding links adenine 1 to Gln 28. The glutamine side chain also lies against the methyl group of the thymine of base pair −1 (TA). At base pair 2 (CG), two hydrogen bonds join the guanine to the -NH₂ of Gln 29. At base pair 3 (AT), there are no hydrogen bonds to protein, but a well-formed van der Waals pocket, created by Thr 27 and Gln 29, receives the methyl group of the
Figure 4  (a) Detailed view of interactions between one R-69 monomer and part of an operator half-site (OR1, left). Amino-acid residues mentioned in the text are designated by single-letter code. With reference to the gallery of domains in Figure 2 and to the diagram in Figure 4d, the direction of view is from the right.

(b) Schematic illustration of the difference between DNA backbone conformations in complexes of OR1 with R-69 and 434 Cro. The R ("repressor") conformation is shown in solid lines; the C ("Cro") conformation, in dashed lines. Compare with Figure 4a. The closer approach of DNA backbone to the HTH turn in 434 Cro permits hydrogen bonds from the peptide -NH of residue 27 and from the side chain of Ser 30 to phosphate 4'.

(c) Detailed view of interactions between one lambda repressor subunit and part of an operator consensus half-site. Amino-acid residues mentioned in the text are designated by single-letter code. Redrawn from Ref. 72.

(d) View, in a direction approximately normal to the one in Figure 4(a)-(c), showing anchoring interactions to DNA backbone that are common to the 434 proteins and lambda repressor. Redrawn from Ref. 76.
thymine. At base pair 4 (AT), thymine forms a single hydrogen bond with Gln 33 and contacts the β-carbon of Gln 29 with its methyl group. The side-chain carbonyl of Gln 33 is also in van der Waals contact with the pyrimidine of base pair 5, in such a way that its position depends on whether there is a methyl group on C5.

Two features of the structure illustrate the correlation between backbone and base-pair interactions in the R1-69/OR1 complex. 1. The -NH2 group Gln 28 donates one proton to N7 of the adenine of base pair 1 and the other proton to the side-chain carbonyl of Gln 17. The -NH2 group of Gln 17 in turn donates a hydrogen bond to phosphate 1. 2. On the other side of the major groove, a tightly bound solvent molecule links the side-chain carbonyl of Gln 33 to phosphate 5' and to Ser 30.

The recently determined structure of an R1-69/OR3 complex (57) shows little difference from the OR1 complex in the conserved, left half-site, but an unexpected conformational change in the altered, right half-site. The DNA backbone around phosphate 4'R lies closer to the helix 2/helix 3 turn of the right-hand R1-69 monomer, with a corresponding difference in the way the protein rests against the DNA. The DNA conformation in the right half-site is actually more like what is seen in the complexes with 434 Cro (see below and Figure 4b). The result suggests that R1-69 is unable to impose upon the nonconsensus 5'-ACAG . . . sequence the same conformation it confers on the consensus 5'-ACAA . . .

**434 CRO/DNA COMPLEXES** Crystal structures have been determined for 434 Cro in complex with the same 14mer (10) and OR1 (14) fragments that were studied with R1-69. The Cro complexes show DNA conformations similar to each other but different from the R1-69 complexes in one important respect: the position of DNA backbone near phosphates 2' and 4' (Figure 4b). The DNA in the OR1 complex is bent, with the curvature concentrated near the end of the operator. The central 14 base pairs superpose well on the 14mer in complex with 434 Cro. In the report of the 14mer complex, the DNA was described as relatively straight (10). Indeed, the central 14 base pairs of the 20-base-pair OR1 fragment do appear straighter when the operator is bound to Cro than when it is bound to repressor. Analysis of a refined structure shows, however, that the overall bend in the two cases is very similar and that the significant difference is a local one around nucleotides 2' and 4' rather than a global change in bend or twist (Figure 4b). For convenience, we denote the conformation of this backbone segment in the repressor complex by R and in the Cro complex by C (Figure 4b). Bending of operator DNA by repressor, R1-69, and Cro can be detected by electrophoretic analysis (60). Recent electrophoretic experiments (G. B. Koudelka, personal communication) require modification of earlier conclusions: Repressor and Cro bend DNA to
equivalent extents, and the overall deviation of the DNA axis is significantly smaller than the one imposed by CAP. These results are therefore in good agreement with the crystal structures.

Noncoplanarity of base pairs is as much a characteristic of the OR1 complex with 434 Cro as it is of its complex with R1-69, but some of the details are different (14). A related pattern of bifurcated hydrogen bonds can be inferred.

The way in which a subunit of 434 Cro interacts with an OR1 half-site is, as expected, very similar to the binding of R1-69 (11, 14). The same two segments of sugar-phosphate backbone serve as anchor points for the protein (Figures 4a, b; 5). 1. In segment 1 (nucleotides -2 to 1), essentially homologous interactions occur. 2. In segment 2 (nucleotides 3' to 6'), hydrogen bonds are present, as in the repressor complex, from the peptide -NH groups to phosphates 5' and 6'. But there is also a hydrogen bond from the peptide -NH of lysine 27 (the last residue in the turn) to one of the nonesterified oxygens of phosphate 4', and a second hydrogen bond to the same phosphate from the side chain of Ser 30. These contacts, not present with R1-69, are possible with 434 Cro because the sugar-phosphate backbone lies closer to the helix 2/helix 3 turn (Figure 4b). Thus, the two very similar proteins bind across the major groove in fundamentally similar ways, but there are significant local differences between imposed DNA conformations that result from differences in the identity and conformation of various residues contacting DNA backbone. Not all of these residues lie in helix 3, and some lie outside the HTH element altogether.

OPERATOR RECOGNITION BY 434 REPRESSOR AND CRO How well do the structures account for the observed specificities of these proteins and their mutants for various operator sequences? The question may be divided into three parts. 1. Why does 5'-ACAA at each end of a 14-base-pair sequence specify a 434 operator? 2. Can we understand the way in which the central six base pairs (5L to 5R) modulate affinity of a site for each of the two proteins? 3. Why is 434 Cro indifferent to a substitution of G (but not of C or T) for A at position 4?

1. The pattern of hydrogen bonds and van der Waals contacts made by Gln 28 and Gln 29 to base pairs 1–3 (Figure 4a) can account for specificity at 5'-ACA (9, 11). The role of Gln 29 is interesting because it interacts with both base pairs 2 and 3. Its hydrogen bond to guanine specifies CG at base pair 2, and its aliphatic chain forms, together with Thr 27, a good van der Waals pocket for the methyl group of the thymine in AT base-pair 3. A GC would create a nonpolar hole by the absence of the 5-methyl substituent, and a TA or CG would place a presumably hydrated N7 opposite a concave hydrophobic surface. At base pair 4 (AT), Gln 29, Ser 30, and Gln 33 all interact with the thymine—the first two through van der Waals contact with
the methyl group, the last by a hydrogen bond. Comparison of the R1-69/Or1
and R1-69/Og3 complexes suggests an additional source of specificity: the
influence of base-pair substitutions at position 4 on the ability of an operator
to adopt the R conformation. This mechanism can also help explain why 434
repressor and Cro respond differently to an AT to GC change (see below).

Direct contacts to DNA bases from residues on helix 3 of 434 Cro are
similar to the homologous contacts from R1-69 (14). Gln 28 and Gln 29 have
essentially identical interactions in the two complexes, so that recognition of
base pairs 1–3 by these conserved residues is the same in the two proteins.
Leu 33cro, the homolog of Gln 33R1-69, contributes to a van der Waals pocket
for the methyl group of the thymine in base pair 4 (AT).

2. The central base pairs must modulate operator affinity through their
effects on the relative free energies of different DNA conformations, because
the proteins have essentially no direct interactions with these bases (9–11,
14). Properties of appropriate mutants show that the two exceptions (a van der
Waals contact between Gln 33 of R1-69 and the pyrimidine of base pair 5, and
water-mediated hydrogen bonds between Arg 43 and base pair 7 in the minor
groove) cannot account for the way base pairs 5, 6, and 7 influence repressor
binding (11, 52). The base-pair noncoplanarity and non-Watson-Crick hydro­
gen bonding seen in the center of the operator can explain many of the
observed affinities (11). Noncoplanarity appears to be an adjustment of the
base pairs to a conformation imposed on the double helix by binding of R1-69
or 434 Cro. Sequences that can compensate for strain with bifurcated hydro­
gen bonds or with improved base stacking will bind the proteins more tightly.
It is really the binding of the repressor dimer that constrains DNA between
base pairs 6L and 6R; the protein dimer contact thus indirectly determines
discrimination among central base-pair sequences. The strain imposed by
dimer binding is probably best described in terms of the detailed con­
formational features of bound operator DNA, rather than by reference to
long-range properties like resistance to twist or bend, measured as an average
over many base pairs. In particular, the suggestion that “overwinding” might
be the principal contribution to conformational strain at the operator center
(60) is probably too imprecise, given the range of estimates for base­
composition dependence of DNA torsional rigidity (61, 61a).

3. The indifference of 434 Cro to a CG substitution at base-pair 4 could in
principle be either a local effect, due only to the changes in residues that
contact base pair 4, or a more distributed one involving conformational
differences between the two complexes. A purely local model can be ruled
out. The one change in a residue making direct contact with base pair 4 is at
position 33: Gln in 434 repressor, Leu in 434 Cro. Substitution of leucine into
position 33 of 434 repressor does not alter its response to the AT to CG
change at base pair 4, nor does substitution of glutamine at position 33 of 434
Cro create a preference for AT (G. B. Koudelka et al, unpublished results).
Since a key shift between R and C conformations of bound DNA is centered just at the nucleotide 4', it is appropriate to suggest that the difference in imposed conformations is part of the mechanism for differential sensitivity (10, 14). The R1-69/O_{R3} structure appears to strengthen this conclusion by showing that repressor cannot impose the R conformation on 5'-ACAG and hence that the conformational response is sequence dependent (57). An analysis of contacts in the Cro/O_{R3} complex will be needed before this suggestion can be evaluated properly.

Lambda Repressor and Cro

Operator Specificities  The six 17-base-pair lambda operator sites are listed in Table 3, together with their relative affinities for lambda repressor and Cro (50, 62). Note that four of the six sites share a "consensus" half-site. Unlike the 434 operators, conserved and variable base pairs are intermingled. Positions 2 and 4 are completely invariant in all twelve half-sites, and position 6 deviates from consensus on only one side of O_{L3}. Positions 3 and 7 are the most variable. The influences of changes in a symmetrical consensus operator site on repressor and Cro affinity in vivo have been determined (63, 64). Any change at base pairs 2, 5, 6, and 7 causes a large reduction in repressor binding, as does a purine at base pair 4 or a cytosine at base pair 8 (63). A nonconsensus C at base pair 3 actually enhances repressor affinity. Lambda Cro has somewhat different preferences: Any substitution at base pair 2, 4, 5, or 6 produces a large reduction in binding; so does C or G at base pairs 1 or 3, or a purine at base pair 7. Measurement of Cro binding to synthetic operators in vitro confirms this pattern of specificities (65). Strong effects correspond to increased binding free energies of about 2 kcal/mol for unilateral changes in O_{R1} and about twice this value for symmetrical substitutions. Taken together, the results of such systematic measurements are consistent with the effects on repressor and Cro binding of changes introduced into O_{R3} at the three positions where it differs from O_{R3}: The affinity of lambda repressor is more sensitive to the difference at base pairs 5 and 8 than it is to the difference at base pair 3, whereas the affinity of lambda Cro is more sensitive to the change at base pair 3 (66).

Site-directed changes in lambda repressor and Cro have been used to associate particular residues with specificity for certain base pairs (66, 67). If alanine is substituted for the conserved serine at position 2 of the second HTH helix in either protein, methylation of the guanine in base pair 4 no longer interferes with binding (67). Moreover, discrimination against substitution at this base pair is reduced while sensitivity to substitutions at base pairs 3 and 5 is unaffected. Residues at positions 5 and 6 in the second HTH helix appear to be critical in enabling the proteins to distinguish between O_{R1} and O_{R3} (66). These residues are Gly-Ala in repressor and Asn-Lys in Cro. Correlations
Table 3  Operator sites of bacteriophage lambda and normalized dissociation constants for lambda repressor and Cro\(^{a,b}\)

<table>
<thead>
<tr>
<th>Site</th>
<th>Sequence</th>
<th>Repressor</th>
<th>Cro</th>
</tr>
</thead>
<tbody>
<tr>
<td>OR1</td>
<td>T A C C T C T G G C G G G T G A T A</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>A T G G A G A C C G C C A C T A T</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1' L</td>
<td>1' L 2' L 3' L 4' L 5' L 6' L 7' L 8' L 9 8'R 7'R 6'R 5'R 4'R 3'R 2'R 1'R</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OR2</td>
<td>T A A C A C C C G T G C G G T G T T G</td>
<td>50</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>A T T G T G G C A C G C A A C A A C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OR3</td>
<td>T A T C A C C C G C C A A G G G A T A</td>
<td>50</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>A T A G T G G C G G T T C C C T A T</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OL1</td>
<td>T A T C A C C C G C C A G T G G T A</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>A T A G T G G C G G G T C A C C C A T</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OL2</td>
<td>C A A C A C C G G C A G A G A T A</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>G T T G T G G C C G C G T C T C T A T</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OL3</td>
<td>T A T C A C C C G C A G A T G G T T</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>A T A G T G G C G G T C T A C C A A</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Consensus bases are shown in boldface.

\(^{b}\)From Refs. 50, 62.
between reciprocal amino-acid substitutions and specificity at various base pairs suggest that the asparagine at position 5 of the second HTH helix in Cro interacts with base pair 3, that the lysine at position 6 of the second HTH helix in Cro interacts with base pair 6, and that the alanine at this position in repressor may have a role in specificity at base pair 5 (66). Deletion of the N-terminal arm eliminates selectivity at base pair 8 and exposes a normally protected guanine to methylation (66, 68, 69).

STRUCTURES OF LAMBDA REPRESSOR N-TERMINAL DOMAIN AND LAMBDA CRO

Unlike their 434 counterparts, these two proteins are quite different from each other outside the HTH elements (see Figure 2 c, d). The N-terminal DNA-binding domain of lambda repressor contains the first 92 residues (70, 71). A 5-residue N-terminal arm, disordered in the free protein, wraps around the double helix when the repressor binds DNA (68). There are five α-helices (43). The first four correspond rather closely to the first four helices in 434 repressor and Cro, with a 4-residue “insertion” between helices 1 and 2 and an additional residue between helices 3 and 4 (13). The fifth helix of lambda repressor has no counterpart in 434 repressor. Its residues have extensive interactions with residues in the twofold related helix 5, and the dimer seen in crystals of the free domain faithfully represents the dimer as bound to DNA (43, 72).

Lambda Cro contains three helices packed against a small three-stranded β-sheet (41). The sheet contains the N-terminal residues and a large hairpin following the HTH motif. There is an extended C-terminal arm. The protein has a reasonably extensive dimer contact in the crystal, and since the protruding third helices lie about 35 Å apart in an orientation appropriate for insertion into the major groove, it was assumed that this dimer would represent the form bound to DNA (73). In fact, significant flexion of the contact occurs on binding, as shown by recent studies of co-crystals with a DNA fragment (21).

When one compares one subunit of lambda Cro with an N-terminal domain of lambda repressor, conserved aspects of the structures are the precise course of the polypeptide chain backbone in the HTH, the glutamine at the beginning of helix 2, and the Gln-Ser sequence at the beginning of helix 3 (74). Helix 1 is much shorter than in lambda repressor and somewhat differently oriented, but the side-chain packing against residues of helix 3 is roughly homologous.

THE LAMBDA REPRESSOR/O1L1 COMPLEX

A crystal structure has been determined and refined at 2.5-Å resolution for the N-terminal domain of lambda repressor in complex with a 20-base-pair fragment containing the O1L1 sequence (59, 72). The results confirm some features predicted by earlier model building from the structure of the protein dimer (75); they also show many unanticipated interactions. The DNA is slightly bent, but less so than in
the 434 structures. The twist and other helical parameters appear to vary less markedly than in the 434 repressor and Cro complexes, and the base-pair planarity is more regular.

The individual lambda repressor domains are anchored both by contacts to the sugar-phosphate backbone and by the way the N-terminal arm wraps around the double helix in the major groove (Figure 4c). The sugar-phosphate contacts are similar to those described for the 434 repressor and Cro and occur on both sides of the major groove (72).

The details of the interactions between lambda repressor and DNA backbone can be summarized as follows.

1. In segment 1 (nucleotides -1 to 1), the DNA backbone lies against the amino terminus of helix 2. Hydrogen bonds link the main-chain -NH of Gln 33 and the phenolic hydroxyl of Tyr 22 (not shown in Figure 4c) to phosphate 1. These residues are homologs of Gln 17 and Arg 10 in 434 repressor, which make similar contacts. Moreover, the similarity extends to the way the side chain of Gln 33 in lambda repressor stretches across the ribose of nucleotide 1 to form hydrogen bonds with phosphate 2 and with Gln 44, as well as to the way Asn 52 (from the C-terminus of helix 3) hydrogen bonds to phosphate 2 (72, 76).

2. In segment 2 (nucleotides 5' to 7'), the peptide -NH of Gly 43 (the last residue in the turn) donates a hydrogen bond to phosphate 5'. Hydrogen bonds to phosphate 6' and 7' come from side chains of Asn 61 and Asn 58 respectively.

3. The N-terminal arm appears to be somewhat disordered, especially on the nonconsensus half-site, but its general course is clear. The arm passes into the groove just beyond the C-terminus of helix 3 and extends up to, but not beyond, the center of the operator.

Lambda repressor contacts DNA base-pairs throughout most of the half-site (Figure 4c). Gln 44 forms two hydrogen bonds with the adenine in base pair 2 (AT), in the bidentate pattern described above for 434 repressor and Cro, and its side chain also lies against the methyl group of the thymine in base pair 1 (TA). Several groups contribute to a hydrophobic surface opposite the methyl of the thymine in base pair 3 (AT) of the consensus half-site, but the repressor is not sensitive to all substitutions at this base. Ser 45 donates a hydrogen bond to N7 of the guanine in base pair 4 (CG), and its β-carbon forms part of a hydrophobic contact for the methyl group of the thymine in base pair 5 (AT). Asn 55, three residues beyond the end of helix 3, and lysine 4, in the N-terminal arm, form hydrogen bonds with the guanine in base pair 6 (CG) of the consensus half-site. There is also contact from a residue in the arm to base pair 8.

THE LAMBDA CRO/OR3 COMPLEX Crystallographic analysis is in progress, at about 3.5-Å resolution, of a complex between lambda Cro and a 17-base-pair fragment containing the OR3 sequence (21). A model has not yet been
published, but a preliminary report indicates that the relationship between subunits within a Cro dimer is altered with respect to what is seen in crystals of the free protein (21). The change corresponds to a rotation of approximately 35° about an axis roughly parallel to the DNA (21; B. W. Matthews, personal communication). The DNA is reported to be bent, and the minor groove compressed near the midpoint, despite the presence of a number of GC base pairs across the operator center (21). Thus lambda repressor and lambda Cro may fix their operators in conformations even more different than those imposed by 434 repressor and 434 Cro.

OPERATOR RECOGNITION BY LAMBDA REPRESSOR AND CRO
The only high resolution-structure of a complex determined so far is of the lambda repressor with O_L1 (72). We can ask at this stage: 1. How does this structure account for observed repressor specificities? In particular, can it explain the negative effect of any substitution at base pairs 2, 5, 6, or 7, or of a purine at base pair 4, and can it explain why base pairs 5 and 8 dictate differential recognition of O_R1 and O_R3? 2. What inferences can be made about recognition by lambda Cro? Note that the overall regularities of lambda operators—conservations at base pairs 2, 4, and 6 and strong preferences at base pairs 1, 5, and 8—must summarize the joint properties of the two proteins.

1. The interactions and conformation of Gln 44 appear to determine both the conserved AT at base pair 2 and the preference among naturally occurring operators for TA at base pair 1 (72). The "immobilization" of the glutamine side chain by the hydrogen-bond network linking Gln 44, Gln 33, and phosphate 2 is probably important, because mutation of Gln 33 reduces operator affinity (78). A hydrogen bond from Ser 45 to N7 of the guanine in base pair 4 (consensus CG) can account for the required pyrimidine:purine, because adenine can participate in the same interaction. At base pair 6, recognition of a conserved CG involves correlated hydrogen bonding from two residues, Lys 4 and Asn 55, neither of which lies in helix 3. Homologous residues are absent in lambda Cro, which lacks an N-terminal arm and which has a different arrangement of the polypeptide chain following helix 3 (41). Lambda Cro is believed instead to interact with base pair 6 through the side chain of Lys 32, the sixth residue of helix 3 (66). The corresponding residue in lambda repressor is an alanine, which is too small to make a contact directly (72). The involvement of Lys 4 can also explain why the effects of changes at base pair 6 are weaker for "armless" repressor (66). At base pair 5, a nonpolar contact between Ser 45 and the methyl group of thymine may give some selectivity. O_R1 and O_R3 depart in different ways from consensus in one half-site at this position, and this difference is thought to be critical for discrimination by repressor (66). The distinction between TA (O_R1) and CG (O_R3) at base pair 5 cannot readily be inferred from the O_L1 structure,
however, which has the consensus AT in both half-sites. The N-terminal arm of lambda repressor contacts base pair 8 in the major groove (68, 72). The details are not clear in the structure as presently resolved, but it is reasonable to assign discrimination at base pair 8 to this interaction. The significance of base pair 7 is not evident from the structure as presently described.

2. An early model for a lambda Cro/operator complex was based on the assumption that the Cro dimer bound in a conformation identical to the one seen in crystals of the free protein (73). This constraint was inconsistent with placing its HTH motif in a position similar to the one found for the HTH of lambda repressor. Since the constraint appears not to apply, the question of whether conserved parts of lambda repressor and Cro make homologous DNA contacts remains unresolved. A number of lines of evidence currently favor homology. Conservation not only of the Gln-Ser sequence in helix 3 but also of the Gln at the N-terminus of helix 2 endows lambda Cro with precisely the same configuration seen in the 434 proteins and in lambda repressor for recognizing the adenine in base pair 2 (AT) and for contacting the thymine in base pair 1 (TA) (75). Moreover, the properties of a mutant Cro, with alanine instead of serine at position 2 of helix 3, are consistent with a contact at base pair 4; so is the sensitivity of this mutant to methylation of the guanine in that base pair (67). The simplest explanation for the very different ways that lambda repressor and Cro respond to changes in base pair 3 is direct interaction of the asparagine at position 5 of Cro helix 3. This view is supported by the observation that substitution of asparagine for glycine at this position in repressor confers Cro-like base-pair-3 specificity (66).

**trp Repressor**

The trp repressor of *Escherichia coli*, a homodimer of 12.5-kd subunits (79, 80), binds to three different operators (79, 81–83). The sites have a central consensus sequence of 16 base pairs (see Figure 5). Conventional numbering is from the center outward. Binding to operators with symmetrical base-pair changes has been studied in vivo: Positions 6 to 3 (5'-CTAG) are most critical for specificity (84). In two of the naturally occurring operators, methylation protection and footprinting extend beyond the 16-base-pair consensus, and cooperative binding of more than one repressor has been proposed (83). The sequences are such that sites 16 base pairs in length could overlap by 8 base pairs, which would lead to an interdigitated arrangement of bound protein. A model of this kind accounts for protection of over 30 base pairs in the trp operator by a protein whose structure shows that it “covers” only 16 base pairs (83, 85).

Specific binding of repressor DNA requires the co-repressor L-tryptophan (80). The protein has therefore been studied crystallographically in several forms: inactive apo-repressor with no bound co-repressor (86), the active
Figure 5  Sequences of 434-, lambda-, trp-, and lac-operator half-sites, aligned according to the register established by structures of repressor/operator complexes. The common features of backbone anchoring interactions, described in the text, establish a correspondence between particular phosphates, contacted by homologous amino-acid residues. The contacts are shown as solid arrowheads, and the numbers of the phosphates are indicated. In general, most of these phosphates can be detected in ethylation interference experiments (130). The spacing between homologous contacts on the 5' strand (segment 1) and on the 3' strand (segment 2) is the same in 434, lambda, and lac operators. In trp operators this spacing is one base pair greater. The alignment of the trp operator shown here is based on contacts made by the second HTH helix of trp repressor: to phosphate 7 in segment 1 (hydrogen bond from Asn 87) and to phosphate 4' in segment 2 (hydrogen bonds from peptide -NH 78 and Thr 81). Operator dyads are shown by the filled symbol . The lac dyad is shown on the left, since the half-site has been aligned with respect to the orientation of the binding domain.

repressor with two L-tryptophans bound per dimer (44), and pseudorepressor with indole-3-propionate bound in place of tryptophan (87). The dimer has an unusual structure (44). Each subunit is a bundle of 6 helices, labelled A–F (see Figure 2f). Helices A and B of one subunit pack tightly against the 4-helix bundle formed by helices C–F of the other subunit, in such a way that the two monomers interlock. Helices D and E form the HTH. Helix E is one turn longer than helix 3 of 434 and lambda repressors. The positions of helices D and E with respect to the rest of the molecule and their relative orientations
with respect to each other change when co-repressor binds. The phrases "solid central core" and "flexible reading heads" have therefore been used to describe the molecule (88). The invariant core comprises helices A–C and F from both monomers. Helices D and E form an HTH unit, flexibly attached at each end. It should be emphasized that this flexibility is very restricted and in no sense comparable to the much greater flexibility of the interdomain hinges in lambda and lac repressors. L-tryptophan binds between the central core and the HTH, in a pocket made available by the presence of glycine instead of a large hydrophobic residue at HTH position 18 (44). It stabilizes the active conformation of the protein by positioning the HTH (86). Comparison of two different crystal forms of the active repressor with each other and with aporepressor shows that tryptophan binding produces a shift in the position of the HTH unit with respect to the core but that internal rearrangements are also possible within the HTH (89). The alternatives involve changes in the turn, resulting in a different angle between helices D and E.

The structure of a complex with operator DNA has been determined at high resolution (40). The DNA fragment is an 18-base-pair duplex augmented by a thymidine at either 5'-end. Its sequence mimics the trp EDCBA operator, except for a symmetrizing change at nonconserved position −8. The DNA is somewhat bent to conform to the repressor (40). The bending is rather sharply localized between base pairs 5 and 4 on either side of the dyad. The conformation of the repressor core is the same as seen in free repressor and aporepressor (44, 86). The conformation of the HTH units is somewhat different from what is observed in either repressor crystal form, however, suggesting that this element can adjust nearly continuously over a small range, even when tryptophan is bound (40).

Hydrogen bonds anchor the protein to segments of sugar-phosphate backbone on both sides of the major groove (40). We can define segments 1 and 2, as in our description of the 434 and lambda repressor complexes.

1. In segment 1 (nucleotides 9, 8, and 7 on the 5' strand), there are a total of 5 hydrogen bonds to nonesterified oxygens of 3 different phosphates. The N-terminal residue of the first HTH helix, Gln 68, donates hydrogen bonds to phosphate 9 from its peptide -NH and to phosphate 8 from its side-chain -NH2. Side chains of asparagine, serine, and lysine residues near the C-terminus of the second HTH helix donate hydrogen bonds to phosphates 7 and 8.

2. In segment 2 (nucleotides 4', 3', and 2' in the 3' strand), there are a total of 7 hydrogen bonds to nonesterified oxygens of 3 different phosphates. The peptide -NH of Gly 78 (the last residue in the turn) and the side-chain -OH of Thr 81 donate hydrogen bonds to phosphate 4'. Only serine can substitute for Thr 81 without abolishing operator binding (90). These interactions are very
similar to those made by -NH$_2$ 27 and Ser 30 of 434 Cro (Figure 4b). The indole nitrogen of the trp co-repressor, protruding from behind the second HTH helix, donates a hydrogen bond to phosphate 3'. A salt bridge from its α-carboxylate also helps anchor the side chain of Arg 84, the sixth residue of the second HTH helix, which in turn donates a hydrogen bond to phosphate 2'. Other hydrogen bonds to segment 2 are from residues outside the HTH. They link Arg 54 in helix C with phosphate 3' and Thr 44 in the BC corner with phosphate 2'. Mutations at 44, 54, and 84 all inhibit specific binding (90).

Only a few direct contacts are observed between a protein side chain and a base pair (40). Arg 69 (the second residue in helix D) forms bidentate hydrogen bonds with the guanine of base pair 9; Gly 78 (the last residue of the turn) lies with its α-carbon tucked against nonpolar edges of the adenine in base pair 5 and the thymine in base pair 4; the side chain of Ile 79 is in van der Waals contact with the thymine methyl group in base pair 8. Since base pairs 8 and 9 can be varied without effect (94) and since the contacts to base pairs 4 and 5 are not unique to particular bases, the interactions just listed cannot explain observed specificities.

Two additional sources of specificity have been suggested (40, 88).

1. The crystal structure shows solvent-mediated hydrogen bonds to base pairs 5, 6, and 7 from three tightly bound waters (40). One of these waters is coordinated by the N7 of adenine in base pair 5, the O6 of guanine in base pair 6, and the peptide -NH of repressor residue 80 (the second in helix E). Another is coordinated by the N7 of guanine in base pair 6, the peptide -NH of residue 79, and the side-chain -NH$_2^+$ of Lys 72. In both cases, the protein ligation requires that the water donate hydrogen bonds to the base. As a result, the guanine in base pair 6 is specified by this network, and so is the purine in base pair 5. These restrictions are consistent with the effects of base-pair substitutions on in vivo operator affinity (84). Base pair 6 must be GC, and base pair 5 can be TA or GC. A third water links Thr 83 and the adenine of base pair 7. All three waters intervene more directly between protein and bases than any identified so far in the phage repressors, and the first two are "caged" in the complex by Ile 79 and Ala 80. Moreover, repression is diminished by mutation at liganding residues 72 and 83 (40, 90). Changes at positions 79 and 80 that lead to modified specificity could do so either by displacing these water molecules or by forming alternative contacts.

2. Indirect readout has been invoked as a further mechanism (40). The DNA indeed appears to be deformed to fit the repressor surface. In particular, the relatively localized bend between base pairs 4 and 5 (a T-A step) may be sensitive to base sequence. The irregularities of the DNA structure are, however, less striking than in the 434 repressor complex. Moreover, in 434
represor, such effects account only for modulations of specific affinities among operator sites rather than for the difference between specific and nonspecific binding.

An additional factor—tandem, overlapping binding—must also be considered (93, 88). Sequences in trp and aroH operators are such that adjoining 16-base-pair consensus elements overlap by 8 base pairs. Repressors bound to these elements would be related by a twofold rotation about an axis between base pairs 4 and 5. That is, adjacent repressor molecules would approach DNA from directions about 90° apart along axes separated by 4 base pairs. The overall effect would be a left-handed helical “decoration” of operator DNA. The crystal structure of the trp repressor/operator complex permits such an arrangement: No elements of the central bound dimer would interfere with an adjacent dimer, interacting as just described (40). One contacted phosphate would be shared. There is good evidence for a similar mode of binding in the case of the met repressor (91). Moreover, even a single 16-base-pair met operator sequence appears to bind three repressor dimers, since a specific half-site interaction is sufficient, together with cooperative protein/protein contacts, to recruit the two flanking dimers (91). If such an arrangement were true of the trp repressor, recognition of base pairs 5 and 6 by the central dimer would be related symmetrically to recognition of base pairs 4 and 3 by the flanking dimer. These are just the base pairs that appear to be specified by water-mediated hydrogen bonds (40). Tandem, overlapping binding might also create further perturbations in the DNA conformation. Thus, the significance of tandem binding must be evaluated before the observed structure of the trp repressor/operator complex can fully be reconciled with available data on recognition.

lac Repressor DNA-Binding Domain

lac repressor binds to a site of approximately symmetric sequence, with its dyad on a central base pair (92). Deletion of this base pair actually increases affinity and shows that the elements that bind to either side of it must be flexibly connected (93, 94). A symmetrical “ideal” operator has been described (93, 95). Its sequence, included in Figure 5, can be taken as a reference. The conventional numbering is from the center outwards. The level of repression in vivo is quite sensitive to changes at base pairs 1–7 and weakly sensitive to changes at base pairs 8 and 9 (95).

lac repressor is a tetramer; its subunits contain 360 residues (96). The N-terminal domain, uneuphoniously known as the “headpiece,” can readily be cleaved from the rest of the protein (97). It is a monomer of between 51 and 59 residues, depending on the cleavage site chosen. It binds specifically to operator DNA (93, 98). Two-dimensional NMR has been used to determine the conformation of the 51-residue fragment (15). It contains three
helices, of which the first two are the HTH element and the third packs against it to complete a hydrophobic core (see Figure 2g). NMR studies of a 56-residue polypeptide in complex with a 14-base-pair half operator, together with docking computations, yield a low-resolution view of the way the domain binds DNA (99–101). Sufficient NOE cross-peaks can be identified between protein and DNA to show that the HTH lies against the major groove approximately as it does in the phage repressors, but that its orientation with respect to the operator dyad is reversed. That is, the N-terminus of the first HTH helix lies against a segment of DNA backbone proximal to the dyad. The local interactions, only roughly defined by the relatively low resolution of the work, are relatively similar to those made by homologous parts of the phage repressors. The first two residues of the second HTH helix lie opposite base pairs 2–5, and the sixth residue could contact base pair 6. The corresponding alignment of the lac and phage operators is shown in Figure 5. The same HTH orientation has been found by NMR in a complex of two lac-repressor DNA-binding domains with a symmetrical 22-base-pair operator (102). Data on altered specificity in vivo of mutant lac repressors are consistent with a “reversed” HTH orientation (95, 103) and support the assumption that the orientation of the domain seen by NMR correctly represents specific binding by intact repressor.

CAP

The catabolite gene activator protein (CAP) interacts with a span of DNA at least 18–20 base pairs long, upstream of a number of E. coli operons (104, 105). The sites vary significantly in sequence, in distance from the transcriptional start position, and in degree of approximate twofold symmetry (106, 107). An “idealized” 18-base-pair symmetrical sequence is xTGTGAx₆TCACAx; a consensus derived from eight sites is xTTGTGAXx-TxxxTCAXA. Ethylation of phosphates eleven and twelve nucleotides from the center of the site interferes with binding and indicates interactions beyond the 18-base-pair consensus region (106). The flanking sequences do influence affinity significantly (108). Contacts over so broad a span require DNA to bend around the protein. CAP indeed bends DNA when it binds, as shown by electrophoretic analysis and by a variety of other physicochemical methods (109–112). The degree of bending, as measured electrophoretically, depends on sequence elements separated by 10.5 and 16 base pairs from the binding-site dyad (108). If the DNA is bent as expected around a CAP dimer centered on the binding site dyad, then the first of these elements will face the protein with its minor groove, and the second will face the protein with its major groove. Dinucleotides that enhance bending in these elements indeed tend to be those believed from analysis of nucleosomal DNA to favor bends toward
the minor and major grooves respectively (39, 108). Models suggest that CAP bends DNA by nearly 180° (113, 114). Changes in the flanking sequences on one side of the dyad can reduce bending by up to 30°, with a corresponding tenfold reduction in affinity (108). Electrostatic interactions are believed to account for these effects (114).

CAP binds specifically to DNA only when cAMP is bound (115). The structure of CAP complexed with cAMP (but not DNA) has been determined and refined to 2.5-Å resolution (116, 117). The protein is a dimer of identical subunits, each of 209 amino-acid residues (118, 119). The polypeptide chain folds into two domains—an N-terminal “large domain” that binds cAMP and mediates dimerization and a C-terminal “small domain” that interacts with DNA (116). The cAMP site in the large domain lies near the dyad, and the nucleotide is believed to influence affinity for DNA by effecting an allosteric change in the dimer. The two small domains have somewhat different orientations with respect to their respective large domains. They are related by an approximate dyad (the rotation angle is 186°), but this twofold axis does not coincide with the much more exact large-domain twofold axis. The significance of this asymmetry is not yet understood. It may be influenced by the way crystal-packing forces constrain a flexible hinge, but it may also be built into the protein and related to the imperfect symmetry of its binding-site sequences. An individual small domain contains three helices (labelled D, E, and F) and a β-sheet (116). The second (E) and third (F) helices form the HTH (Figure 2e). Two of the four strands of the sheet form a turn between the first and second helices, and the remaining two, a hairpin following the third helix. Helix F is longer by about one turn than helix 3 in 434 and lambda repressors.

A model for a CAP/DNA complex (113) has been proposed (supplanting earlier suggestions of complementarity to left-handed DNA, now clearly inconsistent with experimental data). A reasonable stereochemical fit, contact over at least 20 base pairs, and electrostatic complementarity all require a bend in the DNA—at least 50° for a 20-base-pair span and closer to 180° for contacts beyond this region (113, 114). An assumption for the model building is conservation of the structure seen in the crystals of CAP-cAMP. Given the apparent flexibility of the inter-domain hinge, this constraint may be unnecessarily restrictive. Nonetheless, the model has several interesting similarities to the 434 and lambda complexes. The N-terminus of helix E, the first in the HTH, approaches one segment of sugar phosphate backbone, and the turn and N-terminus of helix D approach a segment on the opposite site of the major groove (113). These contacts anchor the HTH in a manner quite similar to the way the N-terminus of helix 2 and the turn and N-terminus of helix 4 anchor the HTH of the phage repressors. Indeed, helix D of CAP and helix 4 of 434 and lambda superimpose rather well when the HTH units are spatially
aligned, despite the difference in their relative positions in the sequences (see section on common features).

The Antp Homeodomain

Homeodomains are related, 60-residue elements, originally found in the products of a number of Drosophila developmental genes and subsequently in a number of mammalian transcription factors (120–122). Amino-acid sequence relationships led to the suggestion that these domains might contain an HTH motif (18). A number of classes have been recognized, based on distinct variations in sequence within a striking overall conservation (122). DNA sites specifically recognized by a variety of homeodomains have recently been identified (123–126). Domains of somewhat different amino-acid sequences appear to recognize similar sites, and a domain of given sequence can bind to quite unrelated sites. Consensus sequences from naturally occurring sites are 8–10 base pairs long (123, 124, 126, 127).

The structure of a 61-residue polypeptide, representing the homeodomain of the Drosophila gene Antennapedia (Antp), has been determined by 2-D NMR methods (16, 17). The presence of an HTH motif is indeed confirmed. The polypeptide chain folds as a 3-helix bundle (17). The second and third helices form the HTH. The former is longer by nearly one turn at its N-terminus than its counterpart in phage repressors; the latter is longer at its C-terminus. The first helix packs into the HTH elbow. Two residues universally conserved in all known homeodomains, Trp 48 and Phe 49 (numbering with respect to the standard homeodomain sequences), are important for this fit (17). They are at positions 7 and 8 of the second HTH helix; the homologous residues in 434 and lambda repressors have comparable structural roles. A short additional helical segment at the C-terminus of the domain is essentially an extension of the second HTH helix, joined at a modest angle (17). This extension contains two further universally conserved residues, Asn 51 and Arg 53. They are exposed at the surface, and it is plausible that they contact DNA backbone. Indeed, the angle between the second HTH helix and helical extension has led to the suggestion that the extension could follow the course of the major groove and augment the interaction surface (17).

The precise mode of DNA binding is, of course, not evident from the structure of the domain alone. In 434, lambda, and trp repressors, important anchoring points to sugar-phosphate backbone are found at the N-terminus of the first HTH helix, the end of the turn, and the C-terminus of the second HTH helix. With respect to these repressors, the N-terminally extended first HTH helix of the Antp homeodomain thus necessarily implies a somewhat different fit of its HTH on DNA. In other homeodomains, changes in the initial residues of the second HTH helix, important for recognition by 434 and lambda repressors, appear to have little effect on the site specificity of DNA
binding (127, 128). By contrast, substitutions at the ninth residue of the second HTH helix can lead to significantly altered preferences (127, 128). A curious feature of the observed specificity changes is that the new target sequence is quite different from the original one (123, 124, 126). It is puzzling how a single residue could determine so extensive a specificity, since it clearly could not contact more than two or three bases.

COMMON FEATURES

Families of HTH Domains

The remarkable similarity of the HTH units in lambda Cro and CAP, the first structures determined, showed that this motif could be found embedded in otherwise rather different domains (1). At least six distinct types of HTH-containing domains have been seen so far (Figure 2).

1. The HTH in the repressors of phages 434 (13) and lambda (43) and the Cro of 434 (12) is part of a 4-helix bundle. Helices 1, 2, and 3 form a roughly triangular unit, and helix 4 closes off its hydrophobic interior. The helix 5 of lambda is not homologous to the helix 5 of the 434 protein. In both cases, the fifth helix serves as a dimerization element rather than as a structural element of the HTH-containing domain. By inference from sequence alignment and from the helix-swap experiment (30), P22 repressor is also a member of this group.

2. The HTH in lambda Cro is part of a three-helix bundle, with a small β-sheet to close off the hydrophobic core (41). The first helix coincides roughly with the C-terminal part of helix 1 in lambda repressor and the 434 proteins. The sheet is formed from an N-terminal segment and from the part of the polypeptide chain that immediately follows helix 3.

3. The DNA-binding domain of CAP is also a three-helix bundle closed off by a β-sheet, but the arrangement of the elements is different (42). The first helix (helix D) is roughly coincident with helix 4 in lambda repressor and the 434 proteins, and the sheet contains a segment lying between helices D and E as well as a C-terminal hairpin.

4. The lac-repressor “headpiece” is a simple three-helix bundle (15). The third helix succeeds the HTH motif in the sequence, and its position is intermediate between those of helices 1 and 4 in lambda repressor and the 434 proteins.

5. The Antp homeodomain is also a three-helix bundle, but the additional helix precedes the HTH in the sequence (17). It lies essentially parallel to the first of the HTH helices.

6. Trp repressor has a different design, because of the interdigititation of helices from two subunits (44). The C-terminal part of its long helix C
coincides rather well with the first helix in the Antp homeodomain, if the HTH units are superimposed (17).

These comparisons suggest that the various HTH domains can be classified by the way in which one or two more helices pack against the HTH unit and by the way in which short additional elements close off the hydrophobic core. The similar placement of additional helices is probably determined as much by the limited number of favorable helix packings as it is by the requirements for DNA binding. It is nonetheless of interest that in several of the structures, the positively charged, amino-terminal ends of at least three α-helices are in position to contact a DNA phosphate (14, 72). These hydrogen bonds are expected to be particularly strong, as a result of oriented peptide dipoles in the α-helices (129).

**Characteristics of Protein/DNA Contacts**

1. Sugar-phosphate backbone. Particularly noteworthy are hydrogen bonds to nonesterified phosphate oxygens, especially from peptide -NH groups, neutral -NH₂ groups of glutamine and asparagine side chains, and -OH groups of serine and threonine. These interactions occur in the context of tight van der Waals complementarity that anchors the proteins very precisely. Hydrogen bonds from positively charged side chains to DNA phosphates appear with only modest frequency. Coulombic interactions from less strongly anchored arginines and lysines do appear to be important, however, since in each structure a number of such residues lie near the DNA backbone.

2. Base pairs. The variety of protein/base-pair contacts is more evident than their regularity. Various side chains (Gln, Asn, Ser, Arg, Lys) donate and accept hydrogen bonds. Some interact in ways previously suggested (26), such as Gln 284_34 and the adenine in operator base pair 1. Others form unanticipated, coordinated hydrogen-bonding patterns: for example, the interaction of Lys 4 and Asn 55 of lambda repressor with the guanine in base pair 6 (72). Nonpolar interactions appear to be equally significant. Extensive van der Waals complementarity between the major groove and the inserted protein elements is a striking characteristic of the phage repressor complexes. A corollary is that many of the contacted base pairs interact with more than one amino-acid side chain, and many of these side chains interact with more than one base pair. If the tightly bound water molecules are included, the same conclusion is also true of trp repressor.

**Anchoring the HTH to DNA**

The several structures examined thus far clearly rule out a universal mode for docking the HTH motif against the major groove of DNA. Not only does the precise “angle of attack” vary from case to case but the major groove itself has
a variable geometry. For example, although 434 repressor and Cro attach similarly to DNA backbone, they create somewhat different groove structures when they bind (9–11, 14). Despite these variations, common features of the 434-, lambda-, and trp-repressor complexes suggest important regularities in the way the HTH element is anchored to DNA (76).

In all three repressors, and in 434 Cro, the peptide -NH of the first HTH residue donates a hydrogen bond to a DNA phosphate on the 5' operator strand, and the -NH₂ group of the corresponding glutamine side chain donates a hydrogen bond to a second, adjacent phosphate (Figures 4, 5). Moreover, in lambda and 434 repressors, an asparagine at the C-terminus of the HTH also donates a hydrogen bond to this second phosphate. In trp repressor, the homologous asparagine bonds instead to the next phosphate. Thus, similar interactions anchor the HTH to one side of the major groove, denoted "segment 1" in the descriptions of individual structures above (76). On the other side of the major groove ("segment 2"), lambda repressor, 434 Cro, and trp repressor all donate hydrogen bonds from the peptide -NH of the last residue in the HTH turn to the nearest DNA phosphate (76). Note that this phosphate shifts away from the turn in the complex of 434 with ORI repressor and eliminates the direct hydrogen bond seen in the other structures. In 434 Cro and trp repressor, another hydrogen bond to this phosphate comes from the third residue in the second HTH helix. Additional phosphates on the 3' strand (segment 2) are contacted mainly by residues outside the HTH, but in quite different ways in the various complexes.

The relative positions, with respect to each other, of the key phosphates in segments 1 and 2 reflect the orientation of the HTH on DNA. In the complexes formed by the 434 proteins and by lambda repressor, the spacing of these phosphates is the same. That is, five base pairs intervene between the phosphate accepting a hydrogen bond from the N-terminus of the first HTH helix and the phosphate accepting a hydrogen bond from the last peptide in the turn. The HTH units therefore lie at similar angles to the major groove. In the trp-repressor complex, there are six intervening base pairs, and its first HTH helix must lie at a different angle to the direction of the major groove than the first helix of the other repressors. The position of the second helix relative to the first is also somewhat different in trp repressor. The combined differences result in different phosphate contacts from homologous asparagines at position 9 of the second helix.

What about other HTH proteins? The structures for lambda Cro/DNA and lac-repressor “headpiece”/DNA complexes are not yet precise enough for definitive descriptions of backbone interactions. The present data on the lac complex are, however, fully consistent with similar anchoring contacts (99, 102). That is, phosphate 2 is near the N-terminus of the first HTH helix, and phosphate 6' (five base pairs away) is near the end of the turn. If present, the
suggested interactions would position the lac-repressor HTH on DNA very much like the phage repressors. The corresponding operator alignment is shown in Figure 5. The published model for a CAP/DNA complex (113) also suggests homologous contacts, since the N-terminus of helix E is near a phosphate eight base pairs from the dyad on the 5' strand and since the turn and Thr 182 are near a phosphate three base pairs from the dyad on the 3' strand. When the HTH of CAP is superimposed on the HTH of 434 repressor (Figure 2a,e), the N-terminus of helix D (CAP) nearly coincides with the N-terminus of helix 4 (434). The N-terminus of the helix 4 contacts a phosphate in the 434 repressor/operator complex. Thus, similar phosphate interactions could be made by a structural element that precedes the HTH in the sequence of one protein (CAP) and follows it in another (434 repressor).

The longer first HTH helix of the Antp homeodomain (16, 17) probably prevents it from forming a set of anchoring interactions like those made by the phage repressors. The apparent indifference of binding specificity in other homeodomains to the initial residues of the second HTH helix also implies a different orientation.

RECOGNITION

From the structures described we can draw a few summary conclusions.

1. Several mechanisms contribute to recognition, to different extents in different cases. (a) Direct contacts, polar and nonpolar, between amino-acid side chains and the edges of base pairs in the major groove are principal sources of specificity in lambda and 434 repressor. There are multiple ways to recognize a given kind of base pair. For example, AT is specified through bidentate glutamine-adenine hydrogen bonds in one case, a van der Waals pocket for a thymine methyl group in another case. This observed multiplicity immediately rules out the simplest sort of code. Changes in directly contacted base pairs generally decrease affinity by at least one or two orders of magnitude. (b) In 434 repressor, the free energetics of DNA conformation make an additional contribution to specificity. Whereas lambda repressor appears to contact central base pairs directly through its N-terminal arm, 434 repressor relies on a differential response to distortions in operator conformation. Individual base-pair changes in the center of the operator have a smaller effect on affinity than changes in contacted base pairs. (c) Recognition by trp repressor appears in part to involve water-mediated hydrogen bonding to operator base pairs. Tightly bound solvent molecules seen in other structures line crevices at the DNA/protein interface (11, 14), but they do not seem to have a direct role in determining specificity.

2. The registration of the HTH motif on DNA, established by partially conserved contacts to sugar-phosphate backbone (76), can vary. Comparisons
are only straightforward when two proteins are anchored with a common local frame of reference relating the HTH to the major groove. As described in the preceding section, the 434 proteins, lambda repressor (and Cro?), and perhaps lac repressor do bind in similar ways. As a result, a correlation exists between particular positions in the HTH and base pairs that can be contacted by residues at those positions. Such correlations suggest that for proteins binding with a common register, certain recurring, local "recognition patterns" might be found. One such pattern is the adenine-glutamine-glutamine-phosphate network, seen in operator complexes of the 434 proteins and of lambda repressor (Figure 4d). This network specifies an adenine. Too few structures are known for any other such patterns to have been observed. The 434 proteins specify a C-A step by a set of contacts between the side chain of Gln 29 and the complementary G-T bases. We note that a similar pattern would be possible in a lac repressor/operator complex, if the register suggested by the present NMR studies were confirmed, because glutamine and C-A bases would be present at homologous positions.

3. The second HTH helix, which lies in the major groove, is obviously important for recognition. Not all the critical interactions come from its amino-acid residues, however. Lambda repressor (72) specifies operator base pair 6 by coordinated interactions from a side chain in the loop following the second HTH helix and one in the N-terminal arm; it specifies base pair 8 by another residue in the N-terminal arm. Trp repressor (40) positions one of the tightly bound water molecules at its DNA interface by a residue in the first HTH helix. The 434 proteins come closest to restricting direct base-pair contacts to the second HTH helix. But the relative free-energy cost of imposing a proper conformation on a 434 operator appears to be important in two respects—for the response of each protein to the sequence at the operator center (52) and for the differential response of the two proteins to a nonconsensus GC in $O_3$ (10, 14). The DNA conformation is established by a matrix of interactions involving residues distributed throughout the domain as well as by the overall geometry of the 434 and Cro dimers. Thus, the unit that "recognizes" DNA is really an entire binding domain, appropriately dimerized, and nearly all the protein/DNA contacts in a complex with a correct target site contribute to specificity.

Acknowledgments

We thank R. Kaptein and K. Wüthrich for sending unpublished coordinates and manuscripts prior to publication; B. W. Matthews and G. B. Koudelka for communicating unpublished results; D. Rodgers, R. Marmorstein, T. Ellenberger, D. C. Wiley for comments on the manuscript; C. Pabo for discussions; M. Ptashne for a continuing collaboration; R. Helmiss for illustrations; and K. Lee for assistance with manuscript preparation.
Literature Cited