Volume 10 Number 16 1982 Nucleic Acids Research

Two helix DNA binding motif of CAP found in *lac* repressor and *gal* repressor

I.T.Weber, D.B.McKay and T.A.Steitz

Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511, USA

Received 31 March 1982; Revised and Accepted 5 July 1982

ABSTRACT

Comparison of both the DNA and protein sequences of catabolite gene activator protein (CAP) with the sequences of lac and gal repressors shows significant homologies between a sequence that forms a two α-helix motif in CAP and sequences near the amino terminus of both repressors. This two-helix motif is thought to be involved in specific DNA sequence recognition by CAP. The region in lac repressor to which CAP is homologous contains many i_d mutations that are defective in DNA binding. Less significant sequence homologies between CAP and phage repressors and activators are also shown. The amino acid residues that are critical to the formation of the two-helix motif are conserved, while those residues expected to interact with DNA are variable. These observations suggest that the lac and gal repressors also have a two α-helix structural motif which is involved in DNA binding and that this two helix motif may be generally found in many bacterial and phage repressors. We conclude that one major mechanism by which proteins can recognize specific base sequences in double stranded DNA is via the amino acid side chains of α-helices fitting into the major groove of B-DNA.

INTRODUCTION

The three dimensional structures of two regulatory proteins that recognize specific nucleotide sequences in double stranded DNA have been solved: the E.coli catabolite gene activator (CAP) and the Cro repressor protein (Cro) of lambda phage[1,2,3]. There are several structural features of these two proteins that are strikingly similar to each other and thus are likely to be important to the function of specific sequence recognition. Both proteins have two subunits each of which contains an alpha helix that protrudes from the surface of the protein. These dimer related alpha helices are separated by 34 Å across a molecular two-fold axis. In the case of Cro, these two protruding alpha helices just fit into successive major grooves of right-handed B-DNA, while in the case of CAP these two alpha helices fit into successive major grooves of left-handed B-DNA. Furthermore, a detailed comparison of the structures of Cro and CAP shows that the DNA binding domain of both proteins contains an identical two-helix motif.[4] The 22 α-carbon atoms of the E and F alpha helices of CAP can be exactly superimposed on the

0305-1048/82/1016-5085$2.00/0

corresponding 22 α-carbon atoms of the α_2 and α_3 helices of Cro with an rms difference of 1.1 Å. Moreover, this two-helix motif does not occur in any other protein structure available from the Brookhaven Protein Data Bank file. Very similar structural features appear to exist in the recently determined structure of the DNA binding domain of lambda repressor (Pabo and Lewis, private communication).

This strong structural homology that occurs between CAP and Cro suggests that the two-helix motif may be involved in the function of DNA recognition and may be a motif that occurs more generally in DNA binding proteins. Anderson et al.[5] have shown that the sequence of Cro repressor in this two helix region is homologous to sequences found in other phage repressors and activators. Also, the phage repressors have a weak, but significant homology to the E. coli lac repressor.[34] We show here that both the DNA and protein sequences of CAP corresponding to this two-helix motif are homologous to sequences in the amino terminal regions of E. coli lac and gal repressors and to a lesser extent to regions in the phage repressor and activator proteins.

METHODS

The Comparison of Sequences

The amino acid sequences and also the DNA coding regions were compared for the following DNA-binding proteins: CAP[6,7], lac[8] and gal[9] repressors from E. coli, Cro[10], cI[11] and cII[10] from λ phage and C2 from phage P22[12]. Only the sequence corresponding to the amino terminal headpiece domain was included for lac and gal repressors since this region is expected to bind to DNA[13,14,20]. Likewise the sequence of the carboxy terminal domain of CAP was used since this is implicated in DNA binding[23,24].

The DNA sequences encoding these proteins were compared in pairs and codon by codon to ensure the correct alignment with the amino acid sequence. The number of identical bases was counted within a window of 66 bases (corresponding to 22 amino acid residues) for every possible alignment of the two sequences. In this way, for example, all possible 66 base (22 amino acid) comparisons between the C-terminus of CAP and the N-terminus of lac repressor were evaluated. For each pair of sequences compared, the number of 66 base comparisons found at each percentage agreement was plotted. Figure 1a shows the distribution of agreements for all possible comparisons between CAP and lac repressor. The mean agreement as well as the standard deviation were calculated and the distribution of comparisons is Gaussian as expected. In this way the statistical significance of any comparison can be derived.

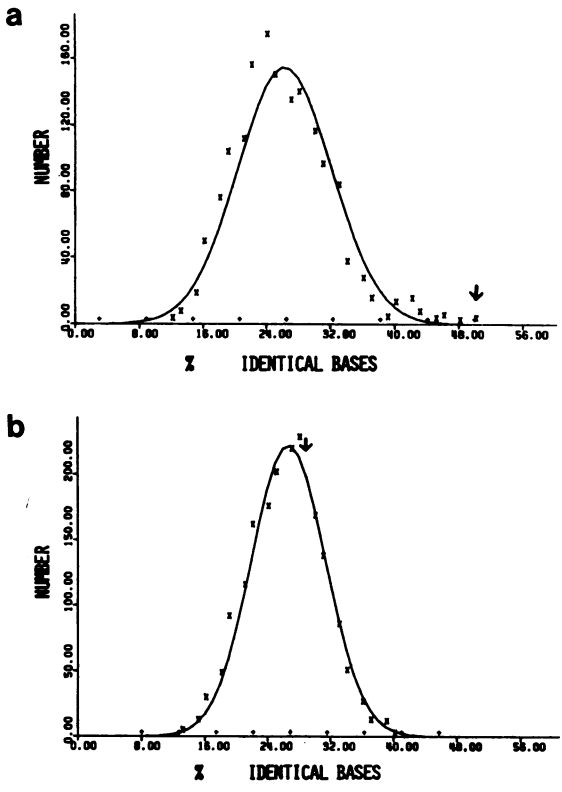

Figure 1a. The distribution of agreements found for all the possible 66 base comparisons between the DNA sequences coding for the N-terminal domain of lac repressor and the C-terminal domain of CAP. The smooth curve is the Gaussian fit to the numbers of comparisons found for each percentage agreement. The crosses at the bottom of the figure indicate the mean value and 1, 2, 3 and 4 standard deviations from the mean. The arrow points to the comparison shown in Table 1. b) The distribution of agreements found for all the possible comparisons of 66 bases between the DNA sequences coding for the C-terminal domain of CAP and Cro repressor.

The mean comparison of these two sequences shows about 26% agreement and one standard deviation is about 5.9%. The best sequence homology between the DNA binding domains of CAP is between bases coding for the two helix motif in CAP and bases coding for residues 4-26 in lac repressor. This comparison is 4.1 standard deviations from the mean. Figure 1b shows the distribution of agreements for a similar comparison between CAP and Cro. Statistics were computed for each pairwise comparison: the mean value for the number of identical bases and the standard deviation for all possible alignments are included in Table 1. No insertions or deletions of bases were considered

Table 1
DNA Sequence Comparisons for the Alignment of Proteins shown in Figure 2

	CAP	lac	gal	Cro	CIλ	CIIλ
lac	50.0*					
	(4.1)					
gal	48.5*	57.6*				
	(4.1)	(4.7)				
Cro	28.8	34.9	36.4			
	(0.5)	(1.5)	(1.8)			
CI	33.3	36.4	34.9	28.8		
	(1.3)	(2.1)	(1.7)	(0.4)		
CII	31.8	33.3	33.3	50.0	33.3	
	(1.0)	(1.1)	(1.3)	(4.2)	(1.3)	
P22 C2	39.4	43.9	31.8	37.9	30.3	34.9
	(2.4)	(3.4)	(1.3)	(2.1)	(0.7)	(1.6)

The percentage of identical bases within a 22 amino acid region is given for each pair of proteins.
The value in brackets is the number of standard deviations from the mean of the distribution of possible alignments.

*The best fit for all possible alignments of the two protein sequences is at this position.

within the 66 base region.

The best alignment of lac and gal repressors with CAP was over the 22 amino acids shown in Figure 2. The amino acid sequences for the best overall alignment of CAP, lac and gal with the phage protein sequences are also shown. Table 1 lists the percentage of identical bases in this 22 amino acid region and the number of standard deviations from the mean for the best overall alignment of Figure 2.

The 66 base homologous region defines probe sequences that were employed to search other sequences of DNA-binding proteins for homology. The new sequence was compared with each probe sequence in turn. For each 66 base region on the new sequence, the number of bases identical to each probe sequence was summed to give the best overall fit. The amino acid sequence of the protein was also searched for a consensus sequence. Thus, each DNA bind-

ing protein was searched for homology with the two-helix region on the DNA and the amino acid level.

Finally the amino acid side chains of the homologous sequences in lac and gal repressors were built into the two helix motif of the CAP structure as shown in Figures 3 and 4. The two helices were displayed on an interactive computer graphics system. The backbone was kept constant and side chains different from those of CAP were moved only if necessary to avoid contact with other atoms.

RESULTS

Figure 2 shows that a statistically significant homology exists between the amino acid sequence of polypeptide that forms the E and F helix of CAP and sequences in the DNA binding domains of lac and gal repressors. Six of the twenty two amino acids are identical between CAP and lac repressor and six are closely similar, such as Ile and Val. The homologies can also be clearly identified in the DNA encoding the proteins. The best of all the possible alignments between the DNA sequences encoding the DNA binding domains of CAP and lac repressor involves the 66 bases that encode the residues forming the E and F helices of CAP. In this alignment, half the bases are identical. No other part of either the lac or gal DNA or protein sequence is as closely homologous. Further, there is no homology (17% on the DNA level) between residues 188 and 200 of CAP and 25 to 37 of lac repressor.

SEQUENCES HOMOLOGOUS TO CAP

	α HELIX E		α HELIX F
	168		
CAP	THR-ARG-GLN-GLU-ILE-GLY-GLN-ILE-VAL-GLY-CYS-SER-ARG-GLU-THR-VAL-GLY-ARG-ILE-LEU-LYS-MET		
	5		
LAC R	THR-LEU-TYR-ASP-VAL-ALA-GLU-TYR-ALA-GLY-VAL-SER-TYR-GLN-THR-VAL-SER-ARG-VAL-VAL-ASN-GLN		
	3		
GAL R	THR-ILE-LYS-ASP-VAL-ALA-ARG-LEU-ALA-GLY-VAL-SER-VAL-ALA-THR-VAL-SER-ARG-VAL-ILE-ASN-ASN		
	15		
CRO	GLY-GLN-THR-LYS-THR-ALA-LYS-ASP-LEU-GLY-VAL-TYR-GLN-SER-ALA-ILE-ASN-LYS-ALA-ILE-HIS-ALA		
	32		
CIλ	SER-GLN-GLU-SER-VAL-ALA-ASP-LYS-MET-GLY-MET-GLY-GLN-SER-GLY-VAL-GLY-ALA-LEU-PHE-ASN-GLY		
	25		
CIIλ	GLY-THR-GLU-LYS-THR-ALA-GLU-ALA-VAL-GLY-VAL-ASP-LYS-SER-GLN-ILE-SER-ARG-TRP-LYS-ARG-ASP		
	20		
P22 C2	ARG-GLN-ALA-ALA-LEU-GLY-LYS-MET-VAL-GLY-VAL-SER-ASN-VAL-ALA-ILE-SER-GLN-TRP-GLU-ARG-SER		

Figure 2. Amino acid sequences of bacterial and viral repressors and activators found to be homologous to CAP. The invariant Glycine residue is triply underlined, the partially invariant residues are either doubly or singly underlined. These residues are important for the two helix structure. Additional homologies are evident.

Thus, the homology exists only with this alignment and over this short stretch of 22 amino acid residues.

In addition to the clear sequence homology that exists between the residues forming the E and F helix region of CAP and the lac and gal repressors, there is a weaker homology between this same region of CAP and a region in the phage repressors and activator. Weaker, but internally consistent sequence homologies have been shown to exist between residues that form the α_2 and α_3 helices of Cro repressor and sequences in other phage re-repressors and activators[5]. An homologous alignment between CAP and the phage repressor sequences can be made either by 1) using the structural superposition of the CAP and Cro two helix motifs or 2) searching for sequence homologies between CAP and the phage repressors. Both the alignment of these sequences on the basis of sequence homologies and on the basis of three dimensional structure gives rise to the same alignment, which is shown in Figure 2. Within the set of compared phage repressor sequences, some pairs show extremely weak homologies. For example, DNA sequences of CAP and Cro have only 29% identical bases in the region encoding the two helix motifs. Yet, comparison of the protein structures have shown the two helix motifs to be identical in these two proteins. The only sequence comparison between CAP and the phage repressors that appears to be significant is between CAP and the P22 repressor. Nevertheless, when taken together, the evidence is strong for homologies among these repressor and activator proteins in the region corresponding to the two helix structure in CAP and Cro.

Lac and Gal Repressor Can Form the Two Helix Motif

It is generally assumed and observed that amino acid sequence homology results in close structural homologies. Thus, we would expect that the sequence homology observed between CAP and the lac and gal repressors would allow us to directly build the structure of the lac and gal repressor sequences that are homologous to the known CAP structure. In Figure 3b and c we show that as expected one can indeed build these sequences into a two-helix motif identical to that in CAP. Inspection of the homologous sequences of the phage repressors and activator indicates that these sequences could also be built into a two-helix motif. We, therefore, conclude that all of the regulatory proteins listed in Table 1 contains a two-helix motif in the region that is homologous to CAP and Cro.

Conserved Residues are Essential for the Structure of the Two Helix Motif

Assuming that the sequences shown in Table 1 all form a two-helix

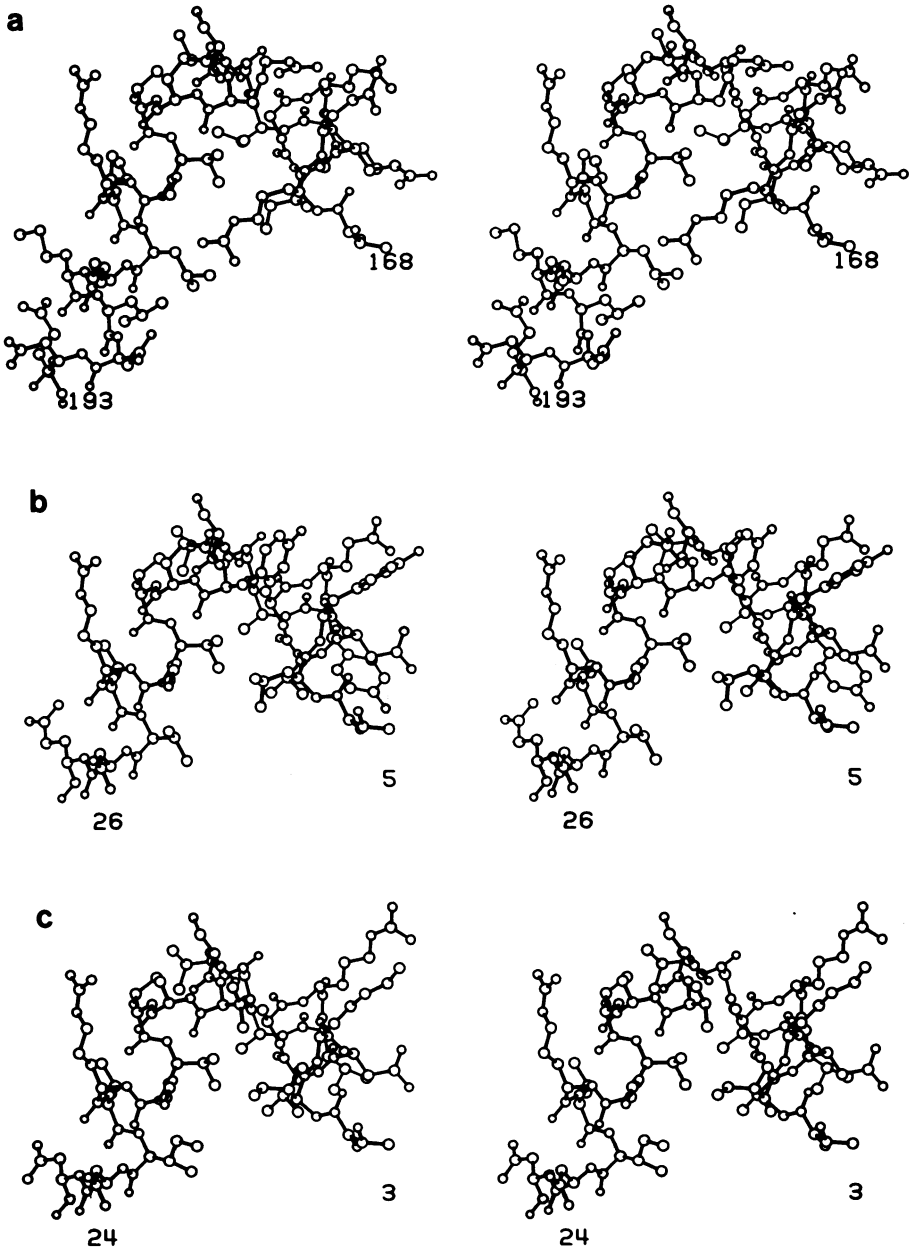

Figure 3. a) A stereo drawing of the E and F α-helices of CAP. b) A stereo drawing to residues 5 to 26 of lac repressor built into the two helix backbone of the CAP structure. c) A stereo drawing of residues 3 to 24 of gal repressor built into the two-helix motif.

motif, we have examined the pattern of conserved and semi-conserved residues to determine where the invariant and variable residues lie. In general the conserved and semi-conserved residues are essential to the formation of the two-helix motif. They are involved in the interaction between the two helices or in forming the turn between the helices. In contrast, the variable positions tend to be the polar residues on the surface interacting with solvent and/or presumably DNA. The situation is exactly analogous to hypervariable and constant regions of immunoglobulins. Figure 4 portrays the two-helix motif including the residues that are homologous between CAP and lac repressor and residues that are semi-invariant among all of the homologous sequences. The three most important invariant residues governing this two helix structure in CAP appear to be 1) a glycine at position 177, 2) a glycine or alanine at position 173 and 3) valine or isoleucine at 183. Gly

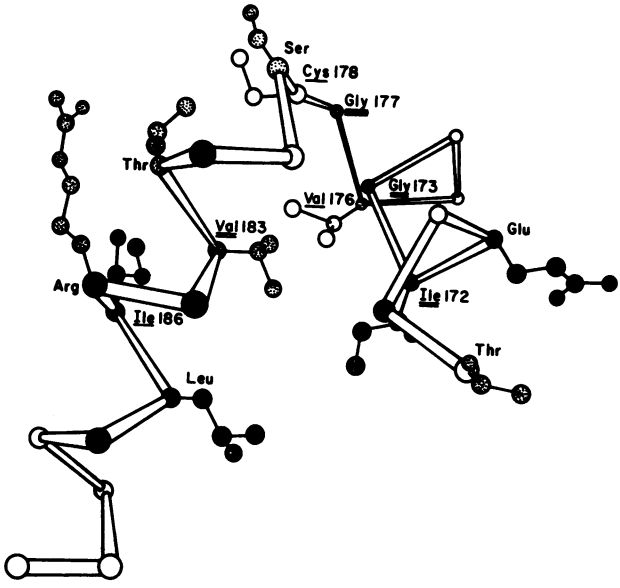

Figure 4. A drawing of the α-carbon backbone of the two helix motif and the homologous side chains. Stippled atoms are side chains that are identical in CAP, lac and gal. Striped atoms are closely similar in CAP and lac repressor and include side chains that seem to be important in interactions between the two helices. The triply underlined glycine is absolutely essential, the doubly underlined residues are either of two types and the singly underlined residues appear to be one of several hydrophobic amino acids. The completely filled α-carbon atoms indicate some side chains that may interact with DNA, derived from model building with CAP and DNA. Some additional interactions with DNA are also possible.

177 is essential for the turn between the two helices and no other amino acid can be put into this position without destroying the structure. That is, the backbone conformational angles, phi and psi, have values that lie in the excluded region of a Ramachandran plot. Only glycine can have these phi, psi values. Position 173 cannot accommodate a larger side chain than alanine since the beta carbon atom is pointed towards an interior pocket of limited volume. The α-carbon of Gly 173 is about 3 Å from two carbonyl oxygens on the F helix. The side chain valine or isoleucine at position 183 is likewise making a significant interaction between the two helices.

There are four other positions which contain hydrophobic residues in all of the structures and appear to be semi-invariant and important for maintenance of the two helix structure. One is at position 172 which can contain either a threonine, valine, isoleucine or leucine - amino acids whose side chains have very similar structures. At position 176, valine is the most commonly found side chain but other hydrophobic side chains such as leucine, alanine and methionine are found there. At position 178 either valine, threonine, isoleucine, methionine, or cysteine are found, while at position 186 isoleucine, valine, leucine, tryptophan or alanine are observed. As can be seen in Figure 4, all of these side chains come together in a hydrophobic pocket between the two alpha helices. Equally important for the proposal that these sequences form the two-helix motif is the observation that there are no residues involved in interaction between the two helices that are not at least partially conserved.

We have also compared the sequence of the two helix region of CAP on both the DNA and protein level with the sequences of other proteins that bind specifically to double stranded DNA. While it is frequently possible to find that as many as 40% of the bases are identical, these other comparisons are not as convincing on the level of protein structure. Most importantly, the protein sequence does not show the same invariant residues required for the maintenance of this two helix structure. The glycine at position 177 and the glycine or alanine at 173 are particularly important in this regard. For example, although there is a strong homology between the N-terminal region of the *E. coli trp* repressor[29] and many of the sequences in Figure 2, there is a Lys at the position corresponding to Gly 173. Other *E. coli* proteins such as *araC*[30] and *lexA*[31] have at least one alignment that is quite homologous to the first helix of the motif (CAP helix E), but is a poor match to the second helix (CAP F). These would not fit easily into the same structure. While it is likely that many more protein repressors and activators besides

the ones described here have this two-helix motif, it appears to be more difficult to find these regions in the sequences currently available for other proteins that recognize specific sequences in double stranded DNA.

DISCUSSION

The structural motif of two alpha helices that has been observed in the small domain of CAP and in Cro repressor[4] and which is presumed to be involved in specific DNA recognition appears to be more generally found in other bacterial and viral repressor and activator proteins. We have shown here that a significant sequence homology exists between that region of CAP sequence corresponding to the two-helix motif and amino terminal sequences of the lac and gal repressor proteins. Figure 3 illustrates the two helix structure in CAP and that predicted for lac and gal repressor proteins. Anderson et al.[5] have shown that rather weaker sequence homologies exist among various phage repressors and activators, again in the region corresponding to this two-helix motif. We have further shown here that some sequence homology exists between the sequence of the two-helix motif in CAP and sequences found in several phage repressors and activators.

These amino acid sequences homologous to the two-helix motif in both CAP and Cro suggest that 1) the two-helix structural motif exists in many phage and bacterial regulatory proteins, 2) this two-helix motif is involved in specific sequence recognition, and 3) the principles of specific interaction between the two-helix motif and DNA will apply to all of these proteins. Knowledge of the basis of sequence recognition in one case may be extrapolated to all of these homologous regulatory proteins.

Lac Repressor Recognition of Operator DNA

There is substantial information available concerning the overall architecture of the lac repressor. The lac repressor is a tetramer of identical 39,000 mol. wt. subunits[32]. Analysis of the products of limited proteolysis and studies of mutants that are defective in operator binding have shown that each subunit consists of at least two domains[14,18,37]. Proteolysis with trypsin yields a monomeric amino terminal 59 residues (called 'headpiece') that binds to operator DNA similarly to intact repressor[13] and a larger carboxy-terminal fragment that forms tetramer and binds inducer[14]. Further, virtually all mutants defective in operator binding while remaining tetrameric (i^{-d}) map in the first 60 to 70 residues[15,20]. Mutants affecting inducer binding map in the latter half of the molecule.

Small angle solution x-ray scattering experiments on intact lac repres-

sor and the 'core' repressor show that the repressor is an elongated molecule with its amino terminal DNA binding domains occurring in pairs at either end of the elongated molecule[16,38]. Further, these data establish that the pairs of headpieces must be separated by a 100 to 120 Å. As the operator itself is only about 70 Å long, a pair of DNA binding 'headpiece' domains rather than all four subunits must be interacting with the operator DNA[16]. This is shown schematically by Figure 5. These data exclude a model for repressor-operator interaction which has the DNA interacting with all four subunits and lying parallel to the long axis of the protein[17]. Further data supporting a model in which only a pair of subunits forms one operator binding site have been provided by experiments with hybrid repressor-core tetramers[18] and with chimeric repressor-β-galactosidase molecules[19].

The mechanism by which the lac and gal repressors recognize their respective operator sequences may be very similar to the mechanism by which CAP and Cro recognize their DNA binding sites. As in the case of CAP, Cro and lambda repressor it is likely that when the DNA binding domains of two subunits of lac repressor interact with the operator, the protein two fold axis is aligned approximately with the DNA two-fold axis. Further, amino acid sequence homologies imply that the lac repressor DNA binding domain contains the two-helix motif and that the second of these two helices protrudes from the surface of the domain as is the case with CAP and Cro. We suggest that this helix is related by a two-fold axis to the corresponding helix in the other subunit and separated from it by 34 Å. We would therefore expect that a major feature of lac repressor recognition of operator DNA constitutes two alpha helices from two 'headpiece' domains fitting into successive major grooves of right-handed B-DNA. Figure 6 shows the two-helix motif in lac

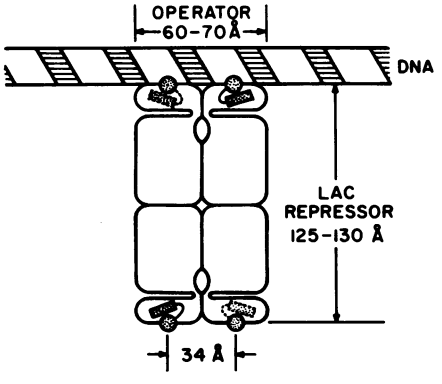

Figure 5. A schematic drawing of the overall structure of lac repressor and its interaction with operator as derived from small angle scattering[16]. The stippled areas indicate the anticipated positions of the two helix motif. The dimensions of the molecule are derived from the small angle scattering. The 34 Å separation of the α-helices is expected by analogy with CAP and Cro.

Figure 6. A drawing of the α-carbon backbone of the two helix motif interacting with the major groove of DNA. The dimer of lac repressor is illustrated. The stippled atoms are positions where i^{-d} mutants are found. The atoms with a vertical line mark amino acid side chains that are expected to interact with DNA.

repressor interacting with DNA.

Thus we would predict that part of the sequence recognition of the lac repressor arises from side chains emanating from an alpha helix fitting into a major groove of DNA. The specific side chains that would be interacting in the major groove include Tyr 17, Gln 18, Ser 21, Arg 22 and Asn 25. Proposals have been made for specific hydrogen bond pairing of Gln or Asn with Adenine and Arg with Guanine in the major groove of B-DNA[25,39]. We point out below that it is possible that pairs of these residues are interacting with each base pair. In addition to the side chains from the second alpha helix that would fit into the major groove it appears that some side chains from the first alpha helix also interact with the DNA. These might include Tyr 7, and Glu 11. A rather similar model involving an alpha helix fitting into the major groove was put forward by Muller-Hill and associates[20] on the basis of i^{-d} mutants and model building. They pointed out that there are mutants in the region corresponding to the alpha helix that fail to bind to DNA and built a specific model showing interactions between the side chains of this alpha helix and an operator sequence that they predicted (in-

correctly as it turns out). Also, Matthews et al.[34] have noted that lac repressor is homologous to the phage repressors and have similarly proposed that the two-helix motif is important in recognition of operator in the manner shown in Figure 6.

The model for lac repressor (Figure 3b) and its interaction with operator (Figures 5 and 6) presented here is consistent with the i^{-d} mutant data obtained in Mueller Hill's and Miller's laboratories[15,21,33]. The change of Thr 5 to Met creates a i^{-d} mutant. This Thr is one of the conserved residues between CAP, lac and gal repressors. The change of Tyr 7 to either Ser, Leu or Lys makes a mutant which may be explained by this Tyr being in a position to interact with DNA. There is evidence from NMR that Tyr 7 and Tyr 17 make a stacking interaction[26]. Such an interaction occurs when the lac repressor sequence is built into the CAP structure (Figure 3). Mutation of Val 9 to Ile creates a i^{-d} mutant presumably because this valine is one of the residues that is very critical for maintaining the structure of the two-helix motif as discussed above. A change of Tyr 12 to Ser, Gln or Leu has no effect on operator binding whereas its change to Lys creates a weak i^{-d} mutant. This Tyr points towards solution in our model and would not be interacting directly with the DNA. Presumably a lysine in this position is able to make some disruptive interaction with the sugar phosphate backbone. The change of Ser 16 to a proline creates a i^{-d} mutant. In our model this serine is at the amino end of the second alpha helix which might be distorted by the incorporation of a proline at this position. Any of a number of changes in Tyr 17 and Gln 18 create i^{-d} mutants presumably because both Tyr 17 and Gln 18 are in a position to interact directly with the bases in the major groove of B-DNA (Figure 6). A change of Thr 19 to an Ala creates a i^{-d} mutant. While it is not immediately obvious what functional role this Thr is playing, it is a conserved residue among CAP, lac and gal repressors.

We would expect that any changes in Ser 21, Arg 22, and Asn 25 residues would lead to an inactive repressor since these residues are in a position to be interacting directly with the DNA bases. Since the change of Gln 26 any one of a number of other residues has no effect,[33] that residue presumably cannot be interacting with operator.

We can predict which region of the operator DNA would be interacting with these alpha helices. By analogy with CAP and Cro the alpha helices are fitting into the two successive major grooves on the same side of the DNA. The interaction would start about three to four base pairs from the two-fold

axis in the operator sequence and would continue for about four to five base pairs on either side. That is, there is a span of approximately seven base pairs in the middle of the operator that would not be interacting with the alpha helices and a span of four or five base pairs on either side of this central seven base pairs that would be interacting with the alpha helices. This is largely consistent with the data on methylation protection[35] and experiments on binding of lac repressor to chemically synthesized 'mutant' operators[36].

There are four operator constitutive mutants in the putative interaction sites on both sides of the two fold axis in the operator. There are also four operator constitutive mutants in the central seven base pair region of the operator[28] which cannot be accounted for by interactions with the two helix motif as proposed here (Figure 5). There must be additional interactions between lac repressor and lac operator to account for these mutants. Since the minor groove of the operator must be facing the repressor in this region, one might suggest that antiparallel beta strands interacting as proposed by Church et al[22]. might account for the additional interaction between repressor and operator. From the mutant data one might expect that the region of lac repressor that would be making this additional interaction would include residues between 54 and 58 since other i^{-d} mutants are located in that region.

Model for Specific Sequence Recognition

The sequence homologies that several repressor and activator proteins show to the region of the two-helix motif suggests that this region is involved in specific sequence recognition. Thus, knowledge of how the two-helix motif of Cro or CAP is involved in specific sequence recognition will allow one to work out some aspect of how specific sequence recognition is achieved in the homologous proteins. Clearly, the only way that a completely reliable model for a CAP or Cro complex with a specific sequence of DNA can be constructed is from a crystal structure of such a complex. Initial attempts to co-crystallize CAP in the presence of a 16 base pair fragment of DNA have yielded small (100 to 150 μ) crystals (Goldman, Steitz, Ikuta and Itakura, unpublished observation, 1982). Until the structure of such a complex is available, however, it is valuable to examine what principles of protein-nucleic acid interaction might be derived on the basis of model building alone.

Model building of possible protein-DNA complexes using the structure of CAP and the known amino acid homologies shows first, that specific

recognition of DNA bases pairs can be achieved by 3 or 4 pairs of side chains spaced 3.4 Å apart and secondly, that the same side chain positions can be used for recognition of either left or and right-handed DNA. The important point is that the geometry of the side chains emanating from an α-helix and the geometry of the major groove of B-DNA are complementary.

Let us first consider an α-helix interacting in the major groove of right-handed B-DNA. The groove is tilted at about 32° to the planes of the base pairs and it is expected that the α-helices would also make an angle of 32° to the base planes. The two amino acid residues adjacent to a base pair have a 1, 4 relationship, so that pairs of amino acid residues (2 and 5, 6 and 9 etc.) could be interacting (Figure 7). Amino acid residues 2 and 5 lie at the same level along the DNA axis and an appropriate pair of side chains could interact with the hydrogen bond donors and acceptors of a single base pair of DNA. As a side chain of an amino acid which is 4 residues along the α-helix is displaced nearly 3.4 Å along the DNA axis, residue 6 is displaced about 3.4 Å along the DNA helix, relative to residue 2. Since the α-helix is nearly a four-fold helix (3.6 residues/turn) over a stretch of 3 or 4 turns, the patterns of residues that will be interacting with the DNA would be alternate pairs of residues.

The fact that pairs of residues can lie almost in the same plane and can interact with each other means that a multiple hydrogen bond donor and acceptor arrangement can be made. Each base pair provides three potential hydrogen bonding donor and acceptor sites exposed in the major groove. With

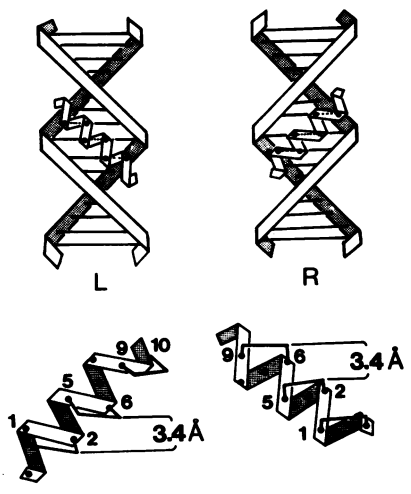

Figure 7. A schematic drawing showing the pairs of residues separated by about 3.4 Å interacting with the edges of the base pairs in either left or right-handed DNA. The same residues would interact with both left and right-handed DNA, but different pairings of protein side chains would occur.

two amino acid side chains it is possible to provide the H-bond donor-donor-acceptor arrangement required to recognize a G C base pair or the donor-acceptor-donor arrangement that is complementary to an A T base pair. For example, an Arg - Glu pair could interact specifically with G C whereas a Ser plus Gln can provide a donor-acceptor-donor arrangement required for interaction with A T. Although pairs of side chains can interact with one base pair, one side chain may also be sufficient in many cases.

Let us now consider what complementarity exists between the side chains of an α-helix and the bases in the major groove of left-handed B-DNA. It has been proposed (1) that CAP binds with α-helices fitting in successive major grooves of left-handed B-DNA. Although the observation of only small change in the linking number produced by binding CAP to closed circular DNA (40) does not support a model of CAP binding to left-handed DNA, data on the size of the DNA sequence that CAP must recognize and the general structural complementarity of CAP and left-handed DNA (1) are more consistent with CAP binding to left-handed DNA. Hence, an unresolved dilemma exists at the moment. The homologies shown here to exist between the sequence of the two helix motif of CAP and that of other regulatory proteins imply that this structural motif is directly involved in the specific DNA sequence recognition, and that it is not some other part of the CAP dimer that is binding to DNA. It is therefore of some interest to see whether this two helix motif could possibly interact specifically with left-handed DNA. That is, could model building eliminate the possibility of CAP binding to left-handed DNA? Unfortunately, model building thus far cannot resolve the dilemma.

It turns out to be as easy to build models of the two helix motif interacting with left-handed DNA as with right using the same amino acid residues to make the interactions (obviously, since it is the side of the α-helices exposed to solvent that can make interactions). However, in this case it is a different combination of pairs of amino acids that could bind to with a single base pair. Rather than side chains that are 4 residues apart along the α-helix, it is adjacent amino acids residues that lie in nearly the same plane as the DNA bases (Figure 7). In this case, the almost co-planar pairs of residues are: 1 and 2, 5 and 6, 9 and 10 and are separated by about 3.4 Å along the DNA axis. Thus, model building shows that the conserved DNA binding two helix motif can interact as well with left as with right-handed B-DNA and model building cannot eliminate the possibility of CAP binding to left-handed DNA.

In conclusion, it appears that it may be possible to build a model of any complex between a protein and the double stranded DNA to which it specifically binds if 1) the characteristic two helix region has been identified and 2) the DNA sequence to which it binds is known. The DNA sequence with which the α-helix will interact will be four base pairs that start either 3 1/2 or 4 base pairs from the two-fold axis in the sequence of the binding site. The protein side chains that are interacting with the base pairs will be at the same positions as in Cro and CAP. Additional interactions between protein and DNA will probably occur in each case and may be different. However, one important general principle of DNA-protein recognition appears to be that pairs of side chains emanating from an α-helix can specifically interact with the edges of base pairs in the major groove.

ACKNOWLEDGEMENTS

We thank Ray Salemme for useful discussions. This research was supported by National Science Foundation grant number PCM-81-10880 and by U.S. Public Health Service grant number GM-22778.

REFERENCES

1. McKay, D.B. and Steitz, T.A. (1981) Nature 290, 744-749.
2. Anderson, W.F., Ohlendorf, D.H., Takeda, Y. and Matthews, B.W. (1981) Nature 290, 754-758.
3. Steitz, T.A., McKay, D.B. and Weber, I.T. (1982) in Nucleic Acid Research: Future Development, Watanabe, I. Ed., Academic Press, Japan Inc., in press.
4. Steitz, T.A., Ohlendorf, D.H., McKay, D.B., Anderson, W.F. and Matthews, B.W. (1982) Proc. Natl. Acad. Sci. USA, 79, 3097-3100.
5. Anderson, W.F., Takeda, Y., Ohlendorf, D.H. and Matthews, B.W. (1982) J. Mol. Biol., in press.
6. Aiba, H., Fujimoto, S. and Ozakai, N. (1982) Nucl. Acid Research 10, 1345-1361.
7. Gicquel-Sanzey, B. and Cossart, P. (1982) Nucl. Acid Research 10, 1363-1378.
8. Farsbaugh, P.J. (1978) Nature 274, 765-769.
9. von Wilcken-Bergmann, B. and Muller-Hill, B. (1982) Proc. Natl. Acad. Sci. USA, 79, 2427-2431.
10. Schwarz, E., Scherer, G., Hobom, G. and Kossel, H. (1978) Nature 272, 410-413.
11. Sauer, R.T. (1978) Nature 276, 301-302.
12. Sauer, R.T., Pan, J., Hopper, P., Hehir, K., Brown, J. and Poteete, A.R. (1981) Biochemistry 20, 3591-3598.
13. Ogata, R.T. and Gilbert, W. (1978) Proc. Natl. Acad. Sci. USA, 75, 5851-5854.
14. Platt, T., Files, J.G. and Weber, K. (1973) J. Biol. Chem. 248, 110-121.
15. Miller, J.H. (1979) J. Mol. Biol. 131, 249-258.
16. McKay, D.B., Pickover, C.A. and Steitz, T.A. (1982) J. Mol. Biol., 156,

175-183.
17. Dunaway, M., Manly, S.P. and Matthews, K.S. (1980) Proc. Natl. Acad. Sci. USA, 77, 7181-7185.
18. Geisler, N. and Weber, K. (1976) Proc. Natl. Acad. Sci. USA, 73, 3103-3106.
19. Kanin, J. and Brown, D.T. (1976) Proc. Natl. Acad. Sci. USA, 73, 3529-3533.
20. Adler, K., Beyreuther, K., Fanning, E., Geisler, N. Gronenborn, B., Klemm, A., Muller-Hill, B., Pfahl, M. and Schmitz, A. (1972) Nature 237, 322-327.
21. Muller-Hill, B. (1975) Prog. Biophys. Molec. Biol. 30, 227-252.
22. Church, G.M., Sussmann, J.L. and Kim, S-H. (1977) Proc. Natl. Acad. Sci. USA, 74, 1458-1462.
23. Houmard, J. and Drapeau, G.R. (1972) Proc. Natl. Acad. Sci. USA, 69, 3506-3509.
24. Aiba, H. and Krakow, J.S. (1981) Biochemistry 20, 4774-4780.
25. Seeman, N.C., Rosenberg, J.M. and Rich, A. (1976) Proc. Natl. Acad. Sci. USA, 73, 804-808.
26. Arndt, K.T., Boschelli, F., Lu, P. and Miller, J.H. (1981) Biochemistry 20, 6109-6118.
27. Muller-Hill, B. (1975) Prog. Biophys. Molec. Biol. 30, 227-252.
28. Gilbert, W., Grall, J., Majors, J. and Maxam, A.M. (1975) in Protein-Ligand Interactions, Sund, H. and Blaur, B. Eds., 193-210, Walter de Gruyter, Bedin.
29. Gunsalus, R.P. and Yanofsky, C. (1980) Proc. Natl. Acad. Sci. USA, 77, 7117-7121.
30. Wallace, R.G., Lee, N. and Fowler, A.V. (1980) Gene. 12, 179-190.
31. Honi, T., Ogawa, T. and Ogawa, H. (1981) Cell 23, 689-697.
32. Gilbert, W. and Muller-Hill, B. (19709) in The Lactose Operon, Beckwith, J.R. and Zipser, D. Eds., Pp. 43-110, Cold Spring Harbor Laboratory, New York.
33. Miller, J.H. (1978) in The Operon, Miller, J.H. and Reznikoff, W.S. Eds., Pp. 31-88, Cold Spring Harbor Laboratory, New York.
34. Matthews, B.W., Ohlendorf, D.H., Anderson, W.F., and Takeda, Y. (1982) Proc. Natl. Acad. Sci. USA, 79, 1428-1432.
35. Ogata, T.R. and Gilbert, W. (1979) J. Mol. Biol. 132, 709-728.
36. Goeddel, D.V., Yansura, D.G. and Caruthers, M.H. (1978) Proc. Natl. Sci. USA 75, 3578-3852.
37. Files, J.G. and Weber, K. (1976) J. Biol. Chem. 251, 3386-3398.
38. Charlier, M., Maurizot, J.C. and Zaccai, G. (1980) Nature (London) 286, 423-425.
39. Helene, C. (1977) FEBS Lett. 74, 10-13.
40. Kolb, A. and Buc, H. (1982) Nucl. Acid Res. 10, 473-485.