Annual Reviews
www.annualreviews.org/aronline

Ann. Rev. Biochem. 1984. 53: 293–321
Copyright © 1984 by Annual Reviews Inc. All rights reserved

PROTEIN-DNA RECOGNITION

Carl O. Pabo

Department of Biophysics, The Johns Hopkins University School of Medicine, Baltimore, Maryland, 21205

Robert T. Sauer

Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

CONTENTS

INTRODUCTION

Sequence-specific DNA-binding proteins regulate gene expression and also serve structural and catalytic roles in other cellular processes. How do these proteins bind to double-helical DNA and how do they recognize a particular base sequence? Here we review recent crystallographic, biochemical, and genetic studies that address these questions. For the most part, we emphasize work published between 1980 and 1983, since the first

0066–4154/84/0701–0293$02.00

Annual Reviews
www.annualreviews.org/aronline

three-dimensional structures of site-specific DNA-binding proteins were reported during this period. Crystal structures are now available for the Cro and cI repressors of bacteriophage lambda and for the CAP protein of *Escherichia coli* (1–3). Each of these three proteins can turn off expression of specific genes by preventing initiation of transcription; CAP and lambda repressor can also enhance gene expression by stimulating transcription from certain promoters. Other recent reviews discuss these proteins and their physiological actions (4–10), the mechanism and control of prokaryotic transcription (11), and general aspects of protein-DNA interactions (12, 13).

Cro, lambda repressor, and CAP interact with DNA in a basically similar manner. Despite differences in size, domain organization and tertiary structure, each of these proteins binds to operator DNA as a dimer and uses α-helices to contact adjacent major grooves along one face of the double helix. Moreover, sequence homologies suggest that many other DNA-binding proteins use similar α-helical regions for DNA recognition. How does each of these proteins recognize its proper binding site? What forces are involved? Is there a "recognition code"? We consider these questions after discussing the structures of Cro, lambda repressor, and CAP and describing the models proposed for the respective protein-DNA complexes.

LAMBDA CRO

Lambda Cro binds to six operator sites in the double-stranded phage DNA (14, 15). These sites are clustered in two operator regions, and each region contains three 17-bp (base pair) sites. The DNA sequences of the six sites are similar but not identical, and Cro's affinity for the different sites varies over a tenfold range (15–17). The sequence of each operator site has approximate two-fold symmetry, and the consensus sequence, shown in Figure 1, is symmetric. The Cro monomer contains 66 amino acid residues (18, 19). Cro exists as a dimer in solution (20) and this is the form active in DNA binding.

A crystallographic study at 2.8 Å resolution by Anderson, Ohlendorf, Takeda, & Matthews (1) showed that the Cro monomer contains three

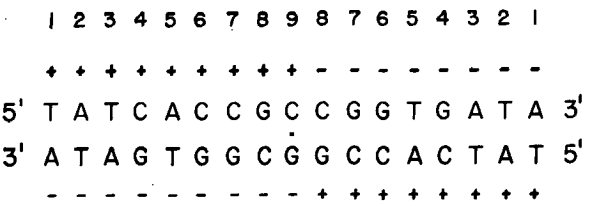

Figure 1 Consensus Operator Sequence for Lambda Cro and Lambda Repressor.

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

Annual Reviews
www.annualreviews.org/aronline

strands of antiparallel β-sheet (residues 2–6, 39–45, and 48–55) and three α-helices (residues 7–14, 15–23, and 27–36). The Cro dimer is stabilized by a region of antiparallel β-sheet that is formed by pairing Glu 54–Val 55–Lys 56 from each monomer. The four C-terminal residues of Cro (residues 63–66) are poorly represented in the electron density map and are probably somewhat disordered both in the crystal and in solution.

A Model for the Cro-Operator Complex

The structure of the Cro dimer immediately suggested its basic mechanism of DNA binding (1). In the dimer, the two copies of α-helix 3 form protruding ridges that are separated by the same center-to-center distance, 34 Å, that separates successive major grooves of B-DNA (Figure 2) (21, 22). The angle between the two Cro helices allows them to fit neatly into successive major grooves of the operator. This arrangement provides an excellent fit between the surface of the protein and the surface of the DNA, and accounts nicely for the observed DNA modification and protection data. This data, which is shown in schematic form in Figure 2, had suggested that Cro bound in a symmetric manner and that it contacted

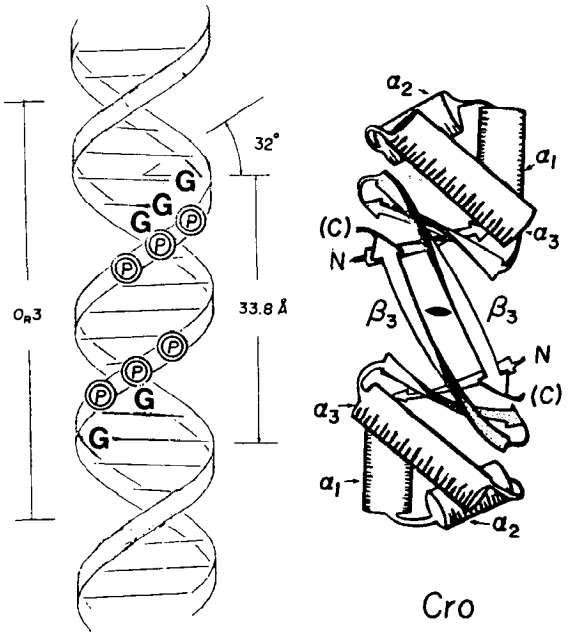

Figure 2 Sketch of the lambda operator site O_R3 and the Cro dimer. P's indicate phosphates that have been implicated as Cro contacts. G's indicate guanines implicated as contacts. Adapted with permission from (1).

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

Annual Reviews
www.annualreviews.org/aronline

adjacent major grooves along one face of the double helix (Figure 3) (15–17). The proposed complex also seems chemically reasonable, since a number of hydrophilic and charged residues can interact with exposed hydrogen-bonding groups in the major groove and with the negatively charged phosphates.

Refinement of the Cro structure at 2.2 Å and further model building by Ohlendorf et al (23) have allowed a detailed analysis of possible Cro-operator interactions. During this modeling of hydrogen-bonding interactions, energy minimization (24) was used to ensure that the stereochemistry of the proposed complex was reasonable. Model building started with a B-form operator (21, 22), since DNA in solution adopts this conformation (25). However, minor adjustments in the DNA structure were allowed, and in the best model for the complex the operator DNA was smoothly bent with a radius of curvature of 75 Å. Thus, the DNA is bent around Cro so that each end of the operator is 5 Å closer to the protein than it would be if the DNA were straight. This bending seems plausible since it should require only a few kcal of energy (26), but Ohlendorf et al (23) point out that a

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

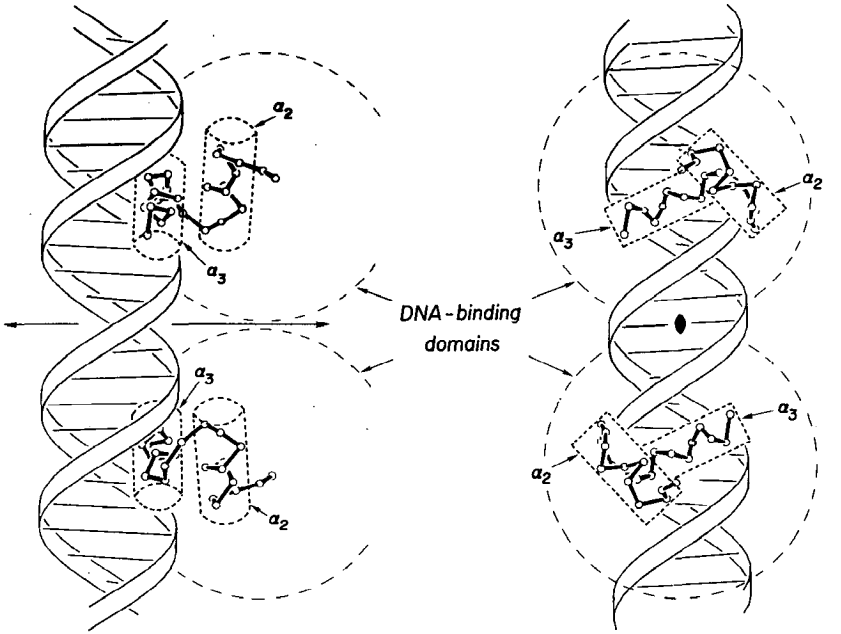

Figure 3 Alpha carbons from helices two and three of the proposed Cro-operator complex. Although the rest of the protein structures are quite different, the corresponding helical regions of repressor and CAP are quite similar and may contact the DNA in a similar manner. Used with permission from (8).

Annual Reviews
www.annualreviews.org/aronline

similar fit could also be obtained by a "hinge-bending" motion of the protein dimer that would allow it to contact a straight operator site.

The Cro-operator model predicts several specific contacts between each Cro monomer and the edges of base pairs in the major groove. These contacts, which all involve side chains in or near α-helix 3, are listed below. The base-numbering scheme is that for the consensus operator site shown in Figure 1:

1. The hydroxyl group of Tyr26 donates a hydrogen bond to 04 of the thymine at +1.
2. The side-chain amide of Gln27 donates a hydrogen bond to N7 of the adenine at +2, and accepts a hydrogen bond from N6 of the same adenine. This interaction is illustrated in Figure 4 (*top*).
3. The hydroxyl group of Ser28 forms two hydrogen bonds with N6 and N7 of adenine −3, as shown in Figure 4 (*center*).
4. The amino group of Lys32 donates one hydrogen bond to 04 of the thymine at −5, donates a second hydrogen bond to 06 of guanine −4, and may donate a third hydrogen bond to N7 of guanine −4.
5. The guanidinium group of Arg38 donates two hydrogen bonds to 06 and N7 of guanine −6, as shown in Figure 4 (*bottom*).

In addition, Gln27, Asn31, and Lys32 seem to make a few van der Waals contacts in the major groove.

The proposed Cro-operator complex also places a large number of residues near the sugar-phosphate DNA backbone. Those with hydrogen-bonding potential include Gln16, Thr17, and Lys21 in helix 2, which is just above the major groove. They also include Asn31, His35, and Lys39, which are in or beyond helix 3, and Glu54 and Lys56 in the C-terminal β-region (23). Several additional polar interactions might also be made by residues Asn61-Lys62-Lys63-Thr64-Thr65 in Cro's flexible C-terminal region.

Evidence Supporting the Cro-Operator Model

The overall fit of Cro against the operator site is supported by a calculation of the electrostatic potential around the Cro dimer (27, 28). There is a weak negative potential on the far side of the dimer, but the overall potential is dominated by a positive region that straddles the two-fold axis. This region of positive charge coincides remarkably well with the presumed DNA-binding site.

Protein modification studies provide general support for the model of the Cro-operator complex (Y. Takeda, J. Kim, C. Caday, D. Davis, E. Steer, D. Ohlendorf, B. Matthews, W. Anderson, manuscript in preparation). As the model predicts, Lys21, Lys32, Lys56, and Lys62/63 are protected from chemical modification when Cro is bound to DNA, but Tyr26 and Lys39,

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

Annual Reviews
www.annualreviews.org/aronline

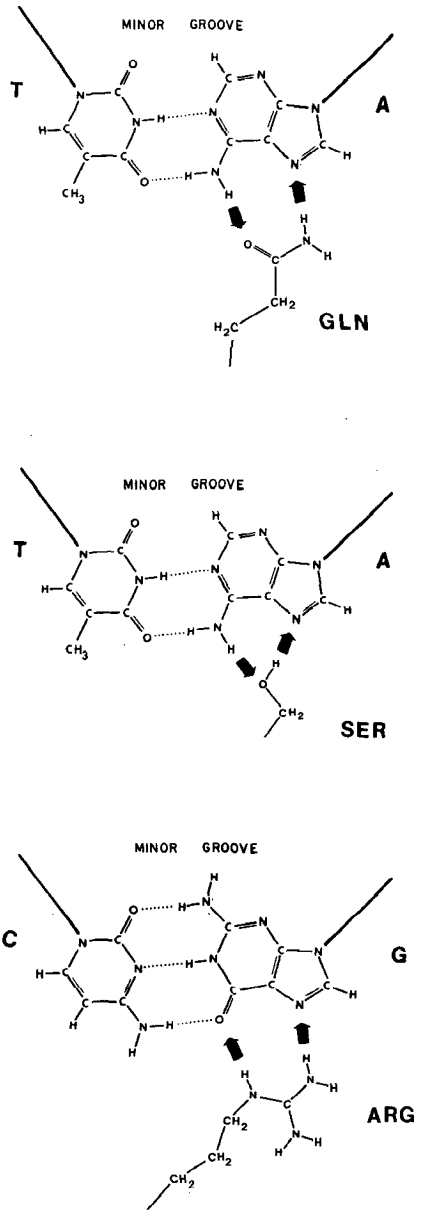

Figure 4 (*top*) Sketch indicating possible hydrogen bonds between glutamine and an A : T base pair (86). (*Center*) Sketch indicating how serine could form a pair of hydrogen bonds with an A : T base pair (23). (*bottom*) Sketch indicating the hydrogen bonds that arginine could form with a G : C base pair (86).

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

Annual Reviews
www.annualreviews.org/aronline

which are also predicted to be contact residues, are not protected. However, detailed interpretation of these results is complicated by the fact that the Cro is largely bound to nonoperator DNA in these studies, and it is likely that the structures of operator and nonoperator complexes may be somewhat different. Other experiments show that the affinity of Cro for operator DNA is reduced by carboxypeptidase digestion, and this result, together with the protection of Lys62/63, suggests that the flexible C-terminal region of Cro makes some DNA contacts.

Genetic studies also support the model proposed for the Cro-operator complex. Most of the proposed DNA contact residues are represented in a collection of mutations that are phenotypically defective in operator binding (A. Pakula, R. Sauer, manuscript in preparation). These include each of the five residues implicated in specific base contacts (Tyr26 → Asp; Gln27 → His; Ser28 → Arg/Asn; Lys32 → Thr/Gln; Arg38 → Gln) and many of the residues implicated in backbone contacts (Gln16 → His; Lys39 → Thr; Glu54 → Lys/Ala; Lys56 → Asn/Gln/Thr). In addition, three of the proposed specific contact residues have been altered by oligonucleotide replacement of the appropriate region of the *cro* gene (M. Nasoff, S. Noble, M. Caruthers, manuscript in preparation). The mutations introduced by this procedure (Tyr26 → Phe/Leu/Asp; Gln27 → Leu/Cys/Arg; Ser28 → Ala) all reduce the operator affinity of Cro. Although further analysis of all the *cro* mutations is needed to show that they do not disrupt Cro folding, it is likely that most of them owe their reduced operator binding to a defect in DNA binding. The correspondence between the positions of the mutations and the proposed DNA contact residues supports the model for the Cro-operator complex.

LAMBDA REPRESSOR

Lambda repressor recognizes the same six operator sites (14) that lambda Cro recognizes, and repressor also binds to each 17-bp operator site as a dimer (29, 30). The affinity of repressor for the six sites varies over a 50-fold range, but the sites for which repressor has highest affinity are not the sites for which Cro has highest affinity (15–17, 31). For example, the site called O_R3 is one of the weakest binding sites for repressor but is the strongest binding site for Cro. The different affinities of repressor and Cro for the six operator sites help explain the contrasting physiological roles of these two proteins (4–6). Chemical protection and modification studies show that Cro and repressor contact many of the same functional groups in the operators, but the Cro contacts seem to be a subset of the repressor contacts (compare Figures 2, 5, and 6) (15–17, 32). The repressor monomer contains 236 amino acids (33) and is thus considerably larger than Cro.

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

Annual Reviews
www.annualreviews.org/aronline

The repressor monomer folds into two domains of similar size, which can be separated by cleavage with papain or other proteases (34). The N-terminal domain, as an isolated proteolytic fragment of residues 1–92, binds specifically to the lambda operator sites and mediates both positive and negative control of transcription (35). Thus the regulatory activities of the intact protein are retained by the N-terminal fragment. However, the N-terminal fragment binds to the operator less tightly than intact repressor because the fragment dimer readily dissociates in solution (34). Intact repressor binds to the operator more tightly because the C-terminal domain stabilizes the dimer and thereby stabilizes the protein-DNA complex.

Intact lambda repressor has not been crystallized, but the structure of the N-terminal operator-binding domain has been solved at 3.2 Å resolution by Pabo & Lewis (3). The N-terminal domain consists of an N-terminal arm and five α-helices, and is a dimer in the crystal. The first eight residues of the domain form an arm that extends away from the globular region. Most of this arm packs against another molecule in the protein crystal, but residues 1–3 are disordered and not visible in the electron density map. The α-helices of the domain include residues 9–23 (helix 1), 33–39 (helix 2), 44–52 (helix 3), 61–69 (helix 4), and 79–92 (helix 5). The first four helices, along with the irregular regions of chain that connect them, form a compact, globular

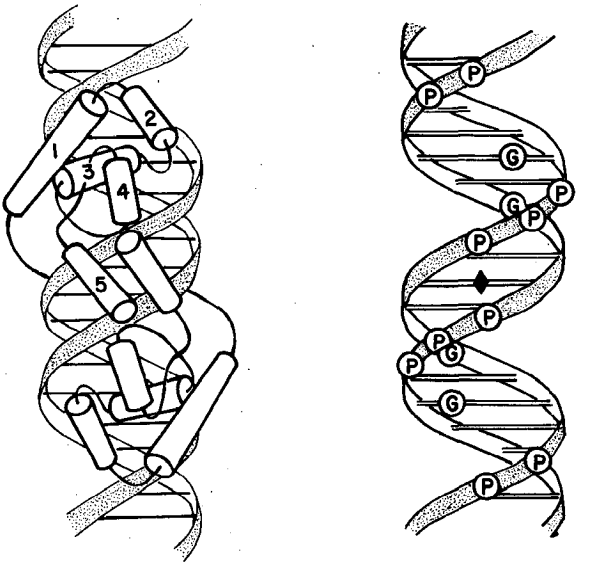

Figure 5 The proposed lambda repressor-operator complex. Panel on right summarizes the results of chemical protection experiments at the site of O_R1.

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

Annual Reviews
www.annualreviews.org/aronline

domain. The fifth helix extends off to one side of the molecule and folds against helix 5 of a neighboring monomer. This helix-helix contact seems to be the major interaction stabilizing the N-terminal dimer.

A Model for the Repressor-Operator Complex

Possible structures for the complex of the N-terminal dimer with operator DNA were evaluated by model building (3). B-form operator DNA was used for these studies because DNA in solution is B-form (25), and repressor does not significantly wind or unwind the operator DNA (36). NMR studies have subsequently shown that the operator conformation in the repressor-operator complex is similar to the expected B-conformation (M. Weiss, D. Patel, R. Sauer, M. Karplus, manuscript in preparation). The model building was also guided by chemical modification experiments that suggested that repressor contacted the major groove and that most contacts were on one side of the double-helical site (16, 17, 31). With these constraints, model building yielded only one arrangement that gave a good fit between the surface of the protein and the surface of the DNA. This complex, which is shown schematically in Figure 5, allows each subunit of the dimer to contact one half of the operator site. In each half site, the N-terminal portion of helix 3 fits directly into the major groove. Helix 2 is just above the major groove, and its N-terminal region is next to the sugar phosphate backbone of the DNA. Figure 5 also summarizes the chemical protection data, showing the guanine N7 and phosphate groups implicated as repressor contacts on the front side of the DNA helix (16, 17, 32).

In the proposed complex, the two N-terminal arms of repressor are set slightly to the sides of the operator helix and extend towards the "back" of the DNA near the center of the 17-bp site. Biochemical studies show that these arms actually make major groove contacts on the back of the operator site (37). The 92-residue N-terminal domain, like intact repressor, protects several operator guanines from chemical methylation; four of the protected sites are visible in the major groove on the "front" of the operator site, and two are visible on the "back" (Figures 5 and 6). However, a shorter N-terminal fragment, containing residues 4–92 and thus missing the first three residues of the arm, protects only the guanines on the front of the operator site. Since NMR studies show that the 1–92 and 4–92 fragments have the same conformation (M. Weiss et al, manuscript in preparation), the different protection patterns imply that the first three residues of the arm must contact the major groove on the back of the operator site. Model building indicates that repressor's arms are long enough to encircle the double helix, possibly by wrapping around the DNA in the major groove, as shown in Figure 6. The first five residues of the arm, Ser1-Thr2-Lys3-Lys4-Lys5, are polar and could readily make hydrogen bonds to bases in the major groove or interact with the sugar-phosphate backbone.

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

Annual Reviews
www.annualreviews.org/aronline

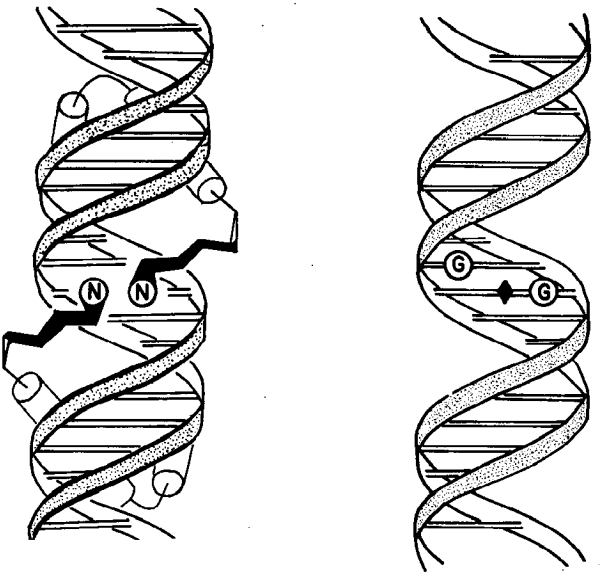

Figure 6 Proposed interaction between lambda repressor's N-terminal arm and the back of the operator site. Panel on right shows guanines that are protected on the back of operator site O_R1.

Except for adjustments in the position of the flexible N-terminal arm, the initial model building (3) simply used the crystallographic coordinates of the N-terminal dimer and searched for an overall fit of the repressor against the DNA. In a subsequent study, Lewis et al (38) made more detailed predictions about the contacts between repressor and the consensus operator site. This phase of model building started with the previous complex but then allowed surface side chains to move and allowed minor adjustments of the relative orientation of the subunits within the dimer. After these adjustments, it appeared that four side chains from each subunit of the dimer could make specific major groove contacts on the front of the operator site. Three of these side chains, Gln44, Ser45, and Ala49, are on α-helix 3, and the fourth, Asn55, is in the irregular region of protein chain just beyond helix 3. The proposed contacts with the consensus site (Figure 1) are summarized below:

1. The side-chain amide of Gln44 makes two contacts with adenine +2. It accepts a hydrogen bond from the N6 of adenine and donates a hydrogen bond to the N7 [Figure 4 (*top*)].
2. The hydroxyl of Ser45 donates a hydrogen bond to the N7 position of guanine −4.

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

Annual Reviews
www.annualreviews.org/aronline

3. The methyl group of Ala49 makes van der Waals contacts with the methyl groups of thymine +3 and thymine −5.
4. The side-chain amide of Asn55 donates two hydrogen bonds to the O6 and N7 positions of guanine −6.

Detailed predictions of contacts that involve repressor's N-terminal arm were not attempted. The conformation of the arm in the crystal seems to be determined by crystal packing forces and there is no reason to believe that it adopts a similar structure in solution or in the repressor-operator complex. In fact, NMR studies show that the arm is flexible in solution (M. Weiss et al, manuscript in preparation). Without reliable structural constraints, detailed model building gives too many possibilities to be useful.

In addition to the major groove contacts, several contacts between repressor and the sugar phosphate backbone of the DNA seem possible (38). Ethylation of any of ten phosphates in the operator interferes with repressor binding (16, 17) and, as shown in Figure 5, these phosphates are symmetrically disposed on the front face of the operator site. Six are clustered near the center of the site and two are near each end of the site. In the proposed complex, Gln33, which is the first residue in helix 2, and Asn52, which is the last residue in helix 3, contact the two phosphates near the outer edge of the operator site. Residues Asn58, Tyr60, and Asn61, in the irregular region between helices 3 and 4, appear to contact the phosphates near the center of the operator site.

None of the proposed contacts between phosphates and amino acids involve ion pairs. However, the Lys24-Lys25-Lys26 sequence, which is part of the loop of irregular chain between helices 1 and 2 (see Figure 5), is near the DNA. These residues could interact with the phosphates on the outer edge of the site if the operator DNA were allowed to partially bend around the protein, in the manner proposed for the Cro-operator complex (23). The salt-dependence of binding (30, 39) suggests that a few ion pairs are formed between the repressor dimer and operator DNA, and residues 24–26 may be responsible for these ion pairs.

Repressor Mutants Defective in DNA Binding

Genetic and biochemical studies of repressor mutants provide strong support for the fundamental features of the proposed repressor-operator complex (40, 41). Twelve mutations, affecting eight residue positions in the N-terminal domain, decrease the operator affinity of repressor but do not disrupt the structure of the mutant N-terminal domain. As shown in Figure 7, seven of the residue positions affected by these "DNA-binding" mutations cluster in the α2-α3 region of the N-terminal domain. Here, the mutations affect each of the four residues predicted to make specific major

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

Annual Reviews
www.annualreviews.org/aronline

groove contacts (Gln44 → Leu/Ser/Tyr; Ser45 → Leu; Ala49 → Val/Asp; Asn55 → Lys); two of the residues predicted to make phosphate contacts (Gln33 → Ser/Tyr; Asn52 → Asp); and one residue that is close to the operator in the proposed complex (Gly43 → Glu). The only "DNA-binding" mutation outside the α2-α3 region alters a residue in the N-terminal arm (Lys4 → Gln).

In Figure 8, the positions of the "DNA-binding" mutations are shown on a space-filling model of the N-terminal dimer. This figure also shows the position of the "DNA contact" residues predicted from model building (3, 38). There is clearly a striking correspondence between the genetic and model-building results. The only significant differences involve residues 58, 60, and 61, which are near the center of the repressor dimer (Figure 8) and which were implicated as backbone contacts by model building. Although mutations at these positions were not obtained in the study cited above, two of these sites were identified by mutations (Asn58 → Ile; Asn61 → Ser) in a study where the mutant proteins were not characterized (42, 43). Thus, almost every residue that has been proposed as a DNA contact residue is altered by one or more repressor mutations. This excellent overall correspondence between the genetic results and the model building would not be expected if the model were seriously wrong, and thus the mutations provide strong experimental support for the model of the repressor-operator complex.

Thus far, we have referred to protein-DNA "contacts" without explicit reference to the energy provided by each contact. This issue has been

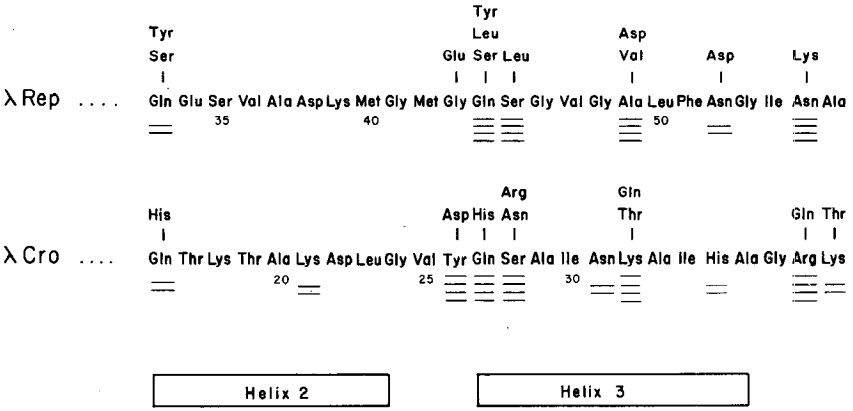

Figure 7 Sequence of the Helix 2/3 Regions from Repressor and Cro. Residues underlined twice are predicted to make contacts with the DNA backbone (23, 38). Those underlined four times are predicted to make major groove contacts. The positions of surface mutations that decrease operator binding are shown.

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

Annual Reviews
www.annualreviews.org/aronline

addressed by studying the affinities of the purified mutant repressors for operator DNA. Mutants may have reduced affinity either because favorable contacts are removed or because unfavorable steric or electrostatic contacts are introduced. Both effects may contribute when the wild-type side chain is replaced by a larger side chain in the mutant (e.g. Ser45 → Leu). However, the Lys4 → Gln, Gln33 → Ser and Gln44 → Ser mutations replace the wild-type side chain with a smaller side chain and thus their decreased operator affinity is likely to reflect the loss of favorable interactions. The operator affinities of these three mutants are about 100-fold less than the wild-type affinity (H. Nelson, R. Sauer, manuscript in preparation). This suggests that the Lys4, Gln33, and Gln44 side chains each contribute about 2.7 kcal/mole of free energy to the interaction between the dimer and the operator. If these energies are additive, then these three side chains contribute a total of about 8 kcal/mole, or half of the 16 kcal/mole free-energy change that occurs upon binding (30). Even if the Lys4, Gln33, and Gln44 contacts are somewhat stronger than average, it seems that the total repressor-operator binding energy can be reasonably explained by the contacts proposed in the model.

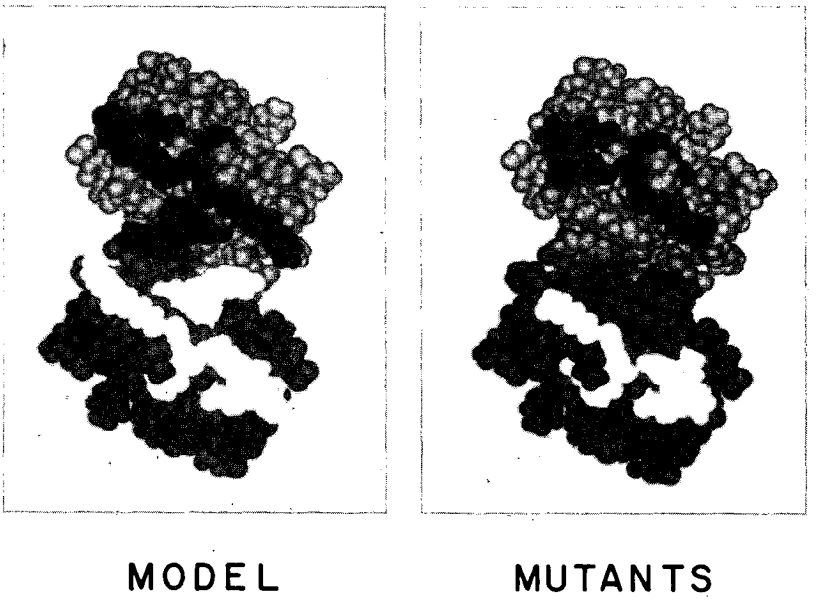

Figure 8 Space-filling models of lambda repressor's N-terminal domain. Residues predicted to contact the operator DNA are highlighted in left panel. Positions of "DNA-binding" mutations are highlighted in right panel. Computer graphics were provided by Richard J. Feldman.

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

Annual Reviews
www.annualreviews.org/aronline

Contacts between a protein side chain and the DNA sugar phosphate backbone are often referred to as "nonspecific" contacts. However, such contacts could readily contribute to the specificity of binding. This point is illustrated by the Gln33 → Ser mutant of lambda repressor. In the proposed complex, Gln33 makes a contact with one of the phosphates near the outer edge of the operator site (3, 38), and substitution of serine at this position reduces the operator affinity of repressor about 100-fold. However, wild-type repressor and the mutant Gln33 → Ser repressor have the same affinity for nonoperator DNA (H. Nelson, R. Sauer, manuscript in preparation). Since specificity depends on the ratio of operator to nonoperator binding (12), Gln33 increases the specificity of operator recognition. How can we rationalize this in molecular terms? One possibility is that Gln33 only contacts the DNA backbone in the specific repressor-operator complex. In complexes of repressor with nonoperator DNA, steric interference in the major groove might prevent Gln33 from approaching the backbone closely enough to make a contact. The Asn52 → Asp repressor mutant also affects a proposed phosphate contact but in this case the mutant has reduced affinity for both operator and nonoperator DNA. Presumably the reduced affinity is caused by electrostatic repulsion between the mutant side chain and the phosphate backbone, and this suggests that Asn52 is close to the DNA in both the operator and nonoperator complexes.

CAP PROTEIN

The catabolite gene activator protein (CAP), also called the cyclic AMP receptor protein (CRP), regulates several catabolite-sensitive gene operons in *E. coli* (7, 44). When cyclic AMP is present at a sufficient concentration, it forms a complex with CAP, and this complex is active in specific DNA binding (45). A consensus sequence has been suggested for the CAP binding sites (46), but many of the individual sites differ considerably from this sequence. The CAP protein contains 209 residues (47, 48) and has two domains (49). The C-terminal domain binds DNA, while the N-terminal domain binds cyclic AMP and provides most of the dimer contacts. CAP is a stable dimer in solution and this dimer is the active DNA binding species (50).

Crystallographic studies by McKay, Weber, & Steitz (2, 51) have determined the structure of the intact CAP dimer in a complex with cAMP. A sketch of the CAP monomer is shown in Figure 9. The N-terminal domain contains 135 residues and consists of a pair of short helices (A and B), an eight-stranded antiparallel β-roll, and a long α-helix (C). The C-terminal domain includes residues 136–209 and contains three α-helices (D, E, and F) and two pairs of short antiparallel β-strands. The CAP

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

Annual Reviews
www.annualreviews.org/aronline

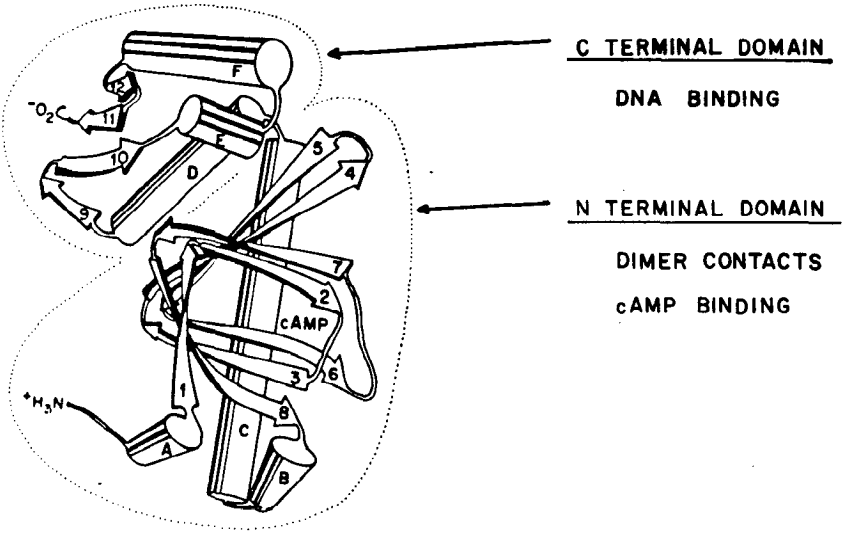

Figure 9 Sketch of the CAP monomer. The approximate extent of each domain is outlined, and the primary functions of each domain are listed. Adapted with permission from (51).

dimer contacts, which are in the N-terminal domain, involve a pairing of the C-helices and some additional contacts between the C-helix of one subunit and the β-roll of the other subunit. The cyclic AMP also occupies part of this dimer interface. It is completely buried within the interior of the N-terminal domain, but it forms hydrogen bonds that bridge the dimer interface. In the crystal form studied, the CAP dimer is somewhat asymmetric, since there are different orientations of the N-terminal and C-terminal domains in the two subunits. One subunit has an "open" conformation with a cleft between the domains, while the other subunit has a "closed" conformation.

Models for CAP-DNA Interactions

Several different models have been proposed for the interaction of CAP with DNA (2, 3, 51–54) but current results suggest that CAP binds to right-handed B-DNA and uses the N-terminal portions of its F-helices to contact the major groove as shown in Figure 10 (3, 54). Calculation of the electrostatic potential at the surface of the CAP dimer provides some support for this model (55), since the only regions of net positive charge are near the amino-terminal portions of the F-helices. Originally, McKay & Steitz (2) had proposed that CAP binds to left-handed B-form DNA. This conformation of DNA (which is quite distinct from Z-DNA) has never been observed. However, it seems conformationally plausible (56) and in model-

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

Annual Reviews
www.annualreviews.org/aronline

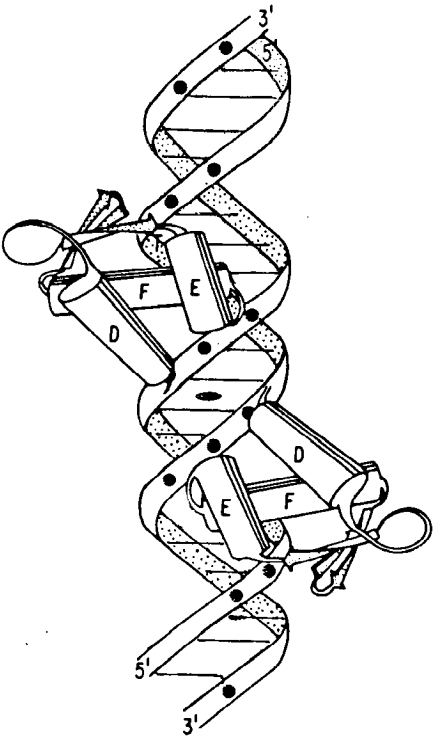

Figure 10 The proposed interaction between the C-terminal domain of CAP and the CAP binding site in the *lac* operon. Black dots indicate contacts with phosphates (45). Adapted with permission from (54).

building experiments, the two F-helices of the CAP dimer can fit neatly into successive major grooves of this left-handed DNA. However, biochemical experiments rule out this model. If CAP were to bind to left-handed DNA, then it should unwind the right-handed DNA found in solution by four turns (1440 degrees). This does not occur. In fact, Kolb & Buc (57) have shown that CAP binding unwinds DNA by no more than 30 degrees.

Several specific CAP-operator interactions have recently been proposed on the basis of model building with right-handed B-DNA (I. Weber, T. Steitz, personal communication). In this model the DNA is bent around the protein with a radius of curvature similar to that predicted for the Cro-operator complex. The specific contacts are listed below using the base numbering scheme of Figure 11:

1. The guanidinium group of Arg180 donates two hydrogen bonds to O6 and N7 of guanine 3 [(Figure 4 (*bottom*)] and donates two hydrogen bonds to the symmetrically related guanine 16.
2. The side chain of Glu181 in one monomer accepts one hydrogen bond

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

Annual Reviews
www.annualreviews.org/aronline

from N4 of cytosine 5 and accepts a second hydrogen bond from N6 of adenine 4. The side chain of Glu181 in the other monomer accepts one hydrogen bond from N6 of adenine 15 and may accept a second hydrogen bond from the N4 of cytosine 14.

3. The amino group of Lys188 donates two hydrogen bonds to O6 and N7 of guanine 5, and two to the symmetrically related guanine 14.
4. Arg185 donates one hydrogen bond to N7 of adenine 6, and donates one hydrogen bond to the symmetrically related adenine 13.

A genetic study of CAP mutants with altered DNA-binding specificity (R. Ebright, J. Beckwith, P. Cossart, B. Gicquel-Sanzey, manuscript in preparation) has suggested specific interactions between CAP and its binding site and also supports the model in which CAP binds to right-handed B-DNA. Figure 11 shows the CAP binding site in the lac operon, and also shows the symmetrically related L8 and L29 mutations in this site. CAP mutants with increased affinity for the L8 site or the L29 site, but with reduced affinity for the wild-type site, were selected and three different mutations were obtained. All three mutations, Glu181 → Leu, Glu181 → Val and Glu181 → Lys, change the same residue in helix F. Since model building suggests that helix F makes major groove contacts, it is likely that Glu181 normally recognizes the G:C base pairs at 5 and 14, while Leu181, Val181, and Lys181 recognize the mutant A:T base pairs at these positions. These contacts can be accommodated in the complex of CAP with right-handed DNA shown in Figure 10, but they would be difficult, if not impossible, to make if CAP bound to left-handed DNA (I. Weber, T. Steitz, personal communications).

Analysis of these CAP mutations also indicates that CAP interacts with its binding site in a symmetric fashion. Two of the CAP mutations were selected using the L8 mutation and one was selected using the L29 mutation. Nevertheless, the mutants bind equally well to the L8 site and the L29 site, which have symmetrically related base changes.

It is not known how cAMP increases the affinity of CAP for its specific DNA sites, but the crystal structure suggests some possibilities and rules

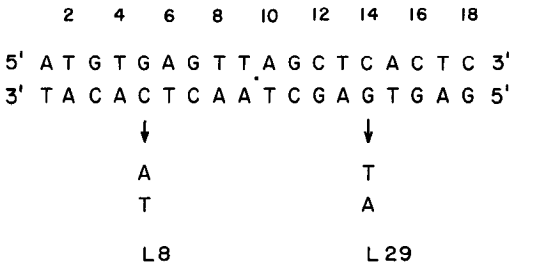

Figure 11 Sequence of the CAP binding site in the *lac* operon.

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

Annual Reviews
www.annualreviews.org/aronline

out others (51). It had been proposed, from cAMP analog studies, that the adenosine ring of cAMP interacts directly with the DNA (52). The structure of the CAP-cAMP complex shows that this proposal must be incorrect, since the cyclic AMP is buried in the interior of the large domain. However, since the buried cAMP interacts with both N-terminal subunits of the dimer, it might affect DNA binding by changing the relative orientation of the two subunits or by changing the orientation of the domains within a subunit.

A CONSERVED α-HELICAL STRUCTURE FOUND IN MANY DNA-BINDING PROTEINS

In the proposed complexes of Cro, repressor, and CAP with DNA, many of the DNA contacts are made by two α-helices that are linked by a tight turn. In both Cro and repressor, these are helices 2 and 3, and in CAP these are designated helices E and F. In each of the three models, the first helix (2 or E) sits above the groove near the DNA backbone while the second helix (3 or F) fits partly or completely into the major groove. This is illustrated in Figure 3.

The structures of these α-helical units within the three proteins are nearly identical. For CAP and Cro, 24 α-carbons from the αE-αF region and 24 α-carbons from the α2-α3 unit can be superimposed with a deviation of only 1.1 Å per α-carbon (58). The agreement between lambda repressor and Cro is slightly better. Here, 20 α-carbons from the two α2-α3 units superimpose with an average deviation of only 0.7 Å (59). Alpha-helices arranged in this way may be unique to DNA-binding proteins as they have not been found in any other protein structures (58, 59).

Superimposing the strictly conserved bihelical unit also reveals a limited structural homology among parts of helix 1 from Cro and repressor and parts of helix D from CAP (58, 59). However, this homology is not extensive and is far less precise than the other homology. In all other regions, the tertiary folds of the three proteins are completely different. It should also be noted that the arrangements of the conserved helical units with respect to the dimer axes are not identical in the three proteins. Thus, the helical units of the protein dimers cannot be superimposed as precisely as the helical units of the monomers, and the different orientations of these regions imply that the proteins could not be docked with their bihelical units contacting the DNA in precisely the same manner.

Sequence Homologies

A number of DNA-binding proteins share sequence homologies with Cro, repressor, and CAP, and several research groups have predicted that these

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

Annual Reviews
www.annualreviews.org/aronline

proteins also use helix-turn-helix structures for DNA interactions (60–62). In some cases, entire protein sequences are homologous. For example, the lambda, P22, 434, and LexA repressor sequences are significantly related (61), and in these cases, the homology almost certainly implies an evolutionary relationship. In other cases, there are only limited regions of homology, but many DNA-binding proteins have regions that are homologous to the α2-α3 sequences of Cro and repressor and the αE-αF sequence of CAP. A set of such sequences is shown in Figure 12.

The pattern of conserved residues and residue types, shown in the alignment of Figure 12, suggests that the homologous proteins also form similar helix-turn-helix structures. Alpha-helices on the surface of a protein often have a characteristic pattern of nonpolar residues that face the hydrophobic core of the protein. Because the helical repeat is 3.6 residues/turn, these residues usually occur at relative helical positions 1-4-5 or 1-2-5 (63). The bihelical units of Cro, repressor and CAP contain such triplets, and in the numbering scheme of Figure 12, these triplets occupy positions 4, 5, and 8 in the first helix and positions 15, 18, and 19 in the second helix. The homologous DNA-binding proteins also tend to have nonpolar residues at these positions and, in addition, have nonpolar residues at position 10, which is part of the hydrophobic core of Cro, repressor and CAP. This means that the homologous proteins could form similar bihelical units and have predominantly nonpolar side chains facing the hydrophobic core. In the proteins of known structure, the residues at positions 1–3, 6–7, 11–14, and 16–17 are solvent exposed and hydrophilic, and the homologous proteins also tend to have hydrophilic residues at these positions. Thus, bihelical units in the homologous proteins would have a number of exposed polar residues that might be used for DNA interactions.

In the alignment of Figure 12, positions 5, 9, and 15 are among the most highly conserved. Alanine is favored at position 5, glycine predominates at 9, and either valine or isoleucine usually occupies 15. Each of these residues seems to have an important role in maintaining the structure of the bihelical unit. In Cro and repressor, the side chains at 5 and 15 in the helix 2/3 unit are in van der Waal's contact and probably help to maintain the proper angle between the two helices. As discussed below, position 9 forms part of the tight turn between the helices. If the homologous sequences also form bihelical structures, then the strong conservation at positions 5, 9, and 15 could be rationalized in structural terms.

The highly conserved glycine at position 9 of the alignment (Figure 12), illustrates an interesting problem in trying to predict structural homology from sequence homology. Originally, it was thought that glycine was required at this position, and most listings of homologies excluded

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

Annual Reviews
www.annualreviews.org/aronline

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
...Gln-	Glu-	Ser-	Val-	Ala-	Asp-	Lys-	Met-	Gly-	Met-	Gly-	Gln-	Ser-	Gly-	Val-	Gly-	Ala-	Leu-	Phe-	Asn...	λ Rep
	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52
					***				***						***			***		
...Gln-	Thr-	Lys-	Thr-	Ala-	Lys-	Asp-	Leu-	Gly-	Val-	Tyr-	Gln-	Ser-	Ala-	Ile-	Asn-	Lys-	Ala-	Ile-	His...	λ Cro
	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
					***				***						***			***		
...Gln-	Ala-	Ala-	Leu-	Gly-	Lys-	Met-	Val-	Gly-	Val-	Ser-	Asn-	Val-	Ala-	Ile-	Ser-	Gln-	Trp-	Gln-	Arg...	P22 Rep
	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
					***				***						***			***		
...Gln-	Arg-	Ala-	Val-	Ala-	Lys-	Ala-	Leu-	Gly-	Ile-	Ser-	Asp-	Ala-	Ala-	Val-	Ser-	Gln-	Trp-	Lys-	Glu...	P22 Cro
	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
					***				***						***			***		
...Gln-	Ala-	Glu-	Leu-	Ala-	Gln-	Lys-	Val-	Gly-	Thr-	Thr-	Gln-	Gln-	Ser-	Ile-	Glu-	Gln-	Leu-	Glu-	Asn...	434 Rep
	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
					***				***						***			***		
...Gln-	Thr-	Glu-	Leu-	Ala-	Thr-	Lys-	Ala-	Gly-	Val-	Lys-	Gln-	Gln-	Ser-	Ile-	Gln-	Leu-	Ile-	Glu-	Ala...	434 Cro
	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
					***				***						***			***		
...Arg-	Gln-	Glu-	Ile-	Gly-	Gln-	Ile-	Val-	Gly-	Cys-	Ser-	Arg-	Glu-	Thr-	Val-	Gly-	Arg-	Ile-	Leu-	Lys...	CAP
	169	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188
					***				***						***			***		
...Arg-	Gly-	Asp-	Ile-	Gly-	Asn-	Tyr-	Leu-	Gly-	Leu-	Thr-	Val-	Glu-	Thr-	Ile-	Ser-	Arg-	Leu-	Leu-	Gly...	Fnr
	187	188	189	190	191	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206
					***				***						***			***		
...Leu-	Tyr-	Asp-	Val-	Ala-	Glu-	Tyr-	Ala-	Gly-	Val-	Ser-	Tyr-	Gln-	Thr-	Val-	Ser-	Arg-	Val-	Val-	Asn...	Lac R
	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
					***				***						***			***		
...Ile-	Lys-	Asp-	Val-	Ala-	Arg-	Leu-	Ala-	Gly-	Val-	Ser-	Val-	Ala-	Thr-	Val-	Ser-	Arg-	Val-	Ile-	Asn...	Gal R
	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
					***				***						***			***		
...Thr-	Glu-	Lys-	Thr-	Ala-	Glu-	Ala-	Val-	Gly-	Val-	Asp-	Lys-	Ser-	Gln-	Ile-	Ser-	Arg-	Trp-	Lys-	Arg...	λ cII
	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43
					***				***						***			***		
...Gln-	Arg-	Lys-	Val-	Ala-	Asp-	Ala-	Leu-	Gly-	Ile-	Asn-	Glu-	Ser-	Gln-	Ile-	Ser-	Arg-	Trp-	Lys-	Gly...	P22 c1
	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
					***				***						***			***		
...Lys-	Glu-	Glu-	Val-	Ala-	Lys-	Lys-	Cys-	Gly-	Ile-	Thr-	Pro-	Leu-	Gln-	Val-	Arg-	Val-	Trp-	Cys-	Asn...	Mat α
	117	118	119	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136
					***				***						***			***		
...Thr-	Arg-	Lys-	Leu-	Ala-	Gln-	Lys-	Leu-	Gly-	Val-	Glu-	Gln-	Pro-	Thr-	Leu-	Tyr-	Trp-	His-	Val-	Lys...	Tet R Tn10
	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46
					***				***						***			***		
...Thr-	Arg-	Arg-	Leu-	Ala-	Glu-	Arg-	Leu-	Gly-	Val-	Gln-	Gln-	Pro-	Ala-	Leu-	Tyr-	Trp-	His-	Phe-	Lys...	Tet R pSC101
	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46
					***				***						***			***		
...Gln-	Arg-	Glu-	Leu-	Lys-	Asn-	Glu-	Leu-	Gly-	Ala-	Gly-	Ile-	Ala-	Thr-	Ile-	Thr-	Arg-	Gly-	Ser-	Asn...	Trp Rep
	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85
					***				***						***			***		
...Arg-	Gln-	Gln-	Leu-	Ala-	Ile-	Ile-	Phe-	Gly-	Ile-	Gly-	Val-	Ser-	Thr-	Leu-	Tyr-	Arg-	Tyr-	Phe-	Pro...	H-inversion
	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180	181
					***				***						***			***		
...Ala-	Thr-	Glu-	Ile-	Ala-	His-	Gln-	Leu-	Ser-	Ile-	Ala-	Arg-	Ser-	Thr-	Val-	Tyr-	Lys-	Ile-	Leu-	Glu...	Tn3 Resolvase
	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180
					***				***						***			***		
...Ala-	Ser-	His-	Ile-	Ser-	Lys-	Thr-	Met-	Asn-	Ile-	Ala-	Arg-	Ser-	Thr-	Val-	Tyr-	Lys-	Val-	Ile-	Asn...	γδ Resolvase
	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180
					***				***						***			***		
...Ile-	Ala-	Ser-	Val-	Ala-	Gln-	His-	Val-	Cys-	Leu-	Ser-	Pro-	Ser-	Arg-	Leu-	Ser-	His-	Leu-	Phe-	Arg...	Ara C
	196	197	198	199	200	201	202	203	204	205	206	207	208	209	210	211	212	213	214	215
					***				***						***			***		
...Arg-	Ala-	Glu-	Ile-	Ala-	Gln-	Arg-	Leu-	Gly-	Phe-	Arg-	Ser-	Pro-	Asn-	Ala-	Ala-	Glu-	Glu-	His-	Leu...	Lex R
	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47
					***				***						***			***		

Figure 12 Sequences homologous to the α2-α3 sequences of lambda Cro and lambda repressor, and the αE-αF sequence of CAP. Sequence references: Fnr protein (92); tetracyline repressors from Tn10 and pSC101 (T. Nguyen, K. Postle, K. Bertrand, manuscript in preparation); H-inversion protein (93); transposon Tn3 resolvase (94); transposon gamma-delta resolvase (95); and arabinose C protein (96). Citations for other sequences are listed in (60–62).

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

Annual Reviews
www.annualreviews.org/aronline

sequences like AraC and Tn3 repressor for this reason. However, it has been shown that a double mutant of lambda repressor, containing Asp38 → Asn (position 6) and Gly41 → Glu (position 9) is functional (64). This proves that formation of the proper turn between helices 2 and 3 does not require glycine. What then do we make of the strong conservation of glycine at position 9? In Cro, this glycine assumes a backbone conformation (phi = 60, psi = 44) commonly observed for glycine but only rarely observed for all other amino acids (65). The side-chain hydrogen atom of glycine allows it to readily assume this conformation, whereas the larger side chains of other residues cause some steric hindrance. This suggests that residues other than glycine might be accommodated at position 9 but at the cost of some modest conformational strain.

For some of the homologous proteins, further evidence suggests that they actually do form bihelical units like those of Cro, repressor, and CAP. Circular dichroism shows that lac repressor, lambda cII protein, and P22 repressor contain substantial regions of α-helix in their DNA-binding domains (66–68) and there are mutants of each protein whose properties can be explained by the proposed helix-turn-helix model (64, 69, 70; Y. Ho, M. Rosenberg, D. Wulff, unpublished). The strongest physical evidence in support of a bihelical unit is for lac repressor. Here, NMR studies have identified tertiary interactions that are predicted by the model (71) and have identified two linked α-helices in the predicted positions (72).

It seems very likely that many DNA-binding proteins use helix-turn-helix units and the question even arises whether there are any specific DNA-binding proteins that do not use bihelical units, or at least α-helical regions, for recognition. There are many DNA-binding proteins that lack obvious homology with the sequences shown in Figure 12. For example, the Mnt and Arc repressors of bacteriophage P22 are two such proteins (73). However, structural homology can be present in proteins that lack sequence homology (74) and circular dichroism studies show that both of these small DNA-binding proteins are substantially α-helical (A. Vershon, P. Youderian, M. Susskind, R. Sauer, manuscript in preparation). Structural studies will be required to determine whether the α-helices of these proteins are used for DNA binding and, if so, whether the helical regions resemble those of CAP, Cro, and lambda repressor.

USE OF α-HELICES IN DNA RECOGNITION

Several early model-building studies predicted that α-helices could fit into the major groove of B-form DNA (75–78), and the structural information now available shows that this is a common mode of DNA recognition. An α-helix with side chains has a diameter of about 12 Å while the major groove

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

Annual Reviews
www.annualreviews.org/aronline

of B-DNA is about 12 Å wide and 6–8 Å deep. Thus, one side of an α-helix can fit snuggly into the major groove. The α-helical backbone can be viewed as a scaffold from which side chains can contact the edges of the base pairs in the major groove. The number of base pairs that could be contacted by a single α-helix depends on the orientation of the α-helix with respect to the groove and on the length of the side chains. Maximum contact is obtained if the helix is parallel to the local direction of the major groove (Figure 3), and with this arrangement side chains on the helix could contact four to six base pairs. (More extensive contacts are not possible because the α-helix is straight whereas the major groove curves away in both directions.) However, model building suggests that the α-helix does not need to be precisely parallel to the major groove. A variety of different orientations would still be sterically reasonable and would allow extensive contacts with the DNA. For example, an α-helix can be positioned with one of its ends, rather than its center, closest to the double-helical axis of the DNA and the helix can be tilted by 15–20 degrees with respect to the major groove. This arrangement, which is similar to the one proposed for lambda repressor, still allows the helix to contact four or five base pairs. An arrangement with the N-terminal rather than the C-terminal end closest to the groove is probably preferred, since an α-helix has a partial positive charge at its N-terminal end, and the major groove carries a partial negative charge (3, 79, 80). Moreover, since the side chains of an α-helix point toward the N-terminal end, this arrangement should help orient the side chains for interactions with the major groove.

Since lambda repressor and Cro recognize the same operator sites, it is instructive to compare the way in which they use their α-helices for DNA binding. Alpha-helix 3 in the Cro-operator complex is almost parallel to the major groove, but its N-terminal end is somewhat closer to the DNA than its C-terminal end (23). This is clear from the pattern of side-chain contacts. Gln27, the first residue in helix 3, fits directly into the major groove and contacts the edge of a base pair. However His35, near the C-terminus of helix 3, is farther from the groove and appears to contact a phosphate. In the repressor-operator complex (38), helix 3 is not as closely parallel to the major groove, but the overall arrangement, including the pattern of side-chain contacts, is similar. For example, as shown in Figure 7, each residue position in the α2-α3 region of repressor that is proposed as a specific or backbone contact is also proposed to make a similar type of contact in Cro. However, the individual side chain contacts made by the two proteins are actually quite different. Half of the common contact positions have different residues and even where the contact residues are identical there are important differences in the proposed complexes. Consider the contacts made by the Gln27-Ser28 side chains of Cro and the homologous Gln44-Ser45

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

Annual Reviews
www.annualreviews.org/aronline

side chains of repressor (Figure 7). The glutamines are both predicted to make the same contacts with adenine +2 [Figure 4 (*top*); (23, 38)]. However the contacts predicted for the serines are different. Ser28 of Cro appears to contact adenine −3 [Figure 4 (*center*)], whereas Ser45 of respressor appears to donate a single hydrogen bond to the N7 of guanine −4. These different predictions arise from differences in the structures of the two protein dimers. In the Cro dimer the α-carbons of Ser28 are some 6 Å farther apart than are the α-carbons of Ser45 in the repressor dimer (59). The α-carbons of Gln27 (in Cro) and Gln44 (in repressor) are also separated by different distances in the two proteins, but these residues can still make the same contacts with adenine +2 because the glutamine side chains, being longer, can reach this base.

Although helix 3 of Cro, helix 3 of repressor, and helix F of CAP are clearly used for recognition of sites in the major groove, it is less clear what role the preceding helices (2 or E) serve and why the two helices are so strictly conserved as a bihelical unit. In the proposed complexes, the axis of helix 2 or helix E is almost perpendicular to the sugar phosphate backbone and the partial positive charges at the N-terminal ends of these helices are close to the phosphates. This should provide a favorable electrostatic interaction. In addition, Gln16 at the N-terminal end of Cro's helix 2 and the corresponding Gln33 at the N-terminal end of repressor's helix 2 appear to hydrogen bond to the phosphates. As discussed with respect to the Gln33 → Ser mutant of lambda repressor, such backbone contacts appear to be directly responsible for some binding specificity. From a structural point of view, these contacts may serve as "clamps" that keep helix 3 from rolling in the groove. By correctly orienting the helices and side chains in the major groove, the backbone contacts could increase the specificity of the interactions with the base pairs.

OTHER MODES OF INTERACTION

Although the recent structural and genetic studies suggest that most of the site-specific contacts are made by residues from the α-helical regions of Cro, repressor and CAP, some contacts seem to be made by regions with an irregular or extended structure. For example, Arg38 of Cro and Asn55 of repressor are thought to make specific major groove contacts and both lie beyond the C-terminal end of helix 3. Since the Cro and repressor α-helices contact only four to five adjacent bases within the major groove, the use of a contact from a nonhelical region allows each protein to contact an additional contiguous base-pair.

Lambda repressor's flexible N-terminal arm provides another example of an extended structure that is used in protein-DNA recognition. The use of

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

Annual Reviews
www.annualreviews.org/aronline

such a structure was first suggested by Feughelman et al (81), who proposed that an extended polypeptide chain could wrap around a double helix and bind in one of the grooves. As previously noted (37), this use of a flexible arm allows repressor to contact both sides of the DNA helix without creating a large kinetic barrier to association, as the arm can wrap around the DNA after the globular portion of the protein has bound to the front of the operator site. However, the use of flexible binding regions may limit specificity by allowing alternative contacts with different sequences, and may also limit interaction energies (82). For example, since repressor's N-terminal arms are disordered in solution but adopt a specific conformation upon binding the operator, their net binding energies will be reduced by the entropic cost of fixing their positions in the complex. The C-terminal residues of Cro apparently provide a second example of a flexible region of protein that is used to make DNA contacts.

Model-building studies have also suggested that β-sheets might be used to bind double-stranded DNA. The right-handed twist of a β-sheet should allow a pair of antiparallel β-strands to fit into the minor (83, 84) or the major groove (85) of B-form DNA. Although these proposals seem plausible, there is no structural evidence to indicate that β-sheets are actually used in site-specific recognition. Initial inspection of the Cro structure (1) had suggested that the antiparallel β-ribbon that is involved in the Cro dimer contacts might bind to the minor groove. However, more detailed model building studies (23) suggest that the β-ribbon does not lie in the minor groove and suggest that contacts from this region are limited to interactions between the side chains and the phosphate backbone.

PROSPECTS FOR A "RECOGNITION CODE"

Even in the absence of high-resolution information from co-crystals it is possible to make a number of reasonable guesses about the general nature of any "recognition code." The structural information, model-building studies, and genetic data make it almost certain that hydrogen bonds between side chains and the edges of base pairs are responsible for much of the specificity in protein-DNA interactions. This has always seemed reasonable, since hydrogen bonds are highly dependent on the position and orientation of the donor and acceptor groups, and since hydrogen bonds are responsible for specificity in so many other biological interactions.

In principle, it could have been possible that site-specific binding proteins used a simple "recognition code," involving a one-to-one correspondence between the amino acid side chains and the bases in the DNA. For example, since a glutamine side chain in the major groove can make two hydrogen bonds to adenine, and arginine can make two hydrogen bonds to guanine

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

Annual Reviews
www.annualreviews.org/aronline

[see Figure 4 (*top* and *bottom*); and (86)], it was conceivable that glutamine would always be used to recognize an A:T base pair and that arginine would always be used to recognize a G:C pair. However, no one-to-one recognition code is consistent with the current data, and it seems inconceivable that any simple code could have escaped notice with all of the structural and genetic information that is now available.

The proposed hydrogen-bonding interactions for Cro, repressor, and CAP seem to indicate that the "recognition code" is degenerate; i.e. it seems that each base pair can be recognized by several different amino acids and that each amino acid can bind to several different bases. The repressor-operator and Cro-operator models certainly support this idea, since adenine is recognized by glutamine (in both complexes), but it is also recognized by serine (in the Cro complex). Serine binds to adenine (in the Cro model), but serine also binds to guanine (in the repressor model). If specificity is still to be maintained, "degeneracy" of this type implies that the "meaning" of a particular amino acid will depend on the conformation and orientation of the protein backbone. This is not surprising. It actually would be impossible to have any simple repeating pattern of contacts made by one α-helix, since the periodicity of an α-helix has no simple relationship to the periodicity of B-DNA. Thus far, in the three proposed complexes a variety of side chains including those of Gln, Asn, Ser, Tyr, Arg, Lys, Glu, Thr, and His have been used to make hydrogen bonds in the major groove or with the DNA backbone. It is likely that these residues will be commonly used for DNA recognition.

However, van der Waals interactions also seem to be an important part of the recognition process. In repressor, the methyl group of Ala49 makes one of the specific major groove interactions and, in CAP, replacement of Glu181 by Leu or Val changes the specificity of DNA binding. Several of the hydrogen-bonding side chains in Cro also seem to make significant van der Waals contacts with the operator. Although the favorable energies obtained from van der Waals interactions, hydrogen bonding, or electrostatic interactions are important aspects of "recognition," the overall fit of the protein and DNA surfaces is also extremely important. For example, Gly48 in repressor's helix 3 seems to play a passive role in recognition since a larger side chain at this position would cause an unfavorable steric contact between the protein and DNA. Thus the "lock and key" analogy that describes the fit of substrates to enzymes also seems to apply to protein-DNA interactions.

At this stage, it is difficult to guess how many different bonding patterns will be used in recognition, and thus we cannot know how "degenerate" the "recognition code" actually is. However, it is still conceivable that the list of possible interactions will be small enough so that the "code" will have a

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

Annual Reviews
www.annualreviews.org/aronline

predictive value and could be used to locate DNA-binding regions within a protein sequence or to predict the preferred DNA sites to which a particular protein binds.

SUMMARY

Several general principles emerge from the studies of Cro, lambda repressor, and CAP.

1. The DNA-binding sites are recognized in a form similar to B-DNA. They do not form cruciforms or other novel DNA structures. There seem to be proteins that bind left-handed Z-DNA (87) and DNA in other conformations, but it remains to be seen how these structures are recognized or how proteins recognize specific sequences in single-stranded DNA.
2. Cro, repressor, and CAP use symmetrically related subunits to interact with two-fold related sites in the operator sequences. Many other DNA-binding proteins are dimers or tetramers and their operator sequences have approximate two-fold symmetry. It seems likely that these proteins will, like Cro, repressor, and CAP, form symmetric complexes. However, there is no requirement for symmetry in protein-DNA interactions. Some sequence-specific DNA-binding proteins, like RNA polymerase, do not have symmetrically related subunits and do not bind to symmetric recognition sequences.
3. Cro, repressor, and CAP use α-helices for many of the contacts between side chains and bases in the major groove. An adjacent α-helical region contacts the DNA backbone and may help to orient the "recognition" helices. This use of α-helical regions for DNA binding appears to be a common mode of recognition.
4. Most of the contacts made by Cro, repressor, and CAP occur on one side of the double helix. However, lambda repressor contacts both sides of the double helix by using a flexible region of protein to wrap around the DNA.
5. Recognition of specific base sequences involves hydrogen bonds and van der Waals interactions between side chains and the edges of base pairs. These specific interactions, together with backbone interactions and electrostatic interactions, stabilize the protein-DNA complexes.

The current models for the complexes of Cro, repressor, and CAP with operator DNA are probably fundamentally correct, but it should be emphasized that model building alone, even when coupled with genetic and biochemical studies, cannot be expected to provide a completely reliable "high-resolution" view of the protein-DNA complex. For example, the use

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

Annual Reviews
www.annualreviews.org/aronline

of standard B-DNA geometry for the operator is clearly an approximation. Recent studies of B-DNA duplexes have revealed sequence-dependent variations in local structure that could affect protein recognition (88–90). Small changes in protein structure, which may occur upon binding to the DNA, could also affect the detailed structure of the complex. Crystallographic studies of the Cro-operator (91) and repressor-operator complexes (38), which are now in progress, should be extremely helpful in evaluating and refining the current models.

ACKNOWLEDGMENTS

We thank our many colleagues for advice, helpful discussions, and unpublished information. Work performed in our laboratories was supported by grants to C. O. P. from the American Cancer Society and the NIH (GM-31471) and to R. T. S. from the NIH (AI-16892, AI-15706).

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

Literature Cited

1. Anderson, W. F., Ohlendorf, D. H., Takeda, Y., Matthews, B. W. 1981. *Nature* 290:754–58
2. McKay, D. B., Steitz, T. A. 1981. *Nature* 290:744–49
3. Pabo, C. O., Lewis, M. 1982. *Nature* 298:443–47
4. Ptashne, M., Jeffrey, A., Johnson, A. D., Maurer, R., Meyer, B. J., et al. 1980. *Cell* 19:1–11
5. Johnson, A. D., Poteete, A. R., Lauer, G., Sauer, R. T., Ackers, G., Ptashne, M. 1981. *Nature* 294:217–23
6. Ptashne, M., Johnson, A. D., Pabo, C. O. 1982. *Sci. Am.* 247:128–40
7. Adhya, S., Garges, S. 1982. *Cell* 29:287–89
8. Ohlendorf, D. H., Matthews, B. W. 1983. *Ann. Rev. Biophys. Bioeng.* 12:259–84
9. Gussin, G., Johnson, A. D., Pabo, C. O., Sauer, R. T. 1983. In *Lambda II*, ed. R. W. Hendrix, J. W. Roberts, F. W. Stahl, R. A. Weisberg, pp. 93–121. Cold Spring Harbor, NY: Cold Spring Harbor Press
10. Takeda, Y., Ohlendorf, D. H., Anderson, W. F., Matthews, B. W. 1983. *Science* 221:1020–26
11. von Hippel, P. H., Bear, D. G., Morgan, W. D., McSwiggen, J. A. 1984. *Ann. Rev. Biochem.* 53:389–446
12. von Hippel, P. H. 1979. In *Biological Regulation and Development*, ed. R. F. Goldberger, 1:279–347. New York: Plenum
13. Helene, C., Lancelot, G. 1982. *Prog. Biophys. Mol. Biol.* 39:1–68
14. Maniatis, T., Ptashne, M., Backman, K. C., Kleid, D., Flashman, S., Jeffrey, A., Maurer, R. 1975. *Cell* 5:109–13
15. Johnson, A. D., Meyer, B. J., Ptashne, M. 1978. *Proc. Natl. Acad. Sci. USA* 75:1783–87
16. Johnson, A. D. 1980. PhD thesis. Harvard Univ., Cambridge, Mass.
17. Johnson, A. D. 1983. *J. Mol. Biol.* In press
18. Hsiang, M. W., Cole, R. D., Takeda, Y., Echols, H. 1977. *Nature* 270:275–77
19. Roberts, T. M., Shimatake, H., Brady, C., Rosenberg, M. 1977. *Nature* 270:274–75
20. Takeda, Y., Folkmanis, A., Echols, H. 1977. *J. Biol. Chem.* 252:6177–83
21. Watson, J. D., Crick, F. H. C. 1953. *Nature* 171:736–38
22. Arnott, S., Hukins, D. W. L. 1972. *Biochem. Biophys. Res. Commun.* 47:1504–9
23. Ohlendorf, D. H., Anderson, W. F., Fisher, R. G., Takeda, Y., Matthews, B. W. 1982. *Nature* 298:718–23
24. Jack, A., Levitt, M. 1971. *Acta Cryst. A* 34:931–35
25. Wang, J. 1979. *Proc. Natl. Acad. Sci. USA* 76:200–3
26. Bloomfield, V. A., Crothers, D. M., Tinoco, I. 1974. *Physical Chemistry of Nucleic Acids.* New York: Harper & Row
27. Matthew, J. B., Richards, F. M. 1982. *Biochemistry* 21:4989–99
28. Ohlendorf, D. H., Anderson, W. F., Takeda, Y., Matthews, B. W. 1983. *J. Biomol. Struct. Dynam.* In press
29. Chadwick, P., Pirrotta, V., Steinberg, R., Hopkins, N., Ptashne, M. 1970. *Cold*

Annual Reviews
www.annualreviews.org/aronline

Spring Harbor Symp. Quant. Biol. 35: 283–94
30. Sauer, R. T. 1979. *Molecular Characterization of the Lambda Repressor and its Gene cI*. PhD thesis. Harvard Univ., Cambridge, Mass.
31. Johnson, A. D., Meyer, B. J., Ptashne, M. 1979. *Proc. Natl. Acad. Sci. USA* 76: 5061–5
32. Humayun, Z., Kleid, D., Ptashne, M. 1977. *Nucleic Acids. Res.* 4:1595–607
33. Sauer, R. T., Anderegg, R. 1978. *Biochemistry* 17:1092–1100
34. Pabo, C. O., Sauer, R. T., Sturtevant, J. M., Ptashne, M. 1979. *Proc. Natl. Acad. Sci. USA* 76:1608–12
35. Sauer, R. T., Pabo, C. O., Meyer, B. J., Ptashne, M., Backman, K. C. 1979. *Nature* 279:396–400
36. Maniatis, T., Ptashne, M. 1973. *Proc. Natl. Acad. Sci. USA* 70:1531–35
37. Pabo, C. O., Krovatin, W., Jeffrey, A., Sauer, R. T. 1982. *Nature* 298:441–43
38. Lewis, M., Jeffrey, A., Wang, J., Ladner, R., Ptashne, M., Pabo, C. O. 1983. *Cold Spring Harbor Symp. Quant. Biol.* 47: 435–40
39. Record, M. T., Lohman, T. M., deHaseth, P. L. 1976. *J. Mol. Biol.* 107: 145–58
40. Nelson, H. C. M., Hecht, M. H., Sauer, R. T. 1983. *Cold Spring Harbor Symp. Quant. Biol.* 47:441–49
41. Hecht, M. H., Nelson, H. C. M., Sauer, R. T. 1983. *Proc. Natl. Acad. Sci. USA* 80: 2676–80
42. Skopek, T. R., Hutchinson, F. 1982. *J. Mol. Biol.* 159:19–33
43. Wood, R. D., Skopek, T. R., Hutchinson, F. 1983. *J. Mol. Biol.* In press
44. de Crombrugghe, B., Busby, S., Buc, H. 1983. In *Biological Regulation and Development*, ed. K. Yamomoto, Vol. 36. New York: Plenum. In press
45. Majors, J. 1978. PhD thesis. Harvard Univ., Cambridge, Mass.
46. Ebright, R. H. 1982. In *Molecular Structure and Biological Function*, ed. J. Griffen, W. Duax, p. 91. New York: Elsevier/North Holland
47. Aiba, H., Fujimoto, S., Ozaki, N. 1982. *Nucleic Acids Res.* 10:1345–61
48. Cossart, P., Gicquel-Sanzey, B. 1982. *Nucleic Acids Res.* 10:1363–68
49. Aiba, H., Krakow, J. S. 1981. *Biochemistry* 20:4774–80
50. Fried, M. G., Crothers, D. M. 1983. *Nucleic Acids Res.* 11:141–58
51. McKay, D. B., Weber, I. T., Steitz, T. A. 1982. *J. Biol. Chem.* 257:9518–24
52. Ebright, R. H., Wong, J. R. 1981. *Proc. Natl. Acad. Sci. USA* 78:4011–15
53. Salemme, F. R. 1982. *Proc. Natl. Acad. Sci. USA* 79:5263–67
54. Steitz, T. A., Weber, I. T. 1983. In *Structures of Biological Molecules and Assemblies*, ed. F. Jurnak, A. McPherson. New York: Wiley. In press
55. Steitz, T. A., Weber, I. T., Matthew, J. B. 1983. *Cold Spring Harbor Symp. Quant. Biol.* 47:419–26
56. Gupta, G., Bansal, M., Sasisekharan, V. 1980. *Proc. Natl. Acad. Sci. USA* 77: 6486–90
57. Kolb, A., Buc, H. 1982. *Nucleic Acids Res.* 10:473–85
58. Steitz, T. A., Ohlendorf, D. H., McKay, D. B., Anderson, W. F., Matthews, B. W. 1982. *Proc. Natl. Acad. Sci. USA* 79: 3097–3100
59. Ohlendorf, D. H., Anderson, W., Lewis, M., Pabo, C. O., Matthews, B. W. 1983. *J. Mol. Biol.* 169:757–69
60. Matthews, B. W., Ohlendorf, D. H., Anderson, W. F., Takeda, Y. 1982. *Proc. Natl. Acad. Sci. USA* 79:1428–32
61. Sauer, R. T., Yocum, R. R., Doolittle, R. F., Lewis, M., Pabo, C. O. *Nature* 298: 447–451
62. Weber, I. T., McKay, D. B., Steitz, T. A. 1982. *Nucleic Acids Res.* 10:5085–5102
63. Schiffer, M., Edmundson, A. B. 1967. *Biophys. J.* 7:121–35
64. Hochschild, A., Irwin, N., Ptashne, M. 1983. *Cell* 32:319–25
65. Ohlendorf, D. H., Anderson, W. F., Matthews, B. W. 1983. *J. Mol. Evol.* 19: 109–14
66. Geisler, N., Weber, K. 1977. *Biochemistry* 16:938–43
67. Ho, Y., Lewis, M., Rosenberg, M. 1982. *J. Biol. Chem.* 257:9128–34
68. Sauer, R. T., Nelson, H. C. M., Hehir, K., Hecht, M., Gimble, F. S., DeAnda, J., Poteete, A. R. 1983. *J. Biomol. Struct. Dynam.* In press
69. Miller, J. H. 1978. In *The Operon*, ed. J. H. Miller, W. Reznikoff, pp. 31–88. Cold Spring Harbor, NY: Cold Spring Harbor Lab.
70. Beyreuther, K. 1978. See Ref. 69, pp. 123–54
71. Arndt, K. T., Boschelli, F., Lu, P., Miller, J. H. 1981. *Biochemistry* 20:6109–18
72. Zuiderweg, E. R. P., Kaptein, R., Wuthrich, K. 1983. *Proc. Natl. Acad. Sci. USA* 80:5837–41
73. Sauer, R. T., Krovatin, W., DeAnda, J., Youderian, P., Susskind, M. M. 1983. *J. Mol. Biol.* 168:699–713
74. Matthews, B. W., Grutter, M. G., Anderson, W. F., Remington, S. J. 1981. *Nature* 290:334–35
75. Zubay, G., Doty, P. 1959. *J. Mol. Biol.* 1:1–20
76. Sung, M. T., Dixon, G. H. 1970. *Proc. Natl. Acad. Sci. USA* 67:1616–23
77. Adler, K., Beyreuther, K., Fanning, E.,

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

Annual Reviews
www.annualreviews.org/aronline

Geisler, N., Gronenborn, B., et al. 1972. *Nature* 237:322–27
78. Warrant, R. W., Kim, S.-H. 1978. *Nature* 271:130–35
79. Hol, W. G. J., van Duijnen, P. T., Berendsen, H. J. C. 1978. *Nature* 273: 443–47
80. Pullman, A., Pullman, B. 1981. *Quart. Rev. Biophys.* 14:289–380
81. Feughelman, M., Langridge, R., Seeds, W. E., Stokes, A. R., Wilson, H. R., et al. 1955. *Nature* 175:38
82. Huber, R. 1979. *Trends Biochem. Sci.* 4:271–76
83. Carter, C. W., Kraut, J. 1974. *Proc. Natl. Acad. Sci. USA* 71:283–87
84. Church, G., Sussman, J. L., Kim, S.-H. 1977. *Proc. Natl. Acad. Sci. USA* 74: 1458–62
85. Blake, C. C. F., Oatley, S. J. 1977. *Nature* 268:115–20
86. Seeman, N. C., Rosenberg, J. M., Rich, A. 1976. *Proc. Natl. Acad. Sci. USA* 73: 804–8
87. Nordheim, A., Tesser, P., Azorin, F., Kwon, Y. H., Moller, A., Rich, A. 1982. *Proc. Natl. Acad. Sci. USA* 79:7729–33
88. Dickerson, R. E., Drew, H. R. 1981. *J. Mol. Biol.* 149:761–86
89. Dickerson, R. E. 1983. *J. Mol. Biol.* 166:419–41
90. Patel, D. J., Ikuta, S., Kozlowski, S., Itakura, K. 1983. *Proc. Natl. Acad. Sci. USA* 80:2184–88
91. Anderson, W. F., Cygler, M., Vandonfelaar, M., Ohlendorf, D., Matthews, B. W., et al. 1983. *J. Mol. Biol.* 168: In press
92. Shaw, D. J., Guest, J. R. 1982. *Nucleic Acids Res.* 10:6119–30
93. Simon, M., Zieg, J., Silverman, M., Mandel, G., Doolittle, R. F. 1981. *Science* 209:1370–74
94. Heffron, F., McCarthy, B. J., Ohtsubo, H., Ohtsubo, E. 1979. *Cell* 18:1153–63
95. Reed, R. R., Shibuya, G. I., Steitz, J. A. 1982. *Nature* 300:381–83
96. Miyada, C. G., Horwitz, A. H., Cass, L., Timko, J., Wilcox, G. 1980. *Nucleic Acids Res.* 8:5267–65

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

Annual Reviews of Biochemistry
Volume 53, 1984

CONTENTS

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.

[1] As noted in the preface (p. vi) to Volume 52 (1983), typesetting problems resulted in the omission of an author index from that volume. That author index appears as an extra index here in Volume 53.

Annu. Rev. Biochem. 1984.53:293-321. Downloaded from arjournals.annualreviews.org
by State University of New York - Buffalo on 01/16/06. For personal use only.