

**CRYSTALLOGRAPHIC STUDIES ON THE PURINE REPRESSOR, AN
ESCHERICHIA COLI DNA-BINDING
PROTEIN**

by

Maria A. Schumacher

A DISSERTATION

**Presented to the Department of Biochemistry and Molecular Biology
and the Oregon Health Sciences University
School of Medicine
in partial fulfillment of
the requirements for the degree of**

Doctor of Philosophy

April 24, 1995

APPROVED:

[Redacted Signature]

(Professor in Charge of Thesis)

[Redacted Signature]

(Chairman, Graduate Council)

TABLE OF CONTENTS

| | |
|---|-----|
| LIST OF FIGURES | v |
| ACKNOWLEDGEMENTS | vii |
| ABSTRACT | ix |
| I. INTRODUCTION | |
| A. Transcriptional regulation in <i>Escherichia coli</i> | 1 |
| B. The purine repressor: master regulator of <i>de novo</i> purine biosynthesis in <i>E. coli</i> | |
| 1. Characterization | |
| a. The cloning of the purine repressor gene and identification of corepressors | 5 |
| b. Description of the <i>pur</i> regulon | 7 |
| C. DNA recognition by regulatory proteins and transcription factors..... | 12 |
| 1. Major groove recognition | |
| a. The helix-turn-helix motif | 14 |
| b. Zinc binding domains | 23 |
| c. b/ZIP, b/HLH and b/HLH/ZIP motifs | 33 |
| d. β Sheet binding proteins | 38 |
| e. Others | 42 |
| 2. Minor groove recognition | |
| a. Homeodomain N-terminal arms | 44 |
| b. TATA-binding protein | 46 |
| D. Structure determination by X-ray crystallography | 48 |
| E. References | 73 |
| II. COLLECTION OF PAPERS | |
| Manuscript #1. | |
| Structural Analysis of the Purine Repressor, an <i>Escherichia coli</i> DNA-binding Protein..... | 123 |
| Manuscript #2. | |
| Crystallization and Preliminary X-ray Analysis of an <i>Escherichia coli</i> Purine Repressor-Hypoxanthine-DNA Complex. | 151 |

| | |
|--|-----|
| Manuscript # 3. | |
| Crystal Structure of the LacI Member, PurR, Bound to DNA: Minor Groove Binding by α -Helices. | 161 |
| Manuscript #4. | |
| Crystallization and Preliminary X-ray Studies on the Co-repressor Binding Domain of the <i>Escherichia coli</i> Purine Repressor. | 200 |
| Manuscript #5. | |
| Unliganded Purine Repressor has an Open Conformation: Implications for DNA Binding..... | 207 |
| | |
| III. CONCLUSION AND DISCUSSION | 233 |
| | |
| IV. APPENDIX I. | |
| Manuscript #6. | |
| Structures of Potential Intermediates of the T to R Transition of Hemoglobin. | 243 |

LIST OF FIGURES

Introduction

| | |
|--|-----|
| Figure 1. Alignment of the proteins of the LacI family and the D-ribose binding protein | .85 |
| Figure 2. Outline of the de novo purine biosynthetic pathway. | 88 |
| Figure 3. Schematic diagram of PurR regulated <i>pur</i> regulon | 90 |
| Figure 4. Examples of helix-turn-helix (HTH) DNA-binding proteins. 4A) λ repressor N-terminal DNA binding domain -DNA structure. 4B). CAP-DNA structure. 4C) Oct-1 POU domain-DNA structure..... | 92 |
| Figure 5. Stereo diagram of canonical helix-turn-helix (HTH) motif..... | 95 |
| Figure 6. λ repressor-DNA contacts | 97 |
| Figure 7. Schematic drawings illustrating the known subfamilies of structural winged-helix DNA binding domains..... | 99 |
| Figure 8. Examples of zinc domain DNA binding proteins A) Zif268-DNA structure. B)glucocorticoid-DNA structure. C) GAL4-DNA structure D) HIV-F1-DNA structure. E) GATA-1-DNA structure..... | 101 |
| Figure 9. Zif268 DNA contacts | 105 |
| Figure 10. Examples of b/ZIP and b/HTH DNA binding domains. A) GCN4-AP1 structure. B) MyoD-DNA structure | 107 |
| Figure 11. b/ZIP and b/HLH contacts compared. A) GCN4-AP1 contacts. B) MyoD-DNA contacts | 109 |
| Figure 12. Examples of β -sheet DNA binding proteins. A) MetJ-DNA structure showing one of two crystallographically related dimers. B) Arc-DNA structure | 111 |
| Figure 13. Schematic diagram comparing MetJ and Arc DNA contacts..... | 113 |
| Figure 14 A) Bragg's Law B) Ewald's Sphere | 115 |
| Figure 15. Reciprocal lattice separations by similar triangles | 117 |
| Figure 16. Eulerian angles α , β , γ used in the rotation function to relate the rotated axes X_1' , X_2' , X_3' to the original unrotated angles, X_1 , X_2 , X_3 | 119 |
| Figure 17 A) Diagram showing vector nature of F_H , F_{PH} and F_P and the lack of closure error. B) Harker construction showing resulting phase ambiguity from SIR and its resolution by MIR | 121 |

Mauscript #1

| | |
|--|-----|
| Figure 1. Alignment of PurR against RBP and GGBP-E | 139 |
| Table I. Identity scores of contiguous segments of RBP and the CBD of PurR | 141 |

| | |
|--|-----|
| Figure 2. Alignment of PurR against <i>E. coli</i> DNA-binding proteins, LacI and CytR | 143 |
| Table II. Observed ligand-binding residues of GBP-E and RBP and the corresponding predicted residues in PurR | 145 |
| Figure 3. CD spectra of the <i>E. coli</i> purine repressor protein, its corepressor binding domain, and the <i>E. coli</i> ribose-binding protein | 147 |
| Figure 4. Predicted secondary structure topology of the purine repressor protein | 149 |

Manuscript # 2

| | |
|---|-----|
| Figure 1. Crystals of PurR-hypoxanthine- <i>purF</i> operator complex | 159 |
|---|-----|

Manuscript # 3

| | |
|---|-----|
| Table 1. PurR-hypoxanthine- <i>purF</i> operator crystallographic analysis | 177 |
| Table 2. DNA helical parameters of the <i>purF</i> operator half-site from the complex | 180 |
| Figure 1A) Sequence of the duplex deoxyoligonucleotide used in PurR-hypoxanthine- <i>purF</i> operator cocrystallization. B) Sequence of perfect palindrome <i>purF</i> operator site | 182 |
| Figure 2A) 2F _o - F _c electron density map showing the how the crystallographically related leucine levers intercalate into the DNA minor groove. B) F _o - F _c omit map of same region | 184 |
| Figure 3 A) Topology diagram of the purR monomer. B) Stereo view of the PurR-hypoxanthine-DNA complex | 186 |
| Figure 4. Stereo diagram of the corepressor binding pocket | 189 |
| Figure 5. Superposition of the HTH motifs of PurR and LacI | 191 |
| Figure 6 A) Stereo view of PurR-DNA interactions showing one PurR monomer B) Schematic view of the PurR-palindromic <i>purF</i> operator interaction C) Overall view of the PurR-hypoxanthine-DNA complex featuring the leucine levers D) View of the twofold related hinge helices and the minor groove of the <i>purF</i> operator | 193 |
| Figure 7. LacI family member operator site specificity | 196 |
| Figure 8. LacI mutants mapped onto the three-dimensional structure of PurR | 198 |

Manuscript # 4.

| | |
|---|-----|
| Figure 1. 12° <i>hk0</i> precession photograph of the corepressor free form of the corepressor binding domain of PurR | 205 |
|---|-----|

Manuscript # 5.

| | |
|---|-----|
| Table 1. Corepressor-free CBD crystallographic analysis | 218 |
| Figure 2. $2F_o - F_c$ electron density map of the intersubunit β -sheet formed between CBD monomers in corepressor free form | 221 |
| Figure 1. Ribbon diagrams comparing the corepressor free CBD and the corepressor bound PurR from the PurR-hypoxanthine-DNA structure | 223 |
| Figure 3. Overlays of the COOH-subdomains of the unliganded onto the liganded CBD | 226 |
| Figure 4. Comparison of the corepressor binding pocket of corepressor free and bound forms | 229 |
| Figure 5. Overlay of the CBD NH ₂ -subdomain of corepressor bound PurR, onto each of the NH ₂ -subdomains of unliganded CBD demonstrating the disruption of the hinge helices in corepressor free CBD | 231 |

APPENDIX I.

Manuscript # 6.

| | |
|---|-----|
| Table 1. Selected crystallographic and biochemical data for chemically cross-linked hemoglobins | 250 |
| Figure 1. $2F_o - F_c$ electron density map of cross-linker region in the triply linked hemoglobin, $\alpha_2\beta^{1,82}\text{Tm}^{82}\beta$ | 253 |
| Figure 2. Stereo view of the overlay of the $\alpha_1\beta_1$ interface of deoxyhemoglobin, $\alpha_2\beta^1\text{Tm}^{82}\beta$, $\alpha_2\beta^{1,82}\text{Tm}^{82}\beta$ and $\alpha_2\beta^1\text{S}^{82}\beta$ onto COHbA using the method of Baldwin and Chothia | 255 |
| Figure 3. Coordinate difference plot (CDP) showing the differences between corresponding C α positions of deoxyhemoglobin, $\alpha_2\beta^1\text{Tm}^{82}\beta$, $\alpha_2\beta^{1,82}\text{Tm}^{82}\beta$ and $\alpha_2\beta^1\text{S}^{82}\beta$ after their $\alpha_1\beta_1$ interfaces are overlaid on COHbA as in Figure 2 | 257 |
| Figure 4 A) The α_1 heme environments of deoxyhemoglobin, $\alpha_2\beta^1\text{Tm}^{82}\beta$, $\alpha_2\beta^{1,82}\text{Tm}^{82}\beta$ and $\alpha_2\beta^1\text{S}^{82}\beta$ compared. B) The β_1 heme environments compared C) The switch region compared. D) The pocket formed by the β F and H helices, contrasting the location of βCys93 and βTyr145 in the hemoglobins | 259 |

ACKNOWLEDGMENTS

I would like to thank, first of all, the many collaborators who allowed me to work on such interesting and stimulating projects; Dr. Zalkin and his former graduate student, Kang-Yell Choi who provided the PurR and CBD protein; Dr. Jones and his lab for providing cross-linked hemoglobins and Cai, Debbie Holland and Ray Jacobson, members of Dr. Matthew's lab, for helpful discussions and programs. Dr. Jones merits special thanks for his helpful comments and support throughout the years. I would also like to thank the army for the fellowship they awarded me, which supported my graduate training.

I would like to say a special thanks to John McClure for his wisdom and support . I am also thankful to the Brennan lab, especially Dr. Brennan himself, for providing a pleasant working environment. I am grateful to Dr. Brennan for many "things" including introducing me to the fascinating world of reciprocal space, helping me gain some self-confidence, his sarcastic remarks, his critical reading of this manuscript and for being a friend as well as a terrific advisor. I, like everyone else who has come into contact with Dr. Brennan, have benefitted greatly from the experience.

Two people, who I am dedicating this thesis to, deserve special recognition, my mom, Alice Schumacher and my dad, Alan Schumacher. They have always provided an example that is impeccable and love that is unconditional.

ABSTRACT

The modern field of regulatory biology was founded on studies on the regulation of lactose metabolism by the *Escherichia coli* Lactose repressor. These studies were the first to demonstrate that proteins function in the regulation of gene expression. Since that time, numerous "transcription factors" have been characterized, including several bacterial proteins homologous to the Lac repressor. The structural characterization of transcription factor-DNA complexes, primarily by X-ray crystallography, has greatly expanded our understanding of how proteins can recognize specifically a DNA site. Unfortunately, these studies have not revealed a simple "DNA-binding code". Additionally, the question of how some transcription factors become activated to bind DNA has not been fully addressed. Therefore, more structural studies are required to understand DNA binding and activation at the atomic level. The crystallographic studies described in this dissertation examine how the LacI family member, the *Escherichia coli* purine repressor or PurR, which is the master regulatory protein of *de novo* purine biosynthesis in *E. coli*, specifically recognizes its operator site and how a small molecule effector in the form of a corepressor activates this protein to bind DNA. The results can serve as a model for understanding the progenitor transcription factor, LacI, which has so far proven refractory to structural elucidation, as well as other LacI members.

To elucidate the basis for the specific binding of dimeric PurR to its operator site, the three-dimensional structure of a ternary complex of PurR, hypoxanthine, and a 16 base pair *purF* operator site was determined to 2.7 Å by X-ray crystallography. The structure was determined by multiple isomorphous replacement and reveals that PurR has a bipartite structure in which each monomer subunit contains an NH₂-terminal DNA-binding domain and a larger COOH-terminal corepressor-binding/dimerization domain that is strikingly similar to that of the bacterial periplasmic binding proteins. The DNA-binding domain contains a helix-turn-helix motif which makes base specific contacts with the DNA major groove. Unexpectedly, base contacts are also provided by residues of symmetry-related α helices, the "hinge" helices, which bind deeply in the minor groove of the DNA. Critical to this binding is the partial intercalation of the side chains of Leu54 and Leu54' into the central CpG step of the operator site. This intercalation causes the central base pairs to become unstacked and kinks the DNA site by 45°.

Significantly, in the ternary complex, the corepressor, which activates PurR to bind DNA, is located in a cleft between the two structurally similar subdomains within the corepressor binding domain (CBD), 40 Å from the DNA-binding domain. To address the

question of how corepressor binding affects DNA binding, the structure of the corepressor-free form of the CBD was determined by X-ray crystallography to 2.2 Å by molecular replacement. The structure reveals that the corepressor-free CBD takes an open conformation whereby the NH₂- and COOH-subdomains of each monomer subunit of the CBD dimer are rotated by more than 18° relative to the corepressor-bound CBD. The open conformation is accompanied by a restructuring of the corepressor binding pocket and a rearrangement of the secondary structural elements of the NH₂-subdomain. This conformational change suggests an activation mechanism whereby corepressor binding leads to a juxtaposition of the previously disordered hinge regions of each subunit, allowing them to form helices when in the presence of DNA.

INTRODUCTION

A. Transcriptional regulation in *Escherichia coli*

The metabolism of *E. coli*, or any cell, is largely dependent on the types, and corresponding concentrations, of proteins that are present. Of the 4,000 genes in the *E. coli* genome, only a fraction are expressed at any given time. Therefore, the regulation of gene expression plays a vital role in the proper orchestration of cellular metabolism. The expression of a given gene into protein is accomplished by two processes, which form part of the central dogma of molecular genetics, transcription whereby the coded genetic message in the DNA is converted to the messenger RNA (mRNA) and translation whereby the message encoded in the mRNA is translated on the ribosome into protein (1).

Six points at which the amount of protein produced can be regulated have been identified: synthesis of the primary mRNA transcript (transcription), postranslational processing of mRNA, mRNA degradation, protein synthesis (translation), postranslational protein modification and protein degradation. Of these, transcriptional regulation is the best characterized and appears to be the most common (1). Given the high energetic cost of protein synthesis, regulation at this first step of the process (transcription), would appear to be the most efficient mode for regulation. Indeed, in the gram negative bacterium, *Escherichia coli*, the control of gene regulation is exerted primarily at the level of transcription. It is this regulation which will form the underlying focus of this thesis.

Transcription is mediated and regulated by protein-DNA interactions with the central component in *E. coli* being a DNA-directed RNA polymerase. The *E. coli* RNA polymerase binds to DNA and initiates transcription at specific DNA sites called promoters, which are usually found near the position where mRNA synthesis begins. Unlike eukaryotes, *E. coli* has a single DNA-directed RNA polymerase which is a large multi-protein enzyme containing five core subunits (α , MW: 36,500; β , MW: 151,000; β' , MW: 155,000 and ω , MW: 11,000) with the composition, $\alpha_2\beta\beta'\omega\sigma$. The sixth subunit, σ , binds transiently to the core enzyme and directs it to the promoter. *E. coli* contains several different σ factors. These factors are what determines the promoter specificity or specificity class of a given holoenzyme (2). Sequence analysis has indicated that there are two main families of sigmas in *E. coli*. The first family contains those related to the primary, "house-keeping" *E. coli* σ , σ^{70} . This σ allows RNA polymerase to recognize promoters containing the -10 consensus sequence, TATAAT, and the -35 consensus sequence, TTGACA (3). Another member of this family, σ^{32} , is induced upon heat shock and promotes binding of the RNA polymerase holoenzyme to promoters containing different -35 and -10 elements. The second family of σ factors in *E. coli* is represented by

σ^{54} , a factor involved in the transcription of nitrogen-related genes. Unlike the σ^{70} family, these σ factors recognize DNA elements found -24 and -12 relative to the transcription start site (3). Recently, it has been established that there is a third sequence element, in addition to the -35 and -10 elements, at some *E. coli* σ^{70} promoters that leads to an increase in promoter strength. This sequence corresponds to an AT-rich region of approximately 20 base pairs located immediately upstream of the -35 region and has been called the upstream element, or UP element (4).

Each cycle of transcription in *E. coli* consists of three steps. 1) Initiation, which itself involves several steps including promoter location and recognition by the RNA polymerase holoenzyme, formation of the closed promoter complex, isomerization to the open promoter-holoenzyme complex, initiation of transcription with concomitant release of the specificity factor, σ and formation of the elongation complex. 2) Elongation which is stabilized by the binding of a protein called NusA to the core polymerase. NusA interacts with the core polymerase when σ dissociates and remains bound to the core polymerase until termination. 3) Termination, which as mentioned, is influenced by NusA (3). Of these three steps, the first step, initiation, is the primary site of transcription regulation in *E. coli*. The regulation of initiation is, in effect, the regulation of the interaction of RNA polymerase with its promoter.

Because promoters of a given specificity class vary considerably in their nucleotide sequence, they do not all bind RNA polymerase with equal affinity. At least two types of proteins, in addition to the σ specificity factors, act to regulate transcription initiation by the core RNA polymerase. These regulatory proteins are repressors, which bind to DNA sites near the promoter called operators, and activators which also bind DNA sites near promoters. It is thought that repressors function by blocking access of RNA polymerase to the promoter whereas activators enhance the RNA polymerase-DNA interaction. The binding of repressors to their corresponding DNA sites is often regulated by small molecule effectors that bind to, and presumably, induce a conformational change in the repressor. When binding of a small molecule effector to a repressor activates the repressor to bind DNA, it is called a corepressor. Alternatively, if small molecule binding decreases a repressors affinity for its cognate DNA site, it is called an inducer. Small molecule effector binding to activators, although not as well studied, also acts to regulate the affinity of certain of these proteins for their cognate DNA sites (5).

Certain bacteria, including *E. coli*, have developed a simple mechanism for coordinating the regulation of genes whose products are involved in related pathways by clustering them together and placing them under the control of a single promoter. These gene clusters, their promoter and the additional regulatory sequences are called operons

(6,7). Operons that contain two to six genes are common in *E. coli*, however, operons containing 20 genes or more have been identified. Also common in bacteria are networks of operons under the control of a common regulator. These networks are called regulons (1).

Many of the principles for the regulation of gene expression in bacteria were defined by studies on the regulation of a specific operon, the *lac* operon, in *E. coli*. This operon encodes proteins involved in the metabolism of lactose. In the absence of glucose, lactose can be utilized as the sole carbon source for growth by *E. coli*. In 1961, Francois Jacob and Jacques Monod, demonstrated, and subsequently published, their finding that the genes involved in lactose metabolism are coordinately regulated by a specific genetic element which was located adjacent to them (8). Out of their studies the operon model was born as was the concept of an operator (6-8). This work had a great impact on biochemistry in that it allowed scientists, for the first time, to think about gene regulation in molecular terms. Furthermore, since its discovery, the *lac* operon has been used as a paradigm for transcriptional regulation in eukaryotes as well as prokaryotes.

Continuation of the work of Jacob and Monod, showed that the "adjacent genetic element" near the *lac* operon encodes a tetrameric repressor protein, the Lac repressor (LacI) (9). Under normal growth conditions of high glucose and low lactose, the Lac repressor binds to operator sites, repressing transcription of the *lac* genes. However, under conditions of low glucose and high lactose, the small levels of β -galactosidase present convert the minute concentrations of lactose found in the cell to allolactose. Allolactose then binds to the lac repressor causing it to dissociate from its DNA sites. Thus, allolactose acts as an inducer of the Lac repressor. Three LacI operator sites have been identified, of which two are critical for full repressor activity: O_1 , located 5' to the promoter and O_2 , located within the protein coding region. Studies indicate that DNA looping allows the tetrameric repressor to bind to both of these sites which are separated by approximately 4000 base pairs. O_3 is occupied by repressor only under conditions favoring DNA supercoiling (10). After the lac repressor dissociates from these sites, the *lac* operon genes are expressed and the level of β -galactosidase in the cell increases by a factor of 1,000.

Full regulation of the *lac* operon is further complicated in that it depends not only on the Lac repressor but also on a positive regulatory factor, the catabolite gene activator protein or CAP. CAP binds its operator site in the *lac* operon only in the absence of glucose. This activation is mediated by the corresponding increase in the small molecule effector, cAMP, which binds to and activates CAP. CAP acts as a control element in several other operons encoding enzymes for the metabolism of other secondary sugars. These operons, thus constitute a regulon (11).

Since the identification of the Lac repressor, many bacterial transcriptional regulators, both repressors and activators, have been isolated. Many of these proteins have been cloned allowing for detailed biochemical studies, including crystallographic studies, to be performed. However, the Lac repressor has resisted efforts to obtain its full three-dimensional structure either complexed or uncomplexed to its DNA site, thus preventing a complete understanding of its mechanism of DNA binding and the changes that must occur upon inducer binding to the protein. The finding that a large and growing group of bacterial repressor proteins (now called LacI members) display significant sequence homology to the Lac repressor presents the possibility that a better understanding of all LacI members could be obtained through the structure determination of any one of these proteins. Members of this family now number over 21 and include the Galactose repressor, GalR; Galactose isorepressor, GalS; Amylase repressor, CcpA; Cytidine repressor, CytR; Evolved β -galactosidase repressor, EbgR; Fructose repressor, FruR; Maltose repressor, MalI; Raffinose repressor, RafR; Ribital repressor, RbtR; Sucrose repressor, ScrR; Opine utilization repressor, OpnR and the Purine repressor, PurR, which is the focus of this thesis (12) (Fig. 1).

These proteins have been shown to consist of two domains: An NH_2 -terminal DNA-binding domain, consisting of the first, approximately 60 residues and a larger, COOH-terminal effector-binding/oligomerization domain. The sequences of these proteins show the greatest homology within their NH_2 -terminal DNA-binding domains. This domain has been predicted to contain a helix-turn-helix DNA-binding motif. Less was known about the larger, COOH-terminal effector-binding/oligomerization domain. However, it has been predicted that this domain may bear a striking structural similarity to the unrelated bacterial periplasmic binding proteins which function in the transport of metabolites across the bacterial membrane and, in some cases, chemotaxis (13). In this thesis, crystallographic studies on PurR were undertaken, in part, to address the questions concerning LacI member structure.

B. The purine repressor: master regulator of *de novo* purine biosynthesis in *E. coli*

1. Characterization

a. The cloning of the purine repressor gene and identification of corepressors

Most organisms, with the exception of certain parasitic protozoa, have the ability for *de novo* synthesis of purine nucleotides. This *de novo* pathway requires 10 steps to proceed from 5'-phosphoribosyl 1'-pyrophosphate (PRPP) to inosine monophosphate (IMP) and two additional steps to proceed to either adenosine monophosphate (AMP) or guanosine monophosphate (GMP). Although this pathway is invariant among organisms, the genetic organization and regulation of expression differ. In *E. coli*, the genes encoding enzymes for *de novo* purine biosynthesis have all been isolated and cloned and are found scattered throughout the *E. coli* genome in the form of small polycistronic and monocistronic operons. In total, the 14 genes required for the synthesis of AMP and GMP are grouped into 10 operons (14,15). Analysis of *E. coli* purine regulatory mutants suggested that these operons may constitute a *purR* regulon, i.e., they may be coregulated (16, 17). This was supported by the identification of similar *cis*-acting control regions in several *pur* genes, a finding which suggested that a *trans*-acting regulatory factor may be involved (16-18). This was confirmed when the *E. coli purR* gene was isolated and cloned (19).

PurR was cloned by functional complementation in a *purRpurF-lacZ* fusion strain employing a mini-Mu cloning procedure. In this procedure, a random library of *E. coli* genes was generated from an *E. coli* strain bearing a Mu containing plasmid. This was carried out by increasing the temperature to 42° leading to inactivation of the temperature sensitive Mu repressor and the corresponding derepression of the plasmid-encoded transposase genes. As a result, the mini-Mu element was randomly transposed into the *E. coli* genome. The DNA was then packaged into phage particles. This Mu lysate, which constituted the *E. coli* library, was then used to infect the *purRpurF-lacZ* strain. *purF*, which encodes glutamine PRPP-amidotransferase, the enzyme that carries out the first step in *de novo* purine biosynthesis, was known to be regulated by *purR* and thus served as an indicator of the presence of the *purR* gene. Transformants were screened on X-Gal plates and colonies that were white, indicating that the *purR* gene was present, were picked. After subcloning, the *purR* gene was localized by deletion studies.

An open reading frame in the *purR* nucleotide sequence showed that the *purR* gene encodes a 341 amino acid protein with calculated molecular weight of 38,179. Gel filtration and cross-linking studies showed that PurR exists as a functional dimer. A *purR* probe was generated from the deduced nucleotide sequence and used in hybridization studies showing

that the *purR* gene lies on coordinate kilobase-pair 17,755 on the *E. coli* restriction map, which corresponds to minute 36 on the *E. coli* chromosome (19).

The cloning of *purR* also allowed for the overproduction and subsequent purification of the PurR protein using a phage T7 expression system (20). As will be described, PurR is autoregulated (21). One of the *purR* autoregulatory operator sites is found within the *purR* coding region. A mutation was made to remove this negative control element by changing the sequence GTTTCC to GTCAGT. This mutation abolishes the operator but does not change the Val-Ser amino acid sequence and resulted in increased PurR production in comparison to a nonmutated plasmid control. The resulting overexpressed repressor constitutes approximately 35% of the soluble cellular protein and can be purified to near homogeneity by two chromatographic steps. Six cycles of Edman degradation carried out on the purified PurR protein revealed no heterogeneity and resulted in the sequence, Ala-Thr-Ile-Lys-Asp-Val, indicating that the initiator methionine is cleaved.

In vitro operator binding studies using purified PurR revealed a requirement for something present in cellular extracts for specific binding of PurR to the *purF* site. This suggested that the high affinity binding of PurR to its operator sites may be dependent upon the presence of corepressor(s). Guided by previous studies carried out by Houlberg and Jensen which indicated a requirement of hypoxanthine and guanine for *pur* gene regulation in *Salmonella typhimurium* (22), a survey was conducted of purine bases as well as pyrimidine bases, nucleosides and nucleotides to determine the identity of PurR's corepressor(s). It was found that only hypoxanthine and guanine could restore binding of purified PurR to the *purF* operator *in vitro* (20). The concentrations of hypoxanthine and guanine required for half maximal binding were determined to be 9.3 μM and 1.5 μM , respectively (23). It was later demonstrated that there is only one corepressor binding site per PurR monomer (24).

b. Description of the *pur* regulon

Studies carried out with the purified PurR protein have established that the 14 genes encoding the enzymes required for *de novo* purine biosynthesis are all coregulated by the Pur repressor (18, 25-28). In addition, PurR is autoregulated (21). In *E. coli* there are seven operons required for IMP synthesis; *cvp purF dedF*; *purL*; *purMN*; *purEK*; *purHD*; *purC*; and *purB*. A single *guaBA* operon is required for conversion of IMP to GMP and the two genes, *purA* and *purB*, are required for the conversion of IMP to AMP. Figure 2 outlines the *de novo* purine biosynthetic pathway and gives gene enzyme relationships in *E. coli*. The differential repression of the *de novo* purine genes by PurR is notable. PurR represses transcription of the genes encoding the enzymes required for *de novo* synthesis of IMP by seven to 18-fold whereas PurR repression of the genes encoding the branch point enzymes required for the synthesis of AMP and GMP from IMP are only two and five fold, respectively (15). It has been suggested that this differential repression in *E. coli* may be important in sustaining the separate pathway of purine salvage while the *de novo* purine pathway is shut down, the notion being that after purines are salvaged by the phosphoribosyltransferases (PRPPs), the branch point enzymes could function in the conversion of IMP to AMP and/or GMP.

The principal features that characterize the promoters and *cis*-acting control sites of the above PurR regulated genes are a -10 promoter which matches the TATAAT σ^{70} RNA polymerase holoenzyme -10 consensus promoter well, but a relatively poor -35 promoter. These coregulated genes also contain a conserved 16 base pair consensus PurR binding or operator site (15). A PurR operator was first identified in the control region of the *purF* gene by gel retardation and DNase I footprinting. A 217-base pair restriction fragment encompassing the *purF* promoter-operator region was used in these studies which showed that cell extracts which were *purR*⁺ demonstrated specific binding to the *purF* site as opposed to cell extracts which were *purR*⁻. The DNase I footprinting studies allowed for the identification of a PurR operator site which was shown to be a 16 base pair imperfect dyad repeat. The subsequent identification of PurR operators from other coregulated genes allowed for the determination of the consensus PurR operator site, aCGCAAAC-GTTTtCNT, where upper-case letters indicate highly conserved positions, lower-case letters indicate less conserved positions and N indicates any nucleotide.

For all the genes encoding the *de novo* purine enzymes, with the exception of *purA* and *purB* and the *purR* gene itself, the PurR operator control site is situated between -46 and +10 relative to the transcription start site (15). In the former cases the binding of the Pur repressor is thought to inhibit transcription initiation. The regulation of the *purB*, *purA*

and *purR* genes is more difficult to explain, however repression of elongation may be the mechanism for at least the *purB* and *purR* genes.

In the *purB* gene, which encodes adenylosuccinate lyase, the PurR operator is located 242 base pairs downstream from the transcription start site, overlapping codons 62 to 67 in the protein-coding sequence. It was shown that the transcription of *purB* is repressed approximately two-fold by PurR. This repression is not influenced by the strength of the promoter as evidenced by *trp* promoter replacement studies in which the strength of the promoter was increased seven-fold, yet two-fold *purR* dependent repression was maintained. This indicated that Pur repressor-operator control is independent of the *purB* promoter. Instead, experiments in which truncated *purB* mRNA of 260 nucleotides was isolated in *purR*⁺ but not *purR*⁻ cells, indicated that purR binding to the *purB* operator inhibits transcription elongation by a roadblock mechanism (25). The size of the truncated transcript (260 nucleotides) was that which would be expected from a block in transcription by PurR. This finding suggests that the transcription elongation complex can approach within approximately 10 base pairs of bound PurR.

Another gene in which a PurR operator is found within the protein coding region is the *purR* gene itself. The autoregulation of PurR requires two *purR* operator sites, *purRo*₁ and *purRo*₂. O₁ is located between base pairs 96 and 111 (relative to the transcription start site) which lies between the transcription start site and the site for translation initiation. O₂ is located in the protein coding region from base pairs 184 to 199. It was shown that PurR binds noncooperatively to these sites and no DNA looping was observed in these experiments. Loops are generally formed when binding sites are separated by an integral number of helical turns. The two PurR operators are separated by 88 nucleotides from dyad center to dyad center resulting in a separation of 8.5 helical turns of B-DNA, a number consistent with the observation of no DNA looping. PurR binds to the two operator sites with different affinities. O₁ has a six-fold higher affinity than O₂ for PurR, thus saturation of O₁ by repressor precedes saturation of O₂, yet the position of O₂, within the protein coding region, suggested that it had an *in vivo* role in repression. Therefore, mutational analysis were carried out to dissect the role of each operator in repression. These studies showed that cells carrying operator mutations only in O₁ (O₁⁻O₂) displayed a 1.9 to 2.0 fold repression compared to the wild type 2.5 to 2.7 while cells carrying operator mutations only in O₂ (O₁O₂⁻) still showed a 1.5 to 1.6 fold repression. The double operator mutant (O₁⁻O₂⁻), as expected, showed little or no activity (21). These results attest that both sites are required for the two- to three-fold autoregulation observed *in vivo*. Interestingly, the two to three-fold repression observed in *purR* is similar to that observed

for the branch point enzymes rather than the higher levels of repression observed in the *de novo* IMP enzymes.

It was originally thought that *purA*, which encodes adenylosuccinate synthetase, was regulated separately from the other *pur* genes, however, recently it has been demonstrated that *purA* is also regulated by the Pur repressor (26). Unlike *purB* and *purR* and similar to the other *pur* genes, the regulation of *purA*, appears to be at the level of initiation. An interesting finding in these studies was that *purA*, unlike the other *de novo pur* genes, is regulated by a dual control mechanism, one of which is *purR* dependent and one of which is postulated to be posttranscriptional. Regulation by PurR was demonstrated by a two fold repression of adenylosuccinate synthetase activity in *purR*⁺ cells, not present in *purR*⁻ cells. Two PurR operator sites, O₁ and O₂, were located from -98 to -83 and from +2 to +17 relative to the transcription start site by DNase I footprinting. Operator O₁ displays a higher affinity for PurR however, both *purA* operators, O₁ and O₂, are required for full repression. The only other *pur* regulated gene containing two operator sites is the *purR* gene (21). As in the *purR* gene, there was no looping observed in *purA* operator binding studies, although the separation of binding sites, 99 base pairs, does not preclude loop formation in B-DNA.

Ongoing studies have identified additional genes, connected with nucleotide synthesis, that are part of the PurR regulated *pur* regulon (Fig.3). These coregulated genes include *pyrC* and *pyrD*, which encode enzymes required for *de novo* pyrimidine biosynthesis, *codBA*, which encodes enzymes required for cytosine transport and salvage, *prs*, which encodes PRPP synthetase, *glyA* and the *gcv* operon, which encode enzymes involved in glycine synthesis, catabolism and synthesis of one-carbon units, *speA*, which encodes arginine decarboxylase, a key enzyme involved in polyamine synthesis, and *glnB* which encodes a regulatory protein for the *gln* operon (29-31). The *purR* operator sites are found at varying positions within the promoter sites of these genes, usually upstream of the -35 promoter element, and lead to two to five-fold repression. All of these genes are subject to other forms of regulation besides PurR mediated regulation. The use of several, low-level, regulatory mechanisms for genes encoding enzymes critical for multiple pathways, such as *speA*, allows *E. coli* to adjust efficiently to a variety of environmental conditions.

The first of these *purR* regulated gene, which are distinct from the *de novo* purine *pur* genes, to be identified was *glyA*, the gene encoding serine hydroxymethyltransferase. A *purR* operator was identified 18 base pairs upstream of the -35 promoter. Studies confirmed that PurR bound this operator and when bound, elicited two-fold repression of the *glyA* gene (29). In addition to the negative regulation mediated by PurR, the *glyA* gene

appears to be under positive control by the MetR protein, although the details of this activation are not well characterized.

A rather surprising and unanticipated finding was that PurR also regulates the transcription of genes encoding enzymes required for *de novo* pyrimidine biosynthesis. These genes are *pyrC*, which encodes dihydroorotase, and *pyrD*, which encodes dihydroorotate dehydrogenase. This finding marked PurR as the first *trans*-acting factor identified that is involved in *de novo* pyrimidine biosynthesis in *E. coli* (30). Indeed, the regulation of the genes encoding enzymes required for *de novo* pyrimidine biosynthesis in *E. coli* appears to be much more complex than the regulation of the genes encoding the *de novo* purine enzymes, all of which are all regulated by PurR alone, with the exception of *purA* (26). As for the *de novo* purine genes, the genes encoding the enzymes required for *de novo* pyrimidine biosynthesis are scattered throughout the chromosome in the form of small polycistronic and monocistronic operons and include; *carAB*, *pyrBI*, *pyrC*, *pyrD*, *pyrE* and *pyrF*. Before 1990, it had been established that the expression of *carAB* is repressed by arginine and pyrimidine nucleotides; *pyrBI* and *pyrE* are regulated by a UTP sensitive attenuation mechanism as well as a secondary attenuation independent mechanism; *pyrF* is repressed by a uracil nucleotide and, *pyrC* and *pyrD* are repressed by both a cytidine nucleotide and PurR.

The identification of PurR operator sites within the promoter regions of these genes (from -31 to -16 in *pyrC* and 18 nucleotides upstream of the -35 promoter in *pyrD*) as well as additional *in vivo* studies demonstrating two-fold *purR* dependent repression of their transcription, confirmed that PurR was a regulator of *de novo* pyrimidine biosynthesis in *E. coli*. (30-31). A potential PurR operator site was also located upstream of the *carAB* gene but it has not yet been established whether it is a PurR regulatory site. PurR also regulates the transcription of *codA*, the gene encoding the enzyme cytosine deaminase, a key pyrimidine salvage enzyme (17).

With mounting evidence that PurR functions to regulate many genes, in addition to the genes encoding the *de novo* purine enzymes, a data base search was conducted to identify PurR *cis*-regulatory elements in the promoters of additional *E. coli* genes (32). This study uncovered putative PurR operator sites within the 5' flanking regions of nine additional *E. coli* genes : *glnB*, which encodes the nitrogen regulatory protein, PII; *cytR*, which encodes the cytidine repressor, also a LacI family member; *dksA*, which encodes the Dna K supressor protein; *speA*, which encodes arginine decarboxylase; *ascF*, which encodes part of the phosphotransferase system; *gshII*, which encodes glutathione synthetase; *ppiA*, which encodes peptidyl-prolyl-*cis-trans*-isomerase; *gltS*, which encodes the Na⁺/glutamate symport carrier and *prsA*, which encodes PRPP synthetase. Initially,

glnB, *prsA*, *speA*, *cytR* and *dksA* were examined for PurR regulation. Work is still in progress on the other candidates. These initial studies showed that of the five, PurR bound the *glnB*, *speA* and *prsA* operators *in vitro* and repressed their transcription *in vivo* as judged by *lacZ* reporter studies. The repression for these genes, was again, in the two to three-fold range. Thus, the *pur* regulon encompasses a large assembly of genes, whose numbers are still growing and all of which are involved in nucleotide synthesis.

C. DNA recognition by regulatory proteins and transcription factors

Protein-DNA interactions are involved in many of the fundamental processes that occur in living cells including packaging, replication, recombination and transcription. Understanding these processes in detail necessitates knowledge of the nature of protein-DNA interactions. Unlike the proteins involved in packaging, replication and recombination, regulatory proteins involved in transcription, such as PurR, display extremely high specificity for their target sites and can distinguish a single site of 10-20 base pairs in a background of 10^6 - 10^9 . Understanding the molecular basis for this exquisite specificity is perhaps one of the biggest challenges in understanding protein-DNA interactions and requires a detailed atomic view of these specific interactions. In recent years significant progress has been made in this area as structures have been determined for prokaryotic as well as eukaryotic regulatory proteins both complexed and uncomplexed to their cognate DNA sites. One of the central observations emerging from these studies is that many DNA-binding proteins can be grouped into classes or families based on the structural motif used in recognition (33-35). The first family to be identified, based on structures of several prokaryotic regulatory proteins, was the helix-turn helix (HTH) family (36). As more structural information accumulated, including the structures of eukaryotic transcription factors, other families were identified. To date, the main families include the HTH family, which can be extended to include HTH variants such as the eukaryotic homeodomains, POU specific domains (37) and winged helix family members (38); the zinc binding proteins (39-40); the basic region/leucine zipper, basic region/helix-loop-helix and basic region/helix-loop-helix/leucine zipper proteins (41); and the β -ribbon proteins (42). Additional structures of DNA-binding regulatory proteins which do not conform to any of the above categories have been determined. Among these are the rel homology protein, NF- κ B (43-44), the eukaryotic tumor suppressor protein, p53 (45) and the papillomavirus 1 protein, E2 (46). Indeed, the large number of distinct structural families of regulatory proteins that is emerging demonstrates that there are multiple solutions to the structural problem of designing a specific DNA-binding protein.

The grouping of specific DNA-binding regulatory proteins into families provides a powerful unifying theme in studying these proteins as they relate not only evolution and gene regulation but structure and DNA recognition as well. Moreover, comparative studies on proteins from the same family have shown that aside from adopting the same fold they often show a conserved docking mechanism onto the DNA and a conserved set of DNA contacts. The similar docking mechanisms within a given family indicate that contacts may be predicted, to some extent, if a given protein's family has been identified. In this context it is important to point out that most of the base contacts that have been observed involve

the types of side chain-base interactions that were predicted by Seeman, Rosenberg and Rich (47). Key examples are the interactions of arginine with guanine and glutamine or asparagine with adenine. Seeman, Rosenberg and Rich also predicted that the major groove would play the key role in DNA-binding recognition because of its greater width, which allows greater access to the bases, and the more specific and discriminatory pattern of hydrogen bonds possible from the edges of its base pairs. This hypothesis has so far been substantiated. This is not to say that specific contacts between protein side chains and bases in the minor groove do not occur, however such interactions are much less common. Two broad principles, called 'direct readout' and 'indirect readout', of sequence-specific recognition in protein-DNA complexes have been proposed based on X-ray crystal structures. 'Direct readout' involves the Seeman, Rosenberg and Rich complementary hydrogen bonding interactions between protein groups and DNA bases exposed in the grooves. 'Indirect readout' is more speculative and is proposed to involve the detection, by the protein, of the sequence-dependent conformational variability or flexibility of the DNA itself. Thus, 'indirect readout' implies a role for the conformation of the DNA in the specific binding by a protein (48).

The motifs involved in major groove binding as well as those involved in minor groove binding that have been structurally characterized will be discussed in the following sections. These structures provide a basis for understanding how regulatory proteins specifically recognize and bind their cognate DNA sites. As will be seen, the use of α -helices in site-specific recognition is quite frequent. So frequent, in fact, that these helices are often designated "recognition helices". This designation can be misleading, however, because although these helices ultimately provide the contacts necessary for specificity, they cannot function independently. Indeed, there is no evidence that an isolated helix from any known motif can bind DNA in a sequence specific manner. Such motifs are always part of a larger domain which is important in docking the recognition element(s) and its stabilization. Key in these docking mechanisms are DNA phosphate contacts which are often made by residues throughout the given domain.

1. Major groove recognition

a. The helix-turn-helix motif

The helix-turn-helix (HTH) motif was the first DNA recognition element identified and because the structures of many HTH containing proteins have been determined, it is likely the best characterized as well. It was initially defined as a 20 residue segment with two α -helices that cross at an angle of 120° where the first helix is composed of residues 1-7 and the second helix, the "recognition helix", is composed of residues 12-20.

Interestingly, this motif, unlike other DNA-binding motifs, does not itself constitute a stable structural domain and is found embedded in a variety of structural environments. For example, the DNA-binding domains from the λ phage cro protein (49) and the *E. coli* catabolite activator protein (CAP) (50), contain β -sheets while the DNA-binding domains from λ repressor (51), 434 repressor (52), 434 cro (53) and the *E. coli trp* repressor (48) are entirely α -helical (Fig. 4). Remarkably, despite the difference in domain structure and a lack of significant sequence homology between these proteins, their HTH motifs are nearly structurally identical.

The strict structural conservation of this motif and its demand for certain stereochemical constraints made possible the development of a method to identify a possible HTH motif within the sequence of a given DNA-binding protein. This method was based on the few structures of DNA-binding domains that were known at the time, including λ cro and CAP (54-56). Shown in Figure 5 is a stereo view of the HTH motif from PurR. Examination of this motif reveals important stereochemical restrictions for residues 4, 5, 8, 9, 10 and 15 if this fold is to be maintained (36). For example, residues 4, 8, 10 and 15 are buried in the core of the protein and therefore, must be hydrophobic. Residue 5, because it is wedged between the two α -helices, should not contain a branched amino acid and helical residues 3-7 and 15-20 should not contain prolines. Also, residue 9, which is the second residue in the turn, takes on the conformation of a residue in a left-handed α -helix and therefore, it was thought, must be a glycine. In fact all of the canonical HTH structures that had been determined contained a glycine at position 9. For this reason, the structure determination of PurR was important because its predicted HTH contains an asparagine at position 9.

These "HTH" rules proved to be remarkably successful in predicting the HTH motifs of proteins that follow the strict definition, however as more structures of DNA-binding proteins were determined, including eukaryotic proteins, it became apparent that not only can there be exceptions to these rules but variations of this motif also exist. Thus in addition to the classic or canonical HTH proteins, subfamilies of HTH proteins have been defined and include the eukaryotic homeodomains, POU specific domains, and

winged-helix domain. Structural studies on these proteins complexed to their cognate DNA sites have shown that they all display distinct docking mechanisms that are dictated by the domains in which the HTH element is embedded.

Comparing the canonical prokaryotic HTH protein-DNA structures that have been determined shows that these complexes share a number of common features: 1) The proteins bind as dimers to pseudo-palindromic operator sites with each monomer recognizing a half site. 2) The conserved HTH element contacts the DNA in each half site such that the first helix of the HTH motif is somewhat "above" the major groove and the NH₂-terminus of this helix contacts the DNA backbone. The second helix of the HTH motif, the recognition helix, fits into the major groove such that the NH₂-terminal portion of this helix is closest to the edges of the base pairs. As a result, most of the base specific contacts made by residues in the first turn of the recognition helix. 3) Each complex has an extensive network of phosphate contacts that are important in docking the HTH motif onto the DNA.

The structure of the NH₂-terminal DNA-binding domain of the repressor from the bacterial phage λ (residues 1-92) complexed to a 20 base pair deoxyoligonucleotide containing the O_L1 operator site,

T₁A₂T₃A₄T₅C₆A₇C₈C₉G₁₀C₁₁C₁₂A₁₃G₁₄T₁₅G₁₆G₁₇T₁₈A₁₉T₂₀, exemplifies the docking mechanism and the types of interactions utilized by canonical HTH proteins (52) (Fig 6). Although the λ repressor dimerizes mainly through its COOH-terminal domain (residues 93 -236), the NH₂-terminal DNA-binding domain can dimerize weakly, allowing each subunit to interact with one half of the operator. Each NH₂-terminal subunit is folded into five α -helices connected by loops. Helices 2 and 3 constitute the HTH motif. Most of the contacts important for specificity are made by helix 3, the λ "recognition helix", to DNA bases in the major groove. Hydrogen bonds are made from Gln44, the first residue in helix 3, to A₄. Ser45 makes a single hydrogen bond with G₆' (where the prime indicates the other DNA strand) and Gly46 and Gln48 make hydrophobic contacts with thymine methyl groups. An additional base specific contact is made by a residue not within the recognition helix, Asn55. Asn55 is located in the loop three residues after helix 3 and hydrogen bonds to G₈' (Fig6). Phosphate contacts are numerous and are essential for both docking and complex stability. Residues from regions other than the recognition helix contribute to these contacts. Key positioning phosphate contacts are made by the amide nitrogen from the first residue in helix 2, Gln33, and Asn52, the last residue in helix 3, to phosphodiester oxygens. Interestingly, contacts corresponding to these have been observed in virtually all canonical HTH proteins, underscoring the significant role these contacts play in the specific docking of the HTH element (33). Unexpectedly, in addition

to its docking role, Gln33 functions in specificity as well by hydrogen bonding to the side chain of Gln44. This hydrogen bond helps position Gln44 to make a base specific contact to A₄. Such complexity in the recognition process complicates the reliability of predictive methods in determining specific protein-DNA interactions, even within a given family.

Additional DNA contacts in the λ -O_L1 complex are provided by an NH₂-terminal arm that becomes ordered only upon DNA binding. This region, which makes contacts with the major groove, can be considered an independent DNA binding module. Lys4 from this arm plays an important role in the recognition process by cooperating with Asn55 to make hydrogen bonds with G₈, while the rest of the arm wraps around the center of the operator making several additional contacts. These interactions and the numerous phosphate contacts underscore the importance of residues outside the recognition helix in the recognition process.

Any description of HTH proteins would be incomplete without mention of the *E. coli* catabolite activator protein (CAP), also called the cAMP (adenosine 3',5'-monophosphate) receptor protein (CRP). CAP functions to activate transcription in the presence of cAMP from more than 20 different *E. coli* promoters by binding to operators located from -41 to -103 relative to the transcription start site. Structural studies on CAP have contributed significantly to the understanding of protein-DNA interactions. Not only was CAP one of the first HTH protein whose structure was determined, but the more recent structure determination of the CAP-cAMP-DNA complex provided the first example of a protein which induces a severe bend in its DNA site via DNA kinking (50) (Fig. 4B). DNA kinking, in which bases are unstacked to provide sharp localized bends, can be distinguished from smooth bends where the base pairs are only partially unstacked so as to spread the bending energy over several base pairs. The unstacking of base pairs seen in DNA kinking is a direct result of unusually large roll angles between base pairs (50). The 90° bending observed in the crystal structure of the CAP-cAMP-DNA complex confirmed a large body of biochemical evidence which indicated that CAP bent its DNA site between 90°-140°.

CAP, like most prokaryotic HTH proteins, is a dimer of identical subunits. Each CAP subunit consists of 209 amino acids and contains an NH₂-terminal cAMP binding domain, consisting of a β roll structure and a very long α -helix that mediates most of the dimer contacts and a COOH-terminal DNA-binding domain, containing a H1-S1-S2-H2-T-H3-S3-S4 motif, in which the H stands for α -helix, S for β -strand, T for the turn of the HTH element and H2 and H3 are the α -helices of the HTH. CAP contacts the DNA in a manner similar to other HTH proteins, by inserting its recognition helix, H3, into the major groove. Three side chains emanating from the recognition helix make direct contacts

to base pairs in the major groove. Although such contacts would allow CAP to interact with 20 base pairs of a straight DNA binding site, bending the DNA allows the large regions of positive electrostatic potential on the sides of CAP to contact the negative phosphate backbone, and thereby permits CAP to interact with a 28 base pair region of DNA. This bending results in greater part from two 40° kinks between TpG base pairs on each side of the DNA dyad axis that are formed by roll angles of approximately 40°. In addition to these kinks, one side of the complex is bent 8° towards the protein. This added bend allows Lys26, which is located on the cAMP binding domain, to interact with a DNA phosphate group. It is proposed that the analogous contact with Lys26', on the other side of CAP, also occurs *in vivo* but is not seen in the structure due to crystal packing influences, a suggestion which is consistent with biochemical studies.

The eukaryotic homeodomain was the first noncanonical HTH observed. This domain is a DNA-binding motif consisting of 60 residues that is present in a large family of eukaryotic regulatory proteins (33, 35) and was first identified in proteins that regulate *Drosophila* development such as Antennapedia. Now it is recognized that this motif has a broader role in eukaryotic gene regulation (57). The structure of the *Drosophila* Antennapedia homeodomain, determined by NMR, provided the first structure of a homeodomain motif and was also important because it provided the first example of a HTH protein that was an exception to the "HTH" rules (58-59). Specifically, it broke the "position 9 always glycine" rule by instead having a cysteine. Despite this substitution, the HTH fold is maintained.

Comparisons of the structures of the homeodomains which have been determined thus far, including the *Drosophila* homeodomains from Antennapedia (58-59) and engrailed (60), the yeast homeodomain of MAT α 2 (61), the Oct-1 homeodomain (62) (see Fig. 4C for Oct-1 homeodomain within the POU domain) and the homeodomain from the rat liver transcription factor, LBF1/HNF1 (63-64), have revealed that despite the lack of sequence homology, these homeodomains have remarkably similar structures. Unlike the canonical prokaryotic HTH motif, which are found embedded in domains of varied structure, the homeodomains, as their names suggest, form a stably folded independent domain. In addition, isolated "monomeric" homeodomains can bind DNA with a specificity similar to the intact protein.

The homeodomain is composed of three α -helices and an extended NH₂-terminal arm. Helix 1 and helix 2 pack against each other in an antiparallel arrangement and helix 3, the recognition helix, is perpendicular to helices 1 and 2. Unlike the canonical prokaryotic HTH proteins, helices 1 and 2 are too far from the DNA to make contacts and most of the specificity determining contacts are made by helix 3, which contacts the major groove, and

an NH₂-terminal arm which contacts the minor groove. The arrangement of the helices in this fold is highly conserved and insertions into the 60 residue homeodomain fold are accommodated by variable-length loops between these helices without disrupting this arrangement. For example, MAT α 2 contains a three residue insertion between helices 1 and 2, which is accommodated in the loop between these helices (61). The atypical homeodomain, LBF1/HNF1 provides the most extreme example of such a case as it contains a 21 residue insertion within the HTH element of the homeodomain (helices 2 and 3) (63-64). Surprisingly, the overall structure of this homeodomain is nearly identical to the other homeodomains with the insertion being accommodated by a lengthening of helix 2 by eight residues and an extension of the loop between the HTH element by 13 residues.

Almost as conserved as the three-dimensional fold of the homeodomain is the way in which proteins containing this motif dock onto DNA. Indeed, the homeodomains appear to have a far more conserved mechanism of docking than their prokaryotic counterparts which, because they are embedded in a variety of domains, all make slightly different phosphate docking contacts. This finding makes the homeodomain, perhaps, the most amenable HTH DNA-binding domain to predictive methods. Such predictive efforts are greatly aided by the presence of 4 "invariant" residues, Trp48, Phe49, Asn51 and Arg53 (WFXNXR), which are found in all homeodomains (33). As has been shown, Trp48 and Arg53 make key phosphate backbone contacts while both Trp48 and Phe49 are critical in the formation of a stable hydrophobic core. However, the key residue is Asn51. This residue makes a conserved bidentate contact to adenine3 in the consensus homeodomain operator site T₁A₂A₃T₄, in all known homeodomain-DNA complexes. This common interaction along with phosphate contacts made by six highly conserved residues, Tyr25, Gln44, Trp48, Arg53, Lys55 and Lys57, fix the position of helix 3 in the major groove (61). Beginning with these conserved interactions, one can attempt to model base contacts made by other side chains in helix 3. It has been shown that DNA specificity is imparted by residues 47, 50 and 54 (33). Interestingly, these residues, as well as Asn51, are located in the middle of the recognition helix. This is different from the canonical prokaryotic HTH proteins, in which residues near the NH₂-terminus of the recognition helix are used for contacting bases, and reflects the significantly different docking mechanisms used by these two HTH subfamilies.

Although isolated homeodomains can bind DNA with a specificity similar to that of the intact protein, it has been demonstrated that the DNA binding affinities and specificities of isolated homeodomains can be altered by several factors, including additional covalently attached DNA-binding domains, different oligomerization states and/or interactions with other proteins (57). DNA-binding by Mat α 2, for example, is modulated not only by

dimerization but also by interactions with at least two other proteins, MCM2 and a1 (61). The DNA-binding activities of homeodomain proteins with a POU domain, which was named for the first POU proteins isolated, Pit-1, Oct-2 and unc86 (37), can be altered by protein-protein interactions such as seen for the Oct-1-VP16 complex (62). However, in addition to the POU homeodomain, DNA-binding by POU domains ultimately relies on a covalently attached POU specific domain and thus provide an example of a homeodomain containing protein whose DNA binding activity is modulated by an additional covalently attached DNA-binding domain (37). In these proteins the POU specific domain is located NH₂-terminal to the POU homeodomain. The NMR structure of the POU specific domain of Oct-1 revealed that this domain also contains an HTH motif (65). In the POU specific domain the HTH motif is embedded in a four helix domain which bears a striking resemblance to the DNA-binding domains of the λ and 434 repressors. Furthermore, several residues in the HTH element are conserved between these quite divergent proteins.

The X-ray structure determination of the entire Oct-1 POU domain complexed to the octamer half site, A₁T₂G₃C₄A₅A₆A₇T₈, provided the first opportunity to view a homeodomain in the context of a larger protein (62) (Fig. 4C). In this structure the POU specific domain contacts the 5' half of the octamer site, **A₁T₂G₃C₄A₅A₆A₇T₈**, (indicated in bold). As predicted from the NMR structure, the docking and contacts made by this domain to the DNA are remarkably similar to those made by the λ and 434 repressors. The Oct-1 homeodomain contacts the 3' half of the octamer half-site, A₁T₂G₃C₄A₅A₆A₇T₈, and even in the context of a larger protein, docks against the DNA and makes the same DNA contacts, including the conserved asparagine to adenine contact, that are observed for all other homeodomains. A much anticipated feature of this structure is that it reveals, for the first time, the NH₂-terminal arm of a homeodomain in the context of a larger protein. Although this "arm" is attached to a flexible linker of 24 residues that connects the POU specific domain to the POU homeodomain, it makes minor groove contacts similar to those previously observed in all other isolated homeodomain-DNA structures (58-61).

Interestingly, there are no protein-protein contacts between the POU specific domain and the homeodomain and the two domains bind major grooves on opposite faces of the DNA (Fig. 4C). This type of arrangement leaves the surface of the POU-specific and the POU homeodomains fully accessible for interactions with other proteins, such as VP16. Despite the lack of POU specific-POU homeodomain interactions, residues from these two domains do make overlapping phosphate contacts near the center of the octamer site which would, perhaps, favor cooperative binding of Oct-1 to the octamer site.

Whereas homeodomain proteins that are monomeric utilize NH₂-terminal arms to promote binding strength and specificity towards their target DNA sites, a recently identified structural family of eukaryotic HTH variants containing a so-called "winged-helix" motif, similarly employs loops (the "wings") to assist in the binding of their monomeric HTH units to target DNA sites. The winged helix motif was first identified in the hepatocyte nuclear factors (HNF-3 α , HNF-3 β and HNF-3 γ) and the *Drosophila* protein, forkhead, and consists of a conserved region of approximately 110 amino acids which is both necessary and sufficient for specific DNA binding (66). Since this motif was first identified, several additional proteins containing this motif have been identified in both vertebrate and *drosophila*. In situ hybridization and antibody studies in various adult and embryo tissues have suggested that these proteins function during development. Moreover, very specific early development roles were found for HNF-3 α and HNF-3 β and the winged helix protein, BF-1 in mammals (38,66).

The crystal structure of the DNA-binding domain of HNF-3 γ (residues 107-223) bound to the 13 base pair transthyretin (TTR) promoter target site, G₁A₂C₃T₄A₅A₆G₇T₈C₉A₁₀A₁₁C₁₂C₁₃, was the first to reveal this new motif and inspired the designation, "winged helix" (67). The protein contains an α/β structure consisting of three helices, H1 (residues 124-134), H2 (residues 142-151) and H3 (residues 160-173) and a three-stranded antiparallel β sheet composed of strands S1 (residues 139-141), S2 (residues 177-183) and S3 (residues 192-194). Two loops, W1 (184-191) and W2 (195-217) are located between S2 and S3 and at the COOH-terminus, respectively. These loops constitute the "wings" of the winged-helix motif and helices, H2 and H3, constitute the HTH motif.

The DNA-binding domain of HNF-3 γ makes extensive contacts with the TTR promoter, mainly by residues from the recognition helix, H3 and the wings, W1 and W2. Interestingly, all side chain-base contacts are restricted to the sense strand and include one minor groove contact, between Arg210 and T₄, and major groove contacts between His169 and T₈ and Asn165 and A₁₀. An unusual water mediated contact is also found between the side chains of Asn214 and Asn174 and G₇. All phosphate contacts are made by residues from H3, W1 and W2 with three exceptions: Leu142 from H2 makes van der Waal interactions with the ribose of G₁₃, the side chain of Tyr193 hydrogen bonds to the phosphate groups surrounding T₈, and the side chain of Trp193 hydrogen bonds to the phosphate of G₁₂. The manner in which HNF-3 γ (107-223) docks onto the TTR promoter site somewhat resembles a butterfly perched on a rod in which H3 is the thorax and W1 and W2 are the wings, hence the name "winged helix".

Comparison of the structure of the winged helix of HNF-3 γ with structures of the globular domain of histone H5 (GH5) (69) and the more recent structures of the DNA binding domain of the human *ets* protein, Fli-1 (70) and the DNA activator binding domain of the bacteriophage Mu transposase, MuA, (71) has revealed that these proteins share striking structural homology despite a lack of sequence homology. This structural homology has prompted the inclusion of all proteins with this similar fold into a structural "winged helix" family. The DNA-binding domains of these proteins are roughly 90-110 residues and are characterized by the presence of three helices, two of which constitute the HTH motif. Members of the structural winged helix family can be divided further into two subfamilies depending on whether there is a single β -strand or a β -hairpin between the first helix and the HTH motif. The first structural subfamily, which includes CAP (50,56), the heat shock transcription factor (72) and Fli-1(70), has a H1-S1-S2-H2-T-H3-S3-W1-S4-W2 fold, where H indicates α -helix, S indicates β -strand and W indicates a loop or wing. The second subfamily includes GH5 (69), the biotin repressor or BirA (72) and the HNF-3/forkhead family (68), and displays a H1-S1-H2-T-H3-S2-W1-S3-W2 topology (Fig. 7). In both subfamilies the H2, T and H3 elements constitute the HTH motif. The angle between the two helices of the HTH varies in these proteins from 90° to 120°, however, the positioning of H1 with respect to H2 is conserved throughout both subfamilies and an angle of 45° is formed between H1 and H2. The most divergent member of the structural winged helix family is MuA(1-76) (71). However, the structural topology of MuA is most consistent with the second subfamily except that H1 has been supplanted from the NH₂-terminus to the COOH terminus of the domain to give the overall topology, S1-H1-T-H2-S2-W1-S3-H3 (Fig. 7).

To date the structures of only three other members of the structural winged helix family have been determined in the presence of their cognate DNA sites. These members include the prokaryotic protein, CAP, MuA(1-76) and the *ets* protein, Fli-1. Like most prokaryotic HTH proteins, CAP binds its DNA site as a dimer. Fli-1, like HNF-3 γ and other members of the "true" winged helix family, binds its DNA site as a monomer. However, MuA exists as a functional tetramer when carrying out its function of transposition on DNA. The structures of these proteins complexed to DNA serve to underscore a recurring difference that has been observed between protein-DNA complexes of prokaryotic canonical HTH and eukaryotic HTH proteins. This is the use of residues from different positions in the recognition helix for making DNA contacts. Again, eukaryotic HTH proteins employ residues located in the middle of their recognition helices for making DNA contacts, while prokaryotic HTH proteins use residues located in the NH₂-terminus of their recognition helices.

An unusual variation on the HTH theme in which two HTH elements are covalently attached and bind to DNA in tandem was revealed by the NMR structure of the minimum DNA-binding domain (called R2R3) of mouse *c-Myb* protein bound to a DNA site containing the consensus Myb sequence, AACTG (73). *c-myb* is a proto-oncogene that controls the proliferation and differentiation of hematopoietic cells. Deregulation of *c-Myb* DNA binding leads to oncogenic activation. *c-Myb* has three functional domains which are responsible for DNA binding, transcriptional activation and negative regulation. Interestingly, the DNA-binding domain consists of three imperfect tandem repeats of 51 or 52 amino acids. However, only the second and third repeat, R2 and R3, are required for DNA binding. The solution structure showed that R2 and R3 each contain three helices with H2, H3 and an extended turn forming an HTH element. However, the manner in which the recognition helix of the HTH of *c-Myb* is docked against the DNA is distinct from all other HTH and variant HTH proteins. Unlike typical HTH proteins in which the recognition helix lies parallel to the edges of the major groove bases or in the case of the homeodomains, slightly tilted, the recognition helices of R2 and R3 are dramatically tilted toward the helical axis of the DNA in manner more akin to the classic, TFIIIA-like zinc fingers (74-75), which are also covalently connected in tandem.

As is evident from the HTH structures just described, the original definition of the HTH element, with four amino acids in the turn, is too restrictive. Antennepedia provided the first example of a divergent HTH with the substitution of a cysteine for the "invariant" glycine 9, suggesting that other residues may be tolerated at this key position within the turn. This was followed by more extreme divergences such as *c-Myb*, GH5, HNF-3 γ and the POU specific domain in which one, two, four and six additional residues are accommodated within the turn. Perhaps the most divergent HTH member to date is the HNF1 homeodomain in which there is a 21 residue insertion with the HTH element. Despite these deviations from the original "rules", the HTH motifs of Antp, *c-Myb*, GH5, HNF-3 γ , HNF1 and the other divergent members all take on the characteristic HTH fold. The large number and diverse members of proteins, from both prokaryotes and eukaryotes, containing the HTH fold attest to its evolutionary success and indicate that it is one of nature's most successful designs for DNA-binding proteins.

b. Zinc binding proteins

A large group of DNA-binding proteins contain zinc as a structural element. To date, four major classes of such proteins, which are found primarily in eukaryotes, have been characterized structurally. The first class of proteins, which will be called the class I zinc binding proteins, contain the original "zinc finger" motif, first identified in the *Xenopus* transcription factor, TFIIIA (75-76). This motif is a 30 residue domain in which one zinc ion is liganded by two cysteines and two histidines following the pattern, Cys-X₂₋₄-Cys-X₁₂-His-X₃₋₅-His. The second class or class II zinc binding proteins, is an approximately 70 residue domain found in the steroid and related hormone-like receptors (77). This domain contains two zinc ions each liganded by four cysteines. The class III zinc binding proteins includes a set of yeast activators including GAL4 (78). The zinc domains of these proteins contain a binuclear cluster which consists of two closely bound zinc ions sharing six cysteines. The class IV zinc binding proteins include the retroviral nucleocapsid proteins (79). These proteins bind a single zinc ion by three cysteines and one histidine found within the sequence, Cys-X₂-Cys-X₄-His-X₄-Cys. In addition to these well characterized structural zinc domain classes, six other zinc binding domains, implicated to be important in DNA binding, have been identified. These include the GATA family of cell specific regulatory proteins (80), the eukaryotic transcription elongation factor, TFIIS (81), MetRS (82), LIM (83-84), RING finger (85) and PKC CRD domains (86).

Three-dimensional structures have been determined for protein-DNA complexes from at least one member of each of the four primary classes of zinc binding proteins (Fig. 8). The class I zinc proteins are perhaps the most prolific class of zinc binding proteins with over 3000 putative members, including, proteins involved in gene development, those induced by differentiation and growth signals, proto-oncogenes and general transcription factors. These proteins generally bind with high affinity to operator sites which are very GC rich. This class of zinc binding proteins is also, structurally, the best characterized, as the small size of their domains (20 residues) makes them highly amenable for study by NMR spectroscopy. Structures of these domains have revealed that they all contain an antiparallel β -sheet and an α -helix (Fig. 8A). Thus, this domain has been called the $\beta\beta\alpha$ fold. Two cysteines near the turn in the β -sheet region and two histidines, in the α -helix, coordinate a central zinc ion forming an independent and compact globular domain (39-40). Most proteins containing these modules have three or more fingers in direct succession.

The direct concatenation of the class I zinc fingers suggested that they form a repeating structure when bound to DNA. This was confirmed by the X-ray structure

determination of the three zinc fingers of the mouse immediate early protein, Zif268 bound to the GC rich Zif268 consensus site, A₁G₂C₃G₄T₅G₆G₇G₈C₉G₁₀T₁₁(74). This structure shows that the zinc fingers bind in the major groove (Figs. 8A and 9). Each finger docks against the DNA in a similar manner and makes base contacts with a three base pair subsite. The bases contacted are on only one strand of the DNA and the first finger contacts the 3' subsite (G₈C₉G₁₀), the second finger contacts the middle subsite (T₅G₆G₇) and the third finger contacts the first or 5' subsite (G₂C₃G₄) (Fig. 8A and 9). The antiparallel β-sheet of each finger is at the back of the α-helix away from the base pairs while the α-helix contacts the major groove. The zinc fingers use an arginine, (indicated by asterisk in Fig. 9) found in all three fingers, that immediately precedes the α-helix as well as the second, third and sixth residues of the α-helix, to contact the base pairs. None of the individual Zif fingers contact all three base pairs in their subsites. In all three fingers, the arginine preceding the α-helix contacts the third base on each subsite (5' _ _ G^{3'}), the third residue of the α-helix of finger 2, which is a histidine, contacts the second base in its subsite (5' _ G^{3'}) and the sixth residue of the α-helices of fingers 1 and 3 contact the first base in their subsites (5'G^{3'} _ _ 3') with an arginine. Phosphate contacts, common to all three fingers, are provided by an arginine in the second β-strand and the first zinc binding histidine. These relatively simple patterns reflect the finding that each of the three fingers docks against the DNA in a very similar manner with each finger related to the next by a simple helical motion. In this motion, neighboring fingers are arranged such that a rotation of approximately 96 degrees (3 X 32 degrees) around the DNA axis and a translation of approximately 10 Å (3 X 3.4 Å / base pair) along the DNA superimposes one finger on the next.

The modular design of zinc fingers and their similar docking mechanisms provided hope that a simple code could be described for the class I zinc finger DNA-binding domains. Indeed, zinc finger polypeptide design and mutagenesis experiments carried out on fingers closely related to Zif provided evidence that the three residues implicated in the Zif268-DNA structure were, indeed, the key specificity determining elements (87-88). However, the hope for this simple code was eventually dashed when structures were determined for two divergent class I members, Tramtrack (89) and GLI (90), bound to DNA.

The *drosophila* protein, Tramtrack , functions as a transcriptional regulator of the fushi-tarazu gene. The structure of the two-fingered DNA-binding domain of Tramtrack bound to the Tramtrack DNA site, C₁T₂A₃A₄T₅A₆A₇G₈G₉A₁₀T₁₁A₁₂A₁₃C₁₄G₁₅T₁₆C₁₇C₁₈G₁₉, provided the first example of a zinc finger bound to an AT rich site (89). The two Tramtrack fingers dock against the DNA just like the Zif fingers. Also, residues at key positions minus one, two, three and

six in the α -helices play the dominant role in base recognition. Despite these similarities, there are notable and unanticipated differences between the Tramtrack and Zif complexes. Although each Tramtrack finger contacts a three base pair subsite as in Zif, the contacts overlap as both the first and second fingers make contacts to base pair 9; Finger 1 contacts G₉, A₁₀ and T₁₁ and finger 2 contacts C₉, G₈ and A₇. The contact made to C₉ by finger 2 also represents a departure from Zif which only makes contacts to a single DNA strand. Additionally there are structural difference between Zif and Tramtrack where finger 1 of Tramtrack contains an extra β -strand. The structure of the Tramtrack-DNA complex indicates that this added β -strand is critical for structural stability and not DNA binding. Unlike the Zif268-DNA complex, the DNA in the Tramtrack structure reveals a large protein-induced DNA deformation. This deformation is manifested in a 20° bend in the A₁₀T₁₁A₁₂ sequence. This bend is towards the protein in the binding site of finger 1. As a result of this deformation, T₁₁ is displaced towards the protein by 2.5 Å compared with the equivalent base in the finger two binding site. This displacement allows Ser124, the second residue in finger 1, to interact with T₁₁, a contact which would not be possible with normal B-DNA.

Another class I zinc protein, GLI, is an oncogene found amplified in glioblastomas and other tumors. The structure of GLI bound to the high affinity DNA site, A₁C₂G₃T₄G₅G₆A₇C₈C₉A₁₀C₁₁C₁₂C₁₃A₁₄A₁₅G₁₆A₁₇C₁₈G₁₉A₂₀A₂₁, confirmed the presence of five fingers in the GLI DNA-binding domain. Unexpectedly, however, this structure showed that despite the fact that all five fingers are of the Zif268 type, only fingers 4 and 5 contact the DNA specifically (90). Finger 1 does not contact the DNA at all but rather interacts with and stabilizes finger 2. Fingers 2 and 3 interact nonspecifically with the DNA and finger 2 makes the only direct base contact of these two fingers. Unlike the Zif268 fingers which contact contiguous three base pair subsites, fingers 4 and 5 of GLI contact noncontiguous subsites. Furthermore, finger 4 contacts the five base pair region (C₁₁C₁₂C₁₃A₁₄A₁₅) and finger 5 contacts the four base pair region (G₆A₇C₈C₉). Like the Zif268 fingers, fingers 4 and 5 of GLI make most of their base specific contacts with residues from the NH₂-terminal portions of their α -helices. Finger 4 uses the residue preceding the α -helix and the first, second third and sixth residues of the α -helix for contacting bases while finger 5 also uses the residue preceding the α -helix and the second, third and sixth α -helical residues. Finger 5 makes an additional base specific contact with the fifth residue of its α -helix, a contact that has no analogy in any other zinc finger-DNA complex. The large differences between the Zif268-DNA and GLI-DNA complexes may indicate that subfamilies within the class I zinc proteins may use quite different binding mechanisms. GLI is a more distant relative of Zif and belongs to a subfamily which

includes GLI3, tra-1 and Cid. Alternatively, it may be that the DNA site used in the crystallographic studies does not represent a physiological GLI site. Indeed, the GLI site used was patterned after a consensus site which was obtained by *in vitro* selection from genomic DNA. Another factor that may influence DNA binding by GLI is its interactions with other proteins. Evidence for the importance of one such interaction was provided by studies which demonstrated that GLI and E1A act synergistically in the transformation of rodent cells (91).

The aforementioned structures of class I zinc finger-DNA complexes imply that DNA interactions made by fingers belonging to homologous class I subfamilies might be reliably modelled. However, as underscored by the Tramtrack-DNA structure, the DNA need not be a passive partner. Therefore, this structure may provide an example where the inherent ability of a given DNA sequence to adopt a certain conformation contributes to the specificity of the complex.

The second class of zinc binding proteins consists of the steroid receptors. These proteins are an important family of regulatory proteins that include receptors for steroid hormones, retinoids, vitamin D and thyroid hormones. Genetic and biochemical studies have shown that these proteins contain a number of separate domains; an NH₂-terminal activating domain and a COOH-terminal hormone-binding/dimerization domain. Found between these two domains is the DNA-binding domain which is approximately 70-80 residues and the most highly conserved domain in these proteins. This region also possesses a weak dimerization property as well as a nuclear localization signal.

While some members of the zinc class II family function as monomers, most bind as dimers to their DNA hormone response elements (HRE) with each monomer identifying a six base pair half site. Of those receptors that bind as dimers, some are able to form both homo- and heterodimers, including the thyroid, retinoic acid and vitamin D receptors. These proteins recognize half sites with a variety of spacings and orientations. Other receptors, such as the estrogen receptor (ER) and the glucocorticoid receptor (GR), bind solely as homodimers to palindromes in which each half site is separated by three intervening base pairs (92).

Structures have been determined for the DNA-binding domains of several hormone receptors including the GR (93), ER (94), retinoid X receptor (95) and retinoic acid receptor (96). The structures showed that all these domains fold into a similar, independent globular domain containing two zinc binding motifs in which each zinc is tetrahedrally coordinated by four conserved cysteines (Fig. 8B). Although each of the zinc binding motifs in these receptors is somewhat reminiscent of an individual finger from the class I zinc binding proteins, they do not form stable independently folded domains. Instead, they

are interwoven into a single globular domain with extensive interactions existing between each individual zinc binding motif.

The GR and ER receptors recognize one of two consensus half sites, either a GRE (glucocorticoid response element) or an ERE (estrogen response element). The combined structures of the GR DNA-binding domain bound to a GRE (glucocorticoid response element) (93) and the ER DNA-binding domain bound to a ERE (estrogen response element) (94) provide insight into how these different subfamilies of hormone receptors differentiate between the two consensus half sites. The structure of the DNA-binding domain of the GR bound to an unnatural GRE site, which was separated by four base pairs (GRE_{4s}) rather than three, revealed for the first time that the DNA-binding domains of a hormone receptor can dimerize upon DNA binding (93). This structure also demonstrated that the dimerization region becomes structurally ordered upon DNA binding as the solution structure of the GR DNA-binding domain without DNA was monomeric and displayed disordered structure for the dimerization region (94). The reason for using the GRE_{4s} was that it provided a target site with exact symmetry, an experimental expedient that seemed justified based on the initial solution structure which demonstrated that the GR DNA-binding domain is monomeric when uncomplexed with DNA. The results of the GRE_{4s} study prompted a study of the GRE_{3s} specific complex. Interestingly, the GR-GRE_{4s} complex affords a view of a complex which is half specific and half non-specific which further implies that the stability of the GR subunit interface must exceed the increment in stability that would be provided by specific base interactions of a half site, otherwise the subunits would be separated into monomers that bind two GRE half sites.

Comparison of the structures of the ER-ERE (94) and the specific half site of the GR-GRE_{4s} complex (93) provides the first important step in our understanding of how these two subfamilies of receptors discriminate between the ER half site, (A₁G₂G₃T₄C₅A₆), and the GR half site, (A₁G₂A₃A₄C₅A₆), which differ at only two base pairs. Such a detailed comparison demands that the specific half site of the GR-GRE_{4s} be used rather than the GR-GRE_{3s} complex because of the limited resolution of the latter (4 Å). Therefore, this comparison relies on the assumption that the contacts observed at the specific half site of the GR-GRE_{4s} complex are the same as would be observed in a native GRE with correct spacing.

Specific contacts from both proteins to the central four base pairs of the six base pair half site are provided by residues on the surface of the first helix. In comparing the ER-ERE and GR-GRE complexes, the logical assumption is that residues, which make DNA contacts and which are conserved between the two proteins, would contact the DNA identically. Consistent with this idea were mutagenesis studies demonstrating that

discrimination is achieved by just three residues (numbered differently in two structures). In the ER these residues are Glu25, Gly26 and Ala29 (94) while in the GR, the corresponding residues are Gly458, Ser459 and Val462 (93). Furthermore, swapping the corresponding amino acids in ER to those found in GR switches the specificity of the ER to a GRE binding protein. Indeed, the GR-GRE structure shows that the valine, Val462, makes hydrophobic contacts to T₄, which is one of the two specificity determining base pairs. The equivalent ER residue, Ala29, makes no DNA contact. In the ER-ERE complex, the base at position 4' is an adenine and is contacted by the divergent glutamic acid, Glu25, through a water mediated interaction. Glu25 further promotes specificity by directly hydrogen bonding with C₃'. In the GR-GRE complex, the equivalent residue is Gly458 and it makes not DNA contacts. Surprisingly, in the GR complex no contacts are made to the second specificity determining base pair, 3.

The DNA interactions made by side chains conserved between GR and ER presents another surprise as the predicted preservation of identical contacts is not found. For example, the base specific contact made by Lys28 to G₂ is similar in both structures but in the ER complex, the interaction is subtly changed by a buttressing salt bridge between Lys28 and Glu25. The interactions made by the conserved residue, Arg33, differ even more between the two structures. In the GR complex, this residue makes the expected bidentate interaction with G₅, but in the ER complex, Arg33 takes on a different conformation such that it makes only one hydrogen bond to G₅, but additionally makes a direct hydrogen bond to a phosphate group, two water-mediated phosphate contacts and two water-mediated contacts to G₅' and T₄. Interestingly, Lys32 is also conserved in ER and GR but it appears to play no role in DNA binding in GR while in ER its side chain interacts with central base pairs which are specific to the ERE. These surprising observations serve to underscore yet again the complexity of specific protein-DNA contacts that can be influenced by interactions between two or more protein side chains and water molecules.

The third class of zinc binding proteins is represented by a group of at least 11 fungal DNA-binding proteins containing the repeated sequence, Cys-X₂-Cys-X₆-Cys and includes the proteins LAC9, PPR1, QA-1F, QUTA1, ARGR11, HAP1, MAL63, PDRa, LEU3, PUT3, AMDR and GAL4 (78,97). GAL4, which activates the transcription of genes required for catabolism of galactose and melibiose, is the best characterized of these proteins. GAL4 recognizes DNA sequences that are 17 base pairs in length containing conserved CCG triplets and separated by 11 base pairs. Functions have been ascribed for various parts of the 881 amino acid GAL4 protein. Residues 1-65 function in DNA binding and residues 66-94 function in dimerization. In addition, three acidic activating

regions, consisting of residues 95-106, 148-196 and 768-881 have been identified. Residues 851-881 also contain the binding site for the inhibitor, GAL80.

The structure of the 65 residue DNA-binding domain of GAL4, GAL(1-65), bound to the 19 base pair deoxyoligonucleotide, $C_1C_2G_3G_4A_5G_6G_7A_8C_9A_{10}G_{11}T_{12}C_{13}C_{14}T_{15}C_{16}C_{17}G_{18}G_{19}$ (where dyad related triplets separated by 11 base pairs are indicated in bold) revealed that, similar to the ER and GR DNA-binding domains, GAL4(1-65) dimerizes upon DNA binding (97) (Fig. 8C). The NH_2 -terminal metal binding domains mediate sequence specific contacts with bases in the major groove while COOH-terminal residues 50-65 form a weak coiled-coiled α -helical dimerization element. The metal binding modules consist of two α -helices in which two zinc ions are tetrahedrally coordinated by six cysteines. Two of the cysteines ligate both metals, creating an unusual binuclear cluster. The two metal binding sites of the dimer lie in the major groove separated by one-and-one-half turns of DNA and are centered over the conserved CCG triplets. Phosphate contacts by Gln9, Arg15, Lys20, Cys21 and Lys23 anchor the recognition modules to the DNA so that the COOH-terminal end of the first α -helix points into the major groove. Interestingly, the side chain of only one residue from each half site of GAL4 (1-65), Lys18, is positioned to make direct base contacts. In this interaction, Lys18 donates hydrogen bonds to two of the conserved bases of the triplets. Additional base specific contacts are provided by the carbonyls of Lys18 and Lys17.

The ability of GAL4 to recognize its binding site also depends on the presence of an 11 base pair spacer between the CCG triplets. This spacer is specified by the length of the linker region, or dimerization region, between the GAL4 DNA-binding modules. The extended conformation of the GAL4(1-65) dimer and the lack of significant contacts outside the metal-binding domain, leave nearly a full turn of DNA exposed with its major groove on the opposite face of the protein. Despite the fact that substitutions in the DNA linker region have little effect on GAL4 binding *in vitro*, they are well conserved in all *in vivo* GAL4 DNA sites, raising the possibility that this portion of the DNA could serve as a binding site for another protein, a potential candidate being GAL11 (97). Cooperative interactions with such an accessory proteins could alter the DNA binding strength and specificity of GAL4.

The fourth class of zinc binding proteins include retroviral fingers, such as the human immunodeficiency virus I (HIV-1) nucleocapsid protein, with the sequence, Cys- X_2 -Cys- X_4 -His- X_4 -Cys (CCHC) (98). Structures of these proteins show that the first two zinc coordinating cysteines are located on an antiparallel hairpin, similar to the class I zinc fingers. The CCHC finger is, however, much smaller than in class I proteins, with four

residues separating the second (Cys) and third (His) zinc ligands. In these proteins the histidine is found on a loop which continues to the fourth zinc ligand.

The NMR structure of a synthetic peptide corresponding to the NH₂-terminal CCHC finger of the HIV-1 nucleocapsid protein, Zn(HIV1-F1), complexed with single stranded deoxyoligonucleotides as RNA mimics revealed a single requirement for guanine (99) (Fig. 8D). Therefore, further structural studies were carried out using the deoxyoligonucleotide, ACGCC which corresponds to a region of the HIV-1 psi-site. This study showed that the finger did not change conformation upon DNA binding and that the DNA was A-form. The resulting HIV1-F1-DNA model revealed that several highly conserved hydrophobic and aromatic residues in the NH₂-terminal portion of the peptide contact the DNA. In the structure, a backbone carbonyl oxygen and two backbone amide nitrogen groups are in position to hydrogen bond to the conserved guanine (Fig. 8D).

In addition to the four established classes of zinc binding proteins, the structures of several new zinc binding motifs, which have been implicated to be involved in DNA binding, have recently been characterized. A potential class V DNA-binding zinc motif is comprised of the so-called GATA family of cell specific regulatory proteins. These proteins are characterized by the presence of two metal binding regions of the form, Cys-X₂-Cys-X₁₄-Cys-X₂-Cys. The first member of this family to be identified was the erythroid specific transcription factor, GATA-1. This protein is responsible for regulating the transcription of erythroid expressed genes and is essential for the development of the erythroid lineage (80). All GATA proteins bind with high affinity to DNA targets containing the consensus site, (T/A)GATA(A/G), and use only one of their two metal binding regions in binding. In GATA-1 mutational and deletion studies established that the NH₂-terminal metal binding region is not required for DNA binding. Thus, unlike the class II or class III proteins which bind DNA as dimers and class I and class IV proteins which use tandem arrays of covalently attached zinc domains, the GATA proteins are unique among the zinc binding/DNA binding classes in their ability to specifically bind DNA as monomers.

The NMR structure of the DNA binding domain of GATA-1 bound to an AGATAA containing deoxyoligonucleotide, showed that this domain binds DNA specifically in a metal dependent manner (Fig. 8E). Interestingly, the *in vivo* metal binding preference of GATA-1 is unknown and although Zn²⁺, Fe²⁺, and Co²⁺ all support DNA binding, Fe²⁺ promotes higher levels of DNA binding than either Zn²⁺ or Co²⁺ (80). The significance of this in the context of erythroid cells is unclear. The NMR structure revealed a protein topology for GATA-1 in which the zinc-binding core contacts the major groove while a COOH-terminal arm contacts the minor groove (100). The zinc binding core

consists of two short, irregular, two-stranded β -sheets connected by a loop and followed by an α -helix. Cysteine residues found in the first β -sheet and the NH_2 -terminus of the α -helix coordinate the metal tetrahedrally (Fig. 8E). This zinc binding core is structurally similar to the NH_2 -terminal zinc-binding module of the hormone receptor family (class II zinc proteins) and the superimposition of the corresponding C α s of GATA-1 and the GR results in an rmsd of 1.4 Å.

GATA-1 makes base specific contacts to the major groove by residues in the α -helix and the loop between β -strands two and three. It is notable that contacts to the AT rich conserved GATA site are primarily hydrophobic in nature. The carboxy terminal tail of GATA-1 is absolutely essential, as indicated by deletion studies, for DNA binding. As revealed in the GATA-1 structure, this extended region makes specific contacts with bases in the DNA minor groove in a manner reminiscent of the NH_2 -terminal arms of homeodomains. Thus, the GATA-1 structure provides the first example of a zinc binding protein which uses an extended region outside the zinc core to specifically interact with DNA.

A novel zinc binding motif was recently uncovered with the structure determination of the Cys₄ DNA binding domain of the eukaryotic transcription elongation factor, TFIIS (81). TFIIS, which also contains an RNA Polymerase II (RNAPII) binding domain, enables RNAPII to read through pause or termination sites. The TFIIS DNA binding domain (residues 231-280) consists of a three-stranded β -sheet and disordered loop in which the zinc is tetrahedrally coordinated between four cysteines, a motif which has been called the zinc ribbon (81). How DNA binding is mediated by this motif and what sites it binds is unclear, but studies indicate that it has a preference for single-stranded oligopyrimidines. As DNA sequences in RNAPII pause and termination sites are unusually A rich, it is possible that this motif may interact with the U rich single-stranded RNA. Sequence analyses indicate that analogous zinc ribbons may occur in other proteins involved in DNA or RNA transactions, including RNAPII itself.

The remaining zinc binding proteins with potential DNA binding activities include the LIM, RING finger, CRD containing proteins and MetRS. The LIM domain has been found in over 20 proteins, often in repeats. The name "LIM" derives from the three transcription factors in which it was first observed, Lin-11, Isl-1 and Mec-3c (83-84). Although several of the proteins containing the LIM domain are transcription factors and often contain homeodomains, there is no evidence that this domain binds DNA. The RING finger ('Really Interesting New Gene') is found in over 40 proteins (85). Despite the location of these proteins in the nucleus, DNA-binding activity has yet to be confirmed for any RING finger member. The CRD, or cysteine rich subdomain, was first observed

within the protein kinase C (PKC), an important enzyme in signal transduction, and homologous CRDs have been found in some oncogene products (86). The zinc binding motif found within the Methionyl-tRNA Synthetase protein, MetRS, is structurally homologous to the zinc binding domains of class IV and may represent an offshoot of that family (82). Members from each have been structurally characterized but are beyond the scope of this introduction.

c. b/ZIP, b/HLH and b/HLH/ZIP

The use of dimerization as a means to expand DNA binding specificity is exploited exquisitely by a group of related eukaryotic transcription factors called basic region leucine zipper (b/ZIP), basic region helix-loop-helix, (b/HLH) and basic region helix-loop-helix leucine zipper (b/HLH/ZIP) proteins (41). The b/ZIP, b/HLH and b/HLH/ZIP motifs of these transcription factors are found in the context of larger proteins and form the minimum essential region for DNA binding. Each of these related motifs is composed of a COOH-terminal dimerization segment which consists of either a leucine zipper (ZIP), a helix-loop-helix (HLH) or a helix-loop-helix leucine zipper (HLH/ZIP) and a NH₂-terminal DNA binding segment called the basic region (b) because it is rich in basic amino acids.

The basic region, which is unstructured in solution, becomes α -helical upon DNA binding and thus, provides another example of DNA induced structural ordering (41). The basic region α -helix is utilized by all these proteins as the DNA binding "recognition helix" and all of these proteins use similar strategies for DNA-binding by inserting a dimer of these recognition α -helices into an eight to ten base pair binding site in the DNA major groove. Because these proteins can associate, in many cases, as either homodimers or heterodimers, the dimerization elements are critical for DNA binding specificity as they help determine which subunits form stable dimers. This ability to form homo- and heterodimers results in a vast number of potential multi-protein complexes with diverse DNA binding properties and thus, diverse biological effects.

Some members of these proteins are cell specific and play central roles in determining cell identity. With a few exceptions, heterodimerization is limited to members of each subfamily, i.e., the b/ZIP proteins only heterodimerize with other b/ZIP proteins, the b/HLH only heterodimerize with other b/HLH proteins and the b/HLH/ZIP proteins only heterodimerize with other b/HLH/ZIP proteins. Interestingly, members of this family that do not contain a DNA binding basic region can function as negative regulators by sequestering functional members into inactive heterodimers.

Crystal structures of the b/ZIP, b/HLH and b/HLH/ZIP domains of several proteins bound to DNA have been determined (Fig. 10). Although these structures are prototypical of b/ZIP, b/HLH and b/HLH/ZIP proteins, they suggest that, once again, a simple DNA-binding code, even for a distinct DNA-binding motif, cannot be readily deduced. The b/ZIP protein, GCN4, which functions in the activation of amino acid biosynthesis, provides a remarkable example of a DNA binding protein that can bind two different DNA sites with near equal affinity (101). Specifically, GCN4 can bind the CREB response element (CRE) and the AP-1 site which differ in the spacing of the -TGA(C/G)-half sites. The CRE is a palindrome of abutting -TGAC- half sites (T₁G₂A₃C₄G₅T₆C₇A₈),

whereas the AP-1 site comprises overlapping -TGAC- and -TGAG- half sites ($T_1G_2A_3C_4T_5C_6A_7$). The crystal structure has been determined for both complexes, GCN4-CREB (102) and GCN4-AP-1 (103). In both structures, the b/ZIP dimer consists of two smoothly curving α -helices in which the basic region α -helix and leucine zipper α -helix actually form a single continuous α -helix. The basic region helices fit into the major groove of half sites on opposite sides of the DNA and the leucine zippers form a parallel coiled-coil dimerization interface.

The primary sequence of the GCN4 leucine zipper, like all leucine zippers, is a repeating heptad, **a-b-c-d-e-f-g**, with predominately hydrophobic and apolar residues at positions **a** and **d** and predominately polar and charged residues at other positions (41). Residues **e** and **g** often contain charged residues which can form intrahelical and interhelical salt bridges (104). These salt bridges play a critical role in determining the dimerization preferences of b/ZIP proteins. In the well characterized case of the selective heterodimerization of the proto-oncogenes, c-Jun and c-Fos, which are sequentially very similar to GCN4 and also bind AP-1 sites, charge repulsion between acidic **e** and **g** residues destabilizes the c-Fos homodimer, shifting the equilibrium in favor of the c-Jun/c-Fos heterodimer (105-106). Residues **e** and **g** are also critical in shielding the hydrophobic dimer interface. This interface is stabilized by interactions between residues in positions **a** and **d** with leucine being the preferred residue at position **d**.

Comparison of the GCN4(b/ZIP)-CRE and GCN4(b/ZIP)-AP-1 complexes reveals the surprising finding that GCN4 accommodates the extra central base pair of the CRE site by a slight distortion of the CRE DNA with minimal changes in the protein (102). This distortion of the DNA enables GCN4 to make similar contacts to both the CRE and AP-1 sites. In both complexes a conserved asparagine, Asn235, from each monomer subunit hydrogen bonds to the two contiguous base pairs at the center of each half site which corresponds to base pairs of G_2A_3 and T_6C_7 in the CRE site and G_2A_3 and T_5C_6 in the AP-1 site. Additional contacts are provided by two alanines from each monomer, Ala238 and Ala239, which make van der Waals contacts with thymine methyl groups. Ser242 also makes van der Waals contacts via its $C\beta$ (Fig. 11A).

The differences in the central base pairs of the two sites are accommodated by slightly different contacts which are mediated in large part by Arg243. In the GCN4(b/ZIP)-CRE complex, symmetric contacts are made by each Arg243 to the N7 of the central guanines, G_4 and G_5 , as well as the phosphate group from the adjacent base pair. Because the pseudosymmetric AP-1 site contains only a single central C-G base pair, the two Arg243 side chains make asymmetric contacts. One of the Arg243 side chains

hydrogen bonds to the central guanine while the other hydrogen bonds to two phosphate groups.

While the DNA in the GCN4(b/ZIP)-AP-1 complex showed little departure from standard B-DNA, the DNA in the GCN4(b/ZIP)-CRE complex is bent symmetrically by 20° at its center. This distortion, which is manifested in a helical shift of the phosphodiester backbone, allows GCN4 to make contacts analogous to those in the AP-1 complex and, thus explains how the GCN4(b/ZIP) can recognize either site. It also underscores the potential importance of DNA flexibility in specifying protein-DNA interactions.

The recent structure determination of the c-Fos/c-Jun heterodimer b/ZIP bound to an AP-1 containing site (107) confirmed the predicted similarity of this heterodimer to GCN4 in both the structure of the b/ZIP as well its DNA contacts. The structure also shows that the c-Fos-c-Jun heterodimer can bind the asymmetric AP-1 site in two orientations which are related by a 180° rotation. The different orientations result in only one asymmetry in DNA contacts: In the first complex it is the conserved arginine in Fos, corresponding to Arg243 in GCN4, which contacts the central guanine, whereas in the second complex it is the conserved arginine in Jun which mediates this contact. The other DNA contacts are exactly homologous to those seen in the GCN4-DNA structures and include the contacts made by the conserved asparagine corresponding to Asn235 in GCN4 and the hydrophobic contacts made by two conserved alanines and a cysteine (mutated from serine) which correspond to Ala238, Ala239 and Ser242 in GCN4. As predicted, g-e salt bridges are critical for the stability of the heterodimer. Also important are a-g contacts.

The b/HLH and b/HLH/ZIP transcription factors, which occur in diverse eukaryotes ranging from mammals to yeast, play key roles in regulating metabolism, cell differentiation and development (108). b/HLH proteins which are expressed ubiquitously include E47, Pho4, E12 and HEB. These proteins bind related DNA sites called E boxes and can dimerize with tissue specific b/HLH proteins such as MyoD, SC and tal-1/SCL. Members of the b/HLH/ZIP group include MAX, MAD, the MYC proteins and USF. These proteins contain an additional leucine zipper COOH-terminal to the HLH element. Structures of four b/HLH and b/HLH/ZIP protein-DNA complexes have been determined: MAX(b/HLH/ZIP), USF(b/HLH), E47(b/HLH) and MyoD(b/HLH).

MAX, a b/HLH/ZIP protein dimerizes with the MYC oncoproteins and recent characterization of other b/HLH/ZIP proteins, which also dimerize with MAX, indicate that it may play a central role in orchestrating the biological activity of the b/HLH/ZIP proteins. The crystal structure of MAX(b/HLH/ZIP) bound to a DNA site containing the adenovirus

major late promoter (109), shows that the DNA binding domain of MAX consists of two long α -helices separated by a loop. The NH_2 -terminal α -helix, consists of the basic region and the first helix of the HLH region, H1. The second α -helix is also contiguous and is composed of the second helix of the HLH element, H2, and the leucine zipper α -helix. In this structure, Max binds to the DNA as a homodimer and the two monomers fold in to a parallel, left-handed four helix bundle. In the dimerization interface, H2 is slightly bent towards the DNA and away from H2' of the opposing subunit, giving the b/HLH region of MAX a less upright appearance than the b/ZIP basic region of GCN4. Apolar residues are found at the a and d positions of H1 and H2 as in GCN4, however, unlike GCN4, the e positions of these regions are also apolar.

Most b/HLH containing proteins bind the DNA hexamer sequence, $\text{C}_1\text{A}_2\text{N}_3\text{N}_4\text{T}_5\text{G}_6$, where the N indicates any base. In the cocrystal structure, the MAX homodimer makes numerous contacts with the conserved base pairs of the hexamer site, the most critical being hydrogen bonds from a conserved glutamic acid, found in the basic region of each subunit, (corresponding to Glu32 in MAX) to the C_1 and A_2 of one half site and the C_6 and A_5 of the other half site (Fig. 11B). This side chain also packs tightly against the thymine methyl groups of T_2 and T_5 . In MAX, His28, located one helical turn, NH_2 -terminal to Glu32 in the basic region, also contributes to the binding at the conserved hexamer bases by donating a hydrogen bond to G_1 and G_6 .

The binding specificity observed between b/HLH and b/HLH/ZIP is dependent upon the identity of the central base pairs, N_3N_4 , and it had previously been established that b/HLH and b/HLH/ZIP proteins exist in two so-called specificity groups. The first specificity group binds hexamer sites containing central CG sequences and the second binds hexamers containing a central GC sequence. Proteins of either specificity class are distinguished by the identity of the residue that is located four residues COOH-terminal to the conserved glutamic acid, implicating this residue as the specificity determining residue. The specificity class containing MAX, contains an arginine at this position and prefer hexamers with central CG sequences, while the second class contains apolar residues at this position and prefer hexamers with a central GC sequence. In the MAX(b/HLH/ZIP)-DNA structure this arginine from one subunit hydrogen bonds to the central guanine, G_3 , while the corresponding arginine from the other subunit makes an equivalent hydrogen bond to G_4 . No other protein contacts are made to the central base pairs in the the MAX(b/HLH/ZIP)-DNA structure. Thus, this arginine appears to play the anticipated role as the DNA specificity determining residue. Finally, numerous contacts are made to the entire backbone of the hexamer sequence by residues from the basic region.

Structures of the b/HLH regions of the b/HLH/ZIP protein, USF (110), and the b/HLH regions of the b/HLH proteins, E47 (111) and MyoD (112) (Figs. 10B and 11B), bound to DNA, have confirmed the critical protein-DNA contacts seen in the MAX structure and have also revealed the unexpected importance of other residues in DNA binding. Like MAX, the b/HLH of USF, the upstream stimulatory factor, contains a "specificity determining" arginine and the crystal structure of the b/HLH of USF bound to a DNA site containing the hexamer sequence, C₁A₂C₃G₄T₅G₆, revealed protein-DNA interactions consistent with the MAX structure (109). Interestingly, USF, which is a b/HLH/ZIP protein, can homotetramerize and this tetramerization depends on the integrity of the leucine zipper. Unfortunately, because the USF-DNA structure was determined in the absence of the ZIP region, a better understanding of this tetramerization and its mechanism must await further structural studies.

The b/HLH proteins E47 and MyoD are from the second specificity determining class, which bind hexamers with a central GC sequence. E47 deviates slightly from other members of this class in preferring the nonpalindromic hexamer site, CACCTG instead of CAGCTG. Structures of the b/HLH regions of E47 (111) and MyoD (112) bound to DNA have begun to reveal the unforeseen means by which discrimination can be conferred at the central base pairs. Rather surprisingly, it turns out that the key "specificity determining" residue, which is apolar for members of this class, does not contact the DNA in either the E47 or the MyoD complex. In the E47(b/HLH)-DNA structure, only one side chain-base contact is observed to the central base pairs. Arg346, from one subunit located one residue NH₂-terminal to the conserved glutamic acid, contacts a central guanine. However, this contact is only seen in one half site due to the nonpalindromic nature of the central base pairs. The MyoD(b/HLH)-DNA structure, although it makes protein-DNA interactions consistent with the MAX structure, is more surprising, as no protein contacts to the central base pairs are observed in this structure (Fig. 11B). Therefore, its preference for the hexamer, CAGCTG, is left unexplained.

d. β -sheet binding proteins

As underscored by the preceding sections, most well-characterized DNA binding transcription factors employ α -helices for the recognition of bases in the major groove. However, the structures of several prokaryotic and viral proteins, which use β -sheets for DNA recognition, have recently expanded the possibilities of what may constitute a DNA major groove recognition element (42). Thus far, two classes of β -sheet DNA binding proteins have been identified (35). The first class includes the *E. coli* MetJ repressor and the arc and mnt repressors from the Salmonella phage P22. The second class includes the bacterial protein, HU, which forms condensed nucleoprotein structures in many prokaryotes, the *E. coli* integration host factor, IHF, and the TF1 transcription factor from *B. subtilis* phage. The eukaryotic TATA-binding proteins comprise a third class of proteins that use β -sheets in DNA binding, however, unlike the preceding classes, these proteins bind in the DNA minor groove and, therefore, will be discussed in section 2b.

Thus far, structures of the class one proteins, MetJ (113) and Arc (114) have been determined bound to DNA (Fig. 12). These structures, unlike most of the protein-DNA structures discussed thus far, include the entire protein. This result is, in part, due to the relatively small size of these proteins. The DNA-binding domains of these proteins contain a similar ribbon-helix-helix structure. Dimerization of these proteins is required for stable folding as both monomers are essential for the formation of a hydrophobic core. β -strands of each monomer pair to form the antiparallel β -sheet that ultimately binds in the DNA major groove. Because the dimer contains two ribbon-helix-helix elements, in which helix 1 is referred to as helix A and helix 2 is helix B, the dimer fold has been named the $(\beta\text{-}\alpha\text{-}\alpha)_2$ motif. The naturally occurring operator sites for the Arc repressor have adjacent binding sites for two $(\beta\text{-}\alpha\text{-}\alpha)_2$ elements as do the MetJ binding sites, which are called Met boxes. The adjacent arrangement of the binding sites in the Arc and MetJ operators, as well as biochemical studies, suggested that the biologically active, DNA bound form of these proteins are tetramers and that each dimer binds in an adjacent binding site (115-117). This was confirmed by the cocrystal structures of both Arc and MetJ bound to DNA.

Comparison of the cocrystal structures of MetJ bound to the 19 base pair MetJ operator site, $T_1T_2A_3G_4A_5C_6G_7T_8C_9T_{10}A_{11}G_{12}A_{13}C_{14}G_{15}T_{16}C_{17}T_{18}A_{19}$, containing two consensus Met boxes (in bold) (Fig. 12A) and Arc bound to its wild type 22 base pair operator site, $T_1A_2T_3A_4G_5T_6A_7G_8A_9C_{10}T_{11}G_{12}C_{13}T_{14}T_{15}C_{16}T_{17}A_{18}T_{19}C_{20}A_{21}T_{22}$ (Fig. 12B), (binding sites for each dimer in bold) reveal important similarities and differences between the two proteins and their DNA complexes (113-114). Similarities include the formation of a tetramer upon DNA binding, the general position and orientation of the $(\beta\text{-}\alpha\text{-}\alpha)_2$ motif

relative to each DNA half-site, the use of residues from the antiparallel β -sheet in making all critical base contacts to the DNA and similar phosphate contacts made by residues from an NH_2 -terminal tandem turn region. Interestingly, the tandem turn region, which precedes the β -strand, becomes ordered only upon DNA binding. Differences between the two complexes include the spacing of the operator half-sites, the DNA structure of the two operators, the requirement of the corepressor, S-adenosylmethionine, for DNA binding by MetJ and no such small molecule binding requirement for Arc and the regions of the proteins used for dimer-dimer interactions. The different dimerization interfaces of MetJ and Arc reflects the different operator half-site spacings of their binding sites. A further notable difference between the two complexes, which can be deduced by comparing structures of complexed and uncomplexed Arc and MetJ, is that DNA binding by Arc, but not MetJ, is accompanied by conformational changes in the β -sheet involved in DNA binding.

In both protein-DNA complexes, phosphate contacts are crucial for proper docking of the $(\beta\text{-}\alpha\text{-}\alpha)_2$ element to the DNA and in orienting the β -sheet properly for making specific base contacts. Taking as a reference frame the β -sheets in the major groove shows that both Arc and MetJ contact three homologous phosphates on each DNA strand. Furthermore, corresponding regions of both proteins contact the same phosphates. These contacts are made by analogous main chain amides and side chains from the tandem turn regions of Arc and MetJ as well as the NH_2 -terminus of α -helix B. There are also non-conserved backbone contacts, although biochemical and mutant analyses indicate that these are not critical for DNA binding. Interestingly, although both proteins contact their DNA sites in a similar manner, there are large difference in DNA conformation between the two complexes. In the Arc-DNA structure, the DNA is bent 50° with the majority of the bending occurring in the center of the operator. Also, the Arc bound operator DNA reveals a widened major groove and compressed minor groove. Although the DNA in the MetJ-DNA complex is also bent by 50° , most of the bending is located near the center of each half site, rather than the center of the operator. As a result the MetJ bound operator DNA displays compressed major grooves. Moreover, the central $T_{10}pA_{11}$ step is overwound and the flanking steps are underwound. This distortion, which may reflect a sequence dependent distortion, is associated with the anomolous positioning of the phosphates directly adjacent to the $T_{10}pA_{11}$ step. These phosphates are displaced by 2 Å from their positions in regular B DNA. This displacement is critical as it allows these phosphate groups to make contacts with the NH_2 -terminus of helices B and B'. As these interactions may rely on the sequence dependent deformability or flexibility of the DNA, they may constitute an example of indirect readout.

The ability of Arc and MetJ to recognize their operator sites with high affinity depends on the direct readout of bases in the major groove by residues in the antiparallel β -sheets of each protein. The general way in which the β -sheets of the two proteins approach the DNA is very similar ($C\alpha$ position of corresponding β -sheet residues have an rmsd of 1.15 Å). Given this similarity, it is surprising that the side chain-base contacts made by β -sheet residues are completely different in the two complexes (Fig. 13). Additionally, the β -sheet region of Arc undergoes conformational adjustments upon DNA binding whereby the side chains of Phe10 and Phe10' (where the prime indicates the other monomer subunit) swing out from their buried positions and pack between adjacent phosphate groups. Concomitantly, the $C\alpha$ atoms of residues 9, 9', 10 and 10' move by 1.0-1.7 Å to fill the cavity left by this movement. The main chain movements of Gln9 and Gln9' help to position these side chains for base contacts. These side chains from one Arc dimer make symmetric contacts with adenines on opposite DNA strands, A₆ and A₇, of the first half site while the corresponding side chains on the other dimer make homologous contacts, although the bases are slightly different. The homologous β -sheet residues in MetJ, Lys23 and Lys23', make symmetric contacts with G₄ and G₉', respectively. As can be seen in Fig. 13, the contacts made by residues from Arc and MetJ show dramatically different spacings, which appear to arise from the differences in the side chain rotations of the residues involved. The only other side chain-base contact found in MetJ is between Thr25 and Thr25' to A₈' and A₅, respectively. The corresponding residue from one monomer subunit in Arc is Asn11. This residue contacts the analogous position in the Arc operator, C₈. However, Asn11', from the other monomer subunit, contacts T₇'. Thus, the contacts made by the two Asn11 and Asn11' are asymmetric. A third residue in Arc, Arg13 and Arg13', makes additional asymmetric contacts with guanines, G₈ and G₄. In MetJ, the corresponding residues are Ser27 and Ser27'. No DNA contacts are provided by these residues as they are too short to reach the bases (Fig. 13).

As discussed, the high affinity binding of MetJ to its operator site depends on corepressor (S-adenosylmethionine or SAM) binding to MetJ. This binding increases the affinity of MetJ for its operator by 1,000 fold. In the MetJ-DNA complex (114), the SAM molecules that bind to each monomer, interact with the faces of the repressors, remote from the DNA, with their positively charged sulfur atoms positioned at the COOH-terminus of the B helices. Unexpectedly, SAM binding does not cause significant structural changes in MetJ (118). Furthermore, the overall structure of the holorepressor when bound to DNA is little changed from the structure of the holorepressor itself. Therefore, these results cannot directly explain corepressor activation of MetJ. A result, which implicates charge as a mechanism, was provided by studies showing that S-adenosylhomocysteine, a SAM

analog lacking the S-methyl group and containing a neutral rather than positively charged sulfur atom, binds to MetJ with high affinity but does not affect its affinity for operator. This suggests that long-range electrostatic interactions between bound corepressor and DNA phosphate groups may explain activation (119).

e. Others

In addition to the aforementioned, well established families of DNA binding proteins, X-ray studies have revealed other DNA binding motifs which, so far, appear to be structurally distinct. Examples include the DNA binding domains of the bovine papillomavirus 1 (BPV-1) E2 protein (46), the eukaryotic tumor suppressor, p53 (45) and the rel homology protein, NF- κ B (43-44). Full length E2, consisting of 410 amino acids, is a *trans*-acting transcription regulator and the truncated forms, E2TR and E8/E2, which are missing the NH₂-terminal *trans*-activation domain, act as repressors. The X-ray crystal structure of the COOH-terminal DNA-binding/dimerization domain, residues 326-410, bound to an E2 DNA target site, revealed that E2(326-410) forms an eight-stranded antiparallel β -barrel made up of four strands from each subunit (42). Residues from a pair of α -helices, which are symmetrically disposed on the outer circumference of the β -barrel make all base-specific contacts to a conserved set of three base pairs found in each half site. The DNA is smoothly bent in the complex forming an extensive network of specific interactions that are coupled with 48 phosphate contacts.

The tumor suppressor, p53, also consists of an NH₂-terminal *trans*-activation domain. However, its oligomerization domain, conferring tetramer formation, is located at the extreme COOH-terminus of the protein, following the highly conserved DNA binding domain, residues 102-292. The crystal structure of the p53 DNA-binding domain bound to a p53 half site, shows that the DNA binding domain of p53 can be divided into two functional regions. The first region contacts the major groove and the second region contacts the minor groove (45). Contacts to the major groove are provided by the so-called "loop-sheet-helix" motif which is composed of noncontiguous regions. Contacts to the minor groove are provided by two large loops which are structurally stabilized by the tetrahedral coordination of a zinc. Three cysteines located on loops and one histidine from a small helix, coordinate this zinc. Residues of the loop-sheet-helix motif and the two loops, perhaps expectedly, constitute the most conserved regions of p53 and contain the majority of the p53 mutations identified in tumors.

The transcription factor NF- κ B, composed of a heterodimer of p50 and p65 proteins, was initially identified as a protein which specifically binds the κ immunoglobulin light chain enhancer. It was subsequently found to regulate a large number of genes involved in the infection response and stress. NF- κ B, unlike E2 and p53, is a member of a large, well-established family of homologous proteins which all contain a rel homology region (named after the rel oncogene) (43-44). Thus, the recent structure determinations of the NF- κ B p50 homodimer bound to κ B sites, can serve as a model for other rel members (43-44). The rel homology region (RHR) of p50, like all rel proteins, contains both an

NH₂-terminal DNA binding domain and a COOH-terminal dimerization domain. It also contains a nuclear localization sequence, (NLS), COOH-terminal to the dimerization domain, that is also largely responsible for the binding of I- κ B, the NF- κ B inhibitor.

The structure of the NF- κ B-DNA complexes resembles a butterfly with the protein playing the part of the wings, and the DNA the body. Interestingly, DNA contacts are made by flexible loop regions rather than regions of regular secondary structure. Most of the contacts are provided by the immunoglobulin-like DNA binding domain. However, the dimerization domain, which also contains an immunoglobulin-like fold, makes several important contacts to the DNA. The use of flexible loops in making DNA contacts may allow for a more adjustable binding specificity. A critical detail, which points out the complexity of the the NF- κ B-DNA interface, is that the base contacts contributed by one subunit of the NF- κ B p50 dimer, interleave with those of the opposing subunit. This presents a problem in predicting the various rel homology protein structures which include homodimeric and heterodimeric proteins.

2. Minor groove recognition

a. Homeodomain NH₂-terminal arms

The structures of homeodomain-DNA complexes have established that base specific contacts to bases in the major groove are made by residues found within helix 3, the recognition helix. However, these few interactions and, even the accompanying phosphate contacts, cannot explain the specificity and affinity of a given monomeric homeodomain for its DNA site. Additional and essential contacts are provided by a highly conserved region, the "NH₂-terminal arm". This arm fits into the minor groove and makes contacts near the 5' end of the core consensus of each homeodomain operator site. In both the engrailed and Oct-1 homeodomain-DNA complexes, two conserved basic residues at positions three and five in the arm make contacts in the minor groove. In engrailed, the side chain of Arg3 appears to make a hydrogen bond to T₂' (where the prime, again, indicates the complementary DNA strand) in the sequence, T₁A₂A₃T₄, as well as contacting a sugar oxygen (60). In Oct-1, the analogous residue is a lysine, Lys3. The side chain of this residue also appears to make minor groove contacts, but, unfortunately, a combination of the relatively low resolution (3 Å) and poor electron density in this area prevent an unambiguous assignment of its contacts (62). The situation is much clearer for residue 5, which is an arginine in both engrailed and Oct-1. This residue makes analogous contacts in both structures, whereby the side chain contacts the base at position 1'. In Oct-1, this base is a thymine as the Oct-1 homeodomain binds the sequence, A₁A₂A₃T₄. In Oct-1, Arg5 hydrogen bonds to the O2 of T₁', whereas in engrailed Arg5 contacts the O2 of T₁. These structures provide an explanation for the observed preference of an AT or TA base pair at position 1 in homeodomain operators as the N3 of adenine and the O2 of thymine represent analogous hydrogen bond acceptors, whereas a GC base pair would interfere with binding as the arginine side chains would clash with the exocyclic N2 group of guanine.

The Oct-1-DNA structure is especially critical in examining contacts made by the "arm" as it provides the opportunity to observe these contacts in the context of a larger protein. For Oct-1, the "arm" is attached to the COOH-terminus of the POU specific domain (62). Despite this constraint, the Oct-1 homeodomain is similarly docked and its "arm" makes similar contacts to the DNA as those described for the engrailed homeodomain-DNA complex (Fig. 4C). The yeast protein, Mat α2, contains a divergent homeodomain that, nonetheless, possesses not only the three-helix homeodomain fold but docks against and contacts the DNA in the same manner as the engrailed and Oct-1 homeodomains (61). The sequence of the "arm" of Mat α2 differs significantly from other homeodomains, despite being rich in basic amino acids. One of the basic residues, Arg7,

contacts base pairs 1 and 2 in the minor groove of the sequence, $T_1A_2A_3A_4T_5$. The contact to base pair 2 is somewhat analogous to the contact to base pair 1 in the engrailed and Oct-1 homeodomains. The contact to base pair 1 in Mat $\alpha 2$, however, has no homologue in the engrailed and Oct-1 DNA complexes.

Hence, contacts made by the "arms" of the homeodomain proteins provide an explanation for the observed discrimination against GC base pairs at positions 1 and 2 in homeodomain operators and thus, provide an example of specificity conferred by minor groove binding.

b. The TATA binding protein

A well documented example of a specific minor groove binding protein is the TATA-box binding protein, TBP. TBP serves as a linchpin for the assembly of basal preinitiation complexes and is required by all three eukaryotic nuclear RNA polymerases (120). The mechanisms of action of TBP are best understood for the initiation of class II nuclear genes that utilize RNA polymerase II and general initiation factors, TFIIA, -B, -E, -F, -G/J, -H and -I (121-123). In pol II initiation complexes, TBP is associated with TBP-associated factors (TAFs) in a 700K complex known as TFIID, which binds the TATA element of class II promoters (124). The amino-terminal segment of TBP varies in length and sequence depending on its cellular source and is not required for basal transcription. By contrast, the COOH-terminal, 180 residue, domain is highly conserved phylogenetically, is extremely sensitive to mutations and is essential for basal transcription. Two crystal structures of the COOH-terminal domain of TBP, one from yeast, γ TBP, and one from *Arabidopsis thaliana*, TBP2, bound to TATA elements, show similar dramatic changes in DNA conformation and provide a basis for understanding the minor groove specificity of TBP (125-126).

In these structures, the COOH-terminal domain of TBP, which contains two 90 residue repeats sharing 28% identity, forms two highly symmetric $\alpha\beta$ regions each comprised of a five-stranded antiparallel β -sheet and two α -helices. The two five-stranded β -sheets of each domain form a continuous ten-stranded β -sheet between domains. The overall structure resembles a molecular saddle with a concave underside lined by eight strands of the interdomain ten-stranded β -sheet. The convex surface of the saddle is composed of four α -helices, a basic peptide which links the two domains and parts of the first β -strands. This surface appears to be responsible for most of the myriad interactions between TBP and other proteins.

In the TBP cocrystal structures, the TATA element (TATAa/tAa/t), is severely bent, providing the first examples of protein induced DNA bending toward the major groove. This bending is mainly mediated by the insertion of two conserved phenylalanines between base pairs 1 and 2 and 7 and 8 in the TATA elements, $T_1A_2T_3A_4T_5A_6A_7A_8$ and $T_1A_2T_3A_4A_5A_6A_7G_8$, of the yeast TBP and *Arabidopsis thaliana*, TBP cocrystal structures, respectively. The insertion of these phenylalanines splays open the minor groove, exposing it to the underside of the TBP molecular saddle. Surprisingly, the TBP-minor groove interface is mainly hydrophobic in nature with only six hydrogen bonds formed between TBP and bases of the DNA. These interactions occur at the center of the TATA box and involve hydrogen bonds from asparagine side chains, which donate hydrogen bonds to base pair 4 and 5. Additionally, the N3 groups of A_5 and A_4 also

accept hydrogen bonds from threonine side chains from each domain. Reminiscent of the homedomain arm, the preference for the AT and TA over GC or CG base pairs in TBP can be explained by steric clash that would result from the presence of the exocyclic N2 group of guanine. This selection against GC or CG base pairs provides part of the explanation for the specificity of TBP for its TATA element. Sigler *et al.* have also implicated 'indirect readout' as being a contributing factor in TBP DNA binding specificity based on the observation that alternating pyrimidine/purine sequences, especially TATA, are especially flexible (125). This, Sigler, speculates, may be important in the 5' half of the TATA element in providing a mechanism for the directionality with which TBP binds to the TATA element.

The most notable features of the DNA in the TBP-DNA complexes are the large kinks (127) which are introduced between base pairs 1 and 2 and between 7 and 8 which result in unusually large roll angles ranging from 37° to 47°. Remarkably, the DNA outside the eight base pair TATA element, in each structure, returns abruptly to B-DNA. These structures, along with the CAP-DNA complex (50) and the PurR-hypoxanthine-*purF* complex, reported in this thesis, provide the only direct examples of protein induced DNA kinking. The unusual DNA kinking toward the major groove observed in the TBP-TATA structures and now the PurR-hypoxanthine-*purF* complex, provides a precedent for other DNA binding proteins, a notable example being the eukaryotic protein, SRY (sex determining region on chromosome Y). Indeed, NMR data has suggested that SRY may kink its DNA site away from itself and towards the major groove (128). The biological implications of DNA kinking towards or away from the major groove have yet to be established.

D. Structure Determination by X-ray Crystallography

Much of the modern field of biology owes its sophistication to the success of macromolecular X-ray crystallography. Indeed, before three-dimensional structures became available through X-ray work no one really knew how proteins such as enzymes might work. In this section, some of the theory behind macromolecular crystallography will be described. A more detailed discussion covering the fine points of crystallography theory can be found elsewhere (129-133).

Macromolecular crystallography may be defined as the application of the techniques of X-ray diffraction to crystals of macromolecules. Such crystals are ordered three-dimensional arrays of macromolecular aggregates that repeat periodically in all directions. The basic building block of the crystal is the unit cell which is characterized by six parameters, three axial lengths and three interaxial angles (130). The lengths of the unit cell edges are designated **a**, **b** and **c** and the interaxial angles are α , β and γ where α is the angle between **b** and **c**, β the angle between **a** and **c** and γ the angle between **a** and **b**. These six cell parameters are used to characterize each of the seven basic crystal systems (triclinic, monoclinic, orthorhombic, trigonal, tetragonal, hexagonal and cubic). The triclinic case represents the system with the lowest symmetry where all six parameters can assume any value. In the other six systems, symmetry decreases the number of independent parameters as indicated in Table 1. In the monoclinic system, which is represented in this thesis by the CBD crystals and the TMMP hemoglobins crystals, the convention has been established that the **b** axis, which is perpendicular to both the **a** and **c** axes, is unique and the angle $\beta > 90^\circ$. This indicates that for the monoclinic case, there are four independent parameters, **a**, **b**, **c** and β , that need to be determined. The other crystal system discussed in this thesis and represented by the PurR-hypoxanthine-*purF* operator crystals, is the orthorhombic lattice. In this system, $\mathbf{a} \neq \mathbf{b} \neq \mathbf{c}$ and $\alpha = \beta = \gamma = 90^\circ$.

The seven crystal systems just discussed contain a lattice point at each corner of the unit cell. Such lattices are termed primitive and are designated by the letter P. In addition to these primitive lattice systems, there are seven other crystal systems termed nonprimitive that contain two or more lattice points per unit cell. Nonprimitive lattices with a pair of lattice points centered on opposite faces of the unit cell are termed A, B or C depending on whether the **bc**, **ac** or **ab** faces are centered. If a lattice has a point at the center of its unit cell it is designated I (interior) and if all faces of the unit cell contain points at their centers, the corresponding lattice is termed F (face centered).

In an array of lattice points it is always possible to define a triclinic, primitive cell regardless of symmetry. However, to disregard symmetry would be to neglect the simplification it provides. Thus, the cardinal rule in the selection of a unit cell is to chose it

such that the lattice conforms to the highest symmetry present. Thus, the origin of the unit cell is determined to take advantage of symmetry elements. Examples of such elements are rotary axes, mirrors, centers of symmetry and rotary inversion axes. Crystals can be classified in terms of the groups of symmetry operators relating their faces. Each of these groups, termed point groups, denotes one of the possible unique combinations of crystallographic symmetry elements. There are 32 possible such point groups. The combination of these 32 point groups with the 14 Bravais lattices leads to 230 unique arrangements of points in three dimensional space. These 230 unique arrangements, called space groups, describe the ways in which identical objects can be arranged in an infinite lattice. They result from a combination of point groups with translational elements and include screw axes and glide planes. Not all 230 space groups are allowed for protein crystals because the application of mirror planes and inversion centers would change the asymmetry of the amino acids from the L to the D isomer. This limits the number of possible space groups to 65, leaving only those without symmetry (triclinic) or exclusively with rotary or screw axes (130-131).

If the crystal lattice has a symmetry higher than triclinic, each molecule in the unit cell will be repeated a number of times as determined by the symmetry operators which define the given system. For example, in $C222_1$, the space group of the PurR-hypoxanthine-*purF* operator crystals, there are eight equal molecular assemblies, called asymmetric units in the unit cell that are all related by crystallographic symmetry. An important aspect of the asymmetric unit is that each one has an identical environment. If a molecule itself has two-fold symmetry, it is possible that its two-fold axis may coincide with a crystallographic two-fold axis. Such a molecule occurs in a "special position". An example of such a case is the PurR-hypoxanthine-DNA structure where there is a PurR monomer, hypoxanthine molecule and *purF* half-site in the asymmetric unit and thus the PurR dimer and full DNA site is generated by the crystallographic two-fold relating the positions x, y, z and $x, -y, -z$. The converse situation is found for the CBD crystals which contain a CBD dimer in the asymmetric unit. In this case, each CBD monomer is in a different environment, as reflected by the slight structural differences between the two subunits.

The diffraction of X-rays by macromolecular crystals was first observed by Max von Laue in 1912. This observation established the wave nature of X-rays. In order for an object to diffract light, the wavelength must be no larger than the object. Visible light, which is electromagnetic radiation with wavelengths of 400-700 nm, cannot produce an image of the individual atoms which are found in macromolecules where bonded atoms are only about 0.15 nm (1.5 Å) apart. Electromagnetic radiation of this wavelength (0.1 - 100

Å) falls in the range of X-rays. Although individual atoms diffract X-rays, it is not possible to obtain an image of a single molecule for two reasons. First, X-rays cannot be focused by lenses and second, a single molecule is a very weak diffractor of X-rays. Analyzing diffraction from macromolecular crystals, however, solves this problem because a crystal contains many ordered molecules in identical orientations such that each molecule diffracts identically and therefore, the diffracted beam for all molecules augment each other to produce strong and detectable diffraction patterns.

W. L. Bragg was the first to describe X-ray diffraction from crystals as reflection from lattice planes and he deduced a simple equation based on this description (134). This equation, known as Bragg's law, is a cornerstone in X-ray diffraction and can be derived by considering an X-ray beam that is incident on a pair of lattice planes, P_1 and P_2 with interplanar spacing d (Fig. 14A). The parallel incident X-rays, 1 and 2, make an angle, θ , with these planes. The electrons at positions C and D are forced to vibrate by the oscillating incident beam and will radiate in all directions. However, for the particular direction where the parallel secondary rays, 1' and 2', emerge at angle, also θ , a diffracted beam of maximum intensity will result if the waves represented by these rays are in phase. Dropping perpendiculars from D to A and B, respectively, shows that $\angle ADC = \angle BDC = \theta$ and thus, $AC = BC$ and waves in ray 2' will be in phase with those in 1' if $AC + CB = 2AC$ is an integral number of wavelengths. This can be expressed as $2AC = n\lambda$ where n is an integer and λ is the wavelength. By definition, $AC/d = \sin\theta$ and appropriate substitution gives:

$$2d\sin\theta = n\lambda \quad (\text{Bragg's law})$$

Examination of Bragg's law in the form $\sin\theta = (n\lambda/2)(1/d)$ reveals that $\sin\theta$ is inversely proportional to d , the interplanar spacing in the crystal lattice. $\sin\theta$ is a measure of the diffracted beams deviation from the direct beam and therefore crystals containing elements with large d with show compressed diffraction patterns. Interpretation of the X-ray diffraction pattern can be facilitated by replacing this inverse relation between $\sin\theta$ and d with a direct one through a theoretical construction defined as the reciprocal lattice. The reciprocal lattice is advantageous in that it allows the direction of scattering to be easily constructed. The reciprocal lattice can be constructed as follows: First consider normals to all direct lattice planes, of the form h,k,l , that radiate from some lattice point taken as the origin. Next terminate each normal at a distance equal to $1/d_{hkl}$ from the origin where d_{hkl} is the perpendicular distance between planes of the set h, k, l . This set of points constitutes the reciprocal lattice.

P. P. Ewald noted a geometrical representation of Bragg's law that links reciprocal space to diffraction (135). As seen in figure 14B, if the Bragg's reflection angle is θ then the total angle of deflection is 2θ . If one places an intersection of X-ray beams with a plane at the center of a sphere, the Ewald's sphere, that has radius of reciprocal wavelength, then one can make the following observation: $1/2(OR)/(1/\lambda) = \sin\theta$ or $\lambda = (2\sin\theta)/OR$. Or if one substitutes $OR = 1/d$ then one obtains Bragg's law with $n = 1$. What is significant about this construction is that the chord OR is perpendicular to the reflecting plane and when one defines its length to be the reciprocal of the distance between diffracting planes, Bragg's law is obeyed. This establishes that whenever a reciprocal lattice point coincides with a circle constructed as described, Bragg's law is satisfied and reflection occurs. Thus, this construction directly connects the concept of the reciprocal lattice to diffraction.

Knowledge of the direction of the diffracted X-rays allows for the determination of the dimensions of the unit cell, however, the direction of X-ray scattering does not reveal the content of the unit cell, which is the information of interest. The molecular structure and arrangement of the molecules in the cell determine the intensities of the diffracted beams and thus, a relationship between the intensities and the structure must be determined. Because X-rays are scattered exclusively by electrons, this relationship is actually between the diffraction data and the electron density distribution within the unit cell. Each diffracted X-ray that is recorded, because of its periodic nature, can be described by a Fourier series (136-137). The Fourier series that describes a diffracted X-ray is called a structure factor equation. The structure factor, F_{hkl} , can be represented as a Fourier series in which each term gives the contribution of each atom to the reflection hkl as:

$$F_{hkl} = \sum_{j=1}^N f_j \exp 2\pi i (hx_j + ky_j + lz_j)$$

where f_j is the scattering factor of the j^{th} atom and x_j , y_j and z_j are the coordinates of each j^{th} atom in the unit cell and hkl denotes the indices of each reflection in reciprocal space. In this equation each atom is treated as a simple sphere of electron density. Alternatively, F_{hkl} can be expressed as the sum of contributions from each volume element of electron density in the unit cell as:

$$F_{hkl} = \int_V \rho(x,y,z) \exp 2\pi i (hx + ky + lz)$$

where V is the volume of the unit cell. These equations describe diffraction in terms of the unit cell's electron density, however, the crystallographer initially is much more interested

in the determination of $\rho(x,y,z)$ rather than F_{hkl} . This is accomplished by Fourier transformation as F_{hkl} is the Fourier transform of $\rho(x,y,z)$ and therefore:

$$\rho(x,y,z) = (1/V) \sum_h \sum_k \sum_l |F_{hkl}| \exp(-2\pi i(hx + ky + lz) + i\alpha'_{hkl})$$

where V , again, is the unit cell volume. This equation tells the crystallographer how to calculate electron density, however, only the structure factor amplitudes, $|F_{hkl}|$, of each reflection can be experimentally obtained. The remaining variable in the above equation that needs to be determined is α'_{hkl} or the phase of each reflection. Therefore, in order to compute $\rho(x,y,z)$, the phase of each reflection must somehow be determined. This is known as the phase problem. In macromolecular crystallography, there are several ways in which to determine phases. The two methods that were utilized in this thesis, multiple isomorphous replacement (MIR) and molecular replacement (MR), are the most common. These two techniques will be described in more detail in subsequent sections.

The determination of the structure of a macromolecule by X-ray crystallography requires several steps:

- I. Growth of data quality crystals.
- II. Determination of cell type/space group and unit cell dimensions.
- III. Collection of three-dimensional intensity data.
- IV. Determination of phases.

The first step in macromolecular structure determination, the growth of data quality crystals, is the least understood step and is mainly a trial-and-error procedure (138). The first parameter that should be checked before any crystallization attempts are made is the purity of the macromolecule under study. Impurities may prevent crystallization or preclude the growth of data quality crystals. Once the purity of the sample has been established and it has been prepared at a sufficiently high concentration, the most common approach to macromolecular crystallization is to bring the macromolecule solution, gradually, to supersaturation, i.e., to force the macromolecule out of solution. The strategies used to do this are to deprive the macromolecule of sufficient ions or water to maintain hydration, to disrupt their hydration layers, or to exclude them from the bulk solvent. These changes are brought about by the addition of certain precipitating reagents, such as high salt or polyethylene glycol. An alternative approach is to decrease the dielectric properties of the solution by the addition of organic solvents. The hope in all of

these strategies is that a condition will be found which favors intermolecular interactions leading to well-ordered crystals.

Several microtechniques have been developed for the crystallization of macromolecules where an important consideration is often the small amounts of sample usually available. However, all of these techniques depend on the same principles of supersaturation. The most common microtechnique, which was utilized in the work described in this thesis, is the hanging drop-vapor diffusion technique. This method, which allows one to screen rapidly a large number of precipitants, is carried out by mixing a small amount of macromolecule solution (2 μ l) with an equal volume of precipitant on a siliconized cover slip, turning the cover slip over and sealing the cover slip over a microtiter dish that contains 1 ml of the precipitant. In this closed system, water is gradually transferred through the vapor phase from the macromolecule "drop" to the more concentrated precipitant in the reservoir until the two are in equilibrium.

The macromolecular crystallographer has not only to crystallize the macromolecule, however, but also to grow crystals of substantial size. This is necessary because the intensity of the X-ray diffraction of a given crystal is roughly proportional to a crystal's volume. There are three general stages in obtaining large macromolecular crystals. The first is the formation of stable nuclei, the second is the growth of the nuclei to form mature crystals and the last is the termination of growth. An important consideration is that the conditions which favor nucleation do not favor optimal growth. Thus, to obtain large crystals, supersaturation must be reduced to the lowest level at which nucleation can occur. Also, crystallization is a multiparameter problem such that several parameters may be varied in searching for optimal crystallization conditions. The most obvious parameters are protein concentration, the nature and concentration of the precipitant, pH and temperature. An appropriate statistical design to approach this multiparameter problem is the factorial or incomplete factorial method in which a table is constructed with a number of values entered for each parameter (139). After conditions are found which yield crystals, the conditions can then be optimized. Although such methods may hasten the identification of favorable crystallization conditions, macromolecular crystallization remains a mostly empirical procedure and, usually a large number of conditions must be tried before any success is gained.

Once suitable crystals have been obtained, the next step is to obtain its cell parameters. To do this the crystal is mounted in a glass capillary which is then sealed at both ends with synthetic mother liquor (usually solution from the crystallization reservoir) to keep the crystal hydrated. Macromolecular crystals are usually anywhere from 30% to 70% solvent so that if they are not kept wet, they will dehydrate and the lattice will be

destroyed. The cell parameters of the crystals in this thesis were all determined by precession photography. Precession photography utilizes a moving crystal-moving film device that allows for the mapping of an undistorted record of a reciprocal lattice plane onto film. The X-ray source which was used in this thesis for this initial characterization is a sealed tube X-ray source.

X-rays in the useful range for crystallography are usually produced by bombarding a metal target, typically copper, with electrons produced from a heated filament and accelerated by an electric field. The collision of high energy electrons with an electron in a low-lying orbital of the target causes the target electron to be displaced. An electron from a higher orbital drops into the now vacant orbital and the excess energy is emitted as an X-ray photon. In the case of copper, electrons from the K shell are displaced and replaced by either an L shell ($L \rightarrow K$, K_{α}) or an M shell electron ($M \rightarrow K$, K_{β}). Typically the K_{β} radiation is removed by a nickel filter, producing a near monochromatic K_{α} radiation with wavelength = 1.5418 Å. There are three X-ray sources used in macromolecular crystallography; sealed X-ray tubes, rotating anodes and particle accelerators. X-ray tubes and rotating anodes are more common. Output from sealed tubes, however, is limited by the amount of heat that can be dissipated by the circulating water used to cool the target. Higher X-ray output is achieved by the rotating anode in which the target is a rapidly rotating copper disk. This allows for much greater heat dissipation by spreading the electron bombardment over a much larger area. As a result, rotating anodes are more than 10 times as powerful as fixed anode tubes. Particle accelerators do not use a copper target to generate X-rays but instead utilize rapidly circulating electrons or positrons. Charged particles, such as these, emit energy called synchrotron radiation when forced into curved trajectories and in accelerators the energy is emitted as X-rays. These systems provide the most powerful X-rays and allow for multiple wavelength selection.

In precession photography, the X-rays strike the crystal which has been mounted on a goniometer. Beyond the crystal are an annular-screen holder and a film holder. The machinery moves the crystal, screen and film in a precessing motion about the X-ray beam. In order to obtain a picture of a reciprocal lattice plane, from which symmetry, unit cell dimensions and the space group can be determined, the crystal must be aligned such that one of the real space axes is parallel to the beam. In this orientation, a zero-level reciprocal lattice plane will be tangent to the sphere of reflection at the origin. For example, in an orthogonal system such as a PurR-hypoxanthine-*purF* operator crystal, when the c axis (hence also the c^* axis in this system) is parallel to the beam, the $hk0$ plane is tangent to the sphere of reflection at the origin. If the crystal is now precessed about an angle, μ , and the film precessed in the same manner, the zero-level reflections of the $hk0$ plane will be

recorded onto the film. Because nets other than the zero-level net will pass through the sphere of reflection, a screen is placed on the screen holder, in front of the film to prevent recording these nets on film. If the crystal to film distance (CF) is accurately known, the separation of the reflections on the film can be converted into the corresponding separations in the reciprocal lattice by similar triangles as shown in Figure 15.

$$RF/CF = PO/CO$$

In this equation, the distance in reciprocal lattice units of reciprocal lattice point P from the origin O is the ratio of the distance of corresponding reflection, R, from the film center, F, to the crystal (CF) times the wavelength, (CO). This relationship can be utilized to obtain the crystal's unit cell dimensions. The space group is determined by examining the recorded reciprocal lattice plane and noting any symmetry and systematic absences. There are 65 possible space groups for macromolecules with symmetry elements being limited to 2, 3, 4 and 6 fold axes along with screw axes and centering. Systematic absences are indicative of screw axes and centering (130). For example, in the CBD $P2_1$ crystals, there is a screw axis along **b**. This is indicated by the absence of reflections when *k* is odd in the $0k0$ reflections. This can be understood mathematically by considering the special form of the structure factor equation for equivalent positions for the $P2_1$ case where *x*, *y*, *z* and $-x, y+1/2, -z$ are equal:

$$F_{hkl} = fe^{2\pi i(hx + ky + lz)} + fe^{-2\pi i(hx - ky - k/2 + lz)}$$

$$F_{hkl} = fe^{2\pi iky} [e^{2\pi i(kx + lz)} + e^{-2\pi i(hx + lz)} e^{\pi i}]$$

Thus, in case of $0k0$ reflections:

$$F_{0k0} = fe^{2\pi iky} (1 + e^{\pi ik})$$

$$F_{0k0} = 2fe^{2\pi iky} \quad \text{for } k = 2n \text{ (even reflections)}$$

$$F_{0k0} = 0 \quad \text{for } k = 2n + 1 \text{ (odd reflections)}$$

A final parameter that can be well estimated from the unit cell volume is the number of molecules in the unit cell (*Z*). This estimation is made by a method proposed by Dr. B.W. Matthews (140). He found that for most protein crystals, the ratio of the unit cell volume to the protein molecular weight is between 1.7 and 3.5 Å³/Da. This number is called *V_m*. This number helps determine how many molecules are contained within the

asymmetric unit. Also, from the V_m , the volume fraction of solvent in the unit cell can be estimated as follows:

$$\begin{aligned} V_{\text{protein}} &= (ZM_r \times \text{specific volume of protein})/N/V_m ZM_r \\ &= \text{specific volume of protein in cm}^3/\text{g}/V_m \text{ in } \text{\AA}^3/\text{Da}\cdot N \text{ mol}^{-1} \end{aligned}$$

where N is Avogadro's number. The specific volume of a protein molecule is approximately $0.74 \text{ cm}^3/\text{g}$ and thus:

$$V_{\text{protein}} = 1.23/V_m \quad \text{and} \quad V_{\text{solvent}} = 1 - (1.23/V_m)$$

Once the unit cell is characterized, an efficient data collection strategy can be devised. The goal of data collection is a set of consistently measured, indexed intensities for as many reflections as possible. After the data are collected they must be processed in a procedure called data reduction which improves the consistency of the data and implements certain corrections. This ultimately provides the structure factor amplitude for each reflection.

Data collection for the projects described in this thesis were carried out on the SDMS area detector using a RIGAKU RU200-H rotating anode generator operating at 40 kV and 150 mA. This system utilizes a four circle goniostat which is a system of movable circles upon which the goniometer containing the crystal is placed. The goniostat allows rotation of the goniometer (angle phi, ϕ), movement of the goniometer around a circle centered on the X-ray beam (angle chi, χ) and rotation of the χ circle around an axis perpendicular to the beam (angle omega, ω). The position of the detector with respect to the beam is denoted by the angle 2θ . This arrangement of moveable angles allows the crystal to be rotated to bring any reciprocal lattice plane that lies within the limiting sphere, as determined by λ , into the plane of the detector and into contact with the sphere of reflection. The position of the detectors determines the maximum resolution that can be obtained. During data collection the detector is locked into place and only the crystal orientation is changed by stepped rotation about the ω axis. The design of the SDMS area detector is based on a gas-filled ionization chamber with two cathodes and an anode. The cathodes and anode consist of parallel wires which act as position sensitive detectors of X-ray ionization (141).

Before data collection can begin an orientation matrix must be determined for each crystal. This is obtained by locating and correctly indexing a small group of reflections in reciprocal space. To do this, the crystal cell parameters are also needed. The alignment

angles, which are determined, define a preliminary matrix from which the angular settings of other reflections can be calculated. After final refinement of the alignment angles and machine parameters, which is carried out by the SDMS software (142), data collection can begin.

The intensity data collected are converted into structure factor amplitudes, $|F_{\text{obs}}|$, through data reduction. The relationship between $|F_{\text{obs}}|$ and the collected intensity depends on certain geometric factors and the data collection apparatus used as follows:

$$|F_{\text{hkl}}| = (KI_{\text{hkl}}/Lp)^{1/2}$$

Where p is the polarization factor, L is the Lorentz factor and K is normally a constant for a given set of measurements. The polarization term, p , is equal to $(1 + \cos^2 2\theta/2)$ and is a function of 2θ and arises because of the nature of the X-ray beam and the manner in which its reflection efficiency varies with reflection angle. The Lorentz factor, L , accounts for the rate at which a reflection passes through Ewald's sphere. This correction is large for reflections which are near the rotation axis and remain in diffracting condition for longer periods of time.

Once the data are processed by the SDMS software, the overall error can be estimated by comparing symmetry related reflections, which would ideally be identical. Thus, for n independent reflections and i observations of a given reflection:

$$R_{\text{sym}}(I) = \frac{\sum_{\text{hkl}} \sum_i^n |I_i(\text{hkl}) - I(\text{hkl})|}{\sum_{\text{hkl}} \sum_i^n I_i(\text{hkl})}$$

where $I(\text{hkl})$ is the average intensity of the reflection, hkl .

Once reflection amplitudes have been determined, what remains to be established are the phases for each reflection. The two major methods used in the determination of phases in macromolecular X-ray crystallography are multiple isomorphous replacement (MIR) and molecular replacement (MR). Both methods utilize a Fourier series called the Patterson function, $P(u,v,w)$. The Patterson function is useful at this point because it is

phase independent. Each term in the Patterson function contains the square of the structure factor amplitude and is thus, proportional to each measured intensity. The general form of the Patterson function is as follows:

$$P(u,v,w) = 1/V \sum_h \sum_k \sum_l |F_{hkl}|^2 \exp(-2\pi i(hu + kv + lw))$$

Unlike a contour map of the electron density, $\rho(x,y,z)$, which displays areas of high electron density, a Patterson map displays peaks corresponding to vectors between atoms. The methods of MIR and MR utilize different forms of the Patterson function.

Of the two methods, MR is generally the most rapid. In this method, molecular coordinates presumed to be identical or nearly identical to the unknown structure are required. These coordinates constitute a probe model or search model. A rough rule of thumb that has been suggested is that the sequence of the probe model and the sequence of the unknown should be at least 50% identical for this technique to be straightforward (132).

A significant problem can result using probes with low homology and corresponding low structural homology because if a solution is obtained, the phases will be poor estimates of the true phases. Because the new structure is highly biased towards the probe structure it may be difficult to correct such a structure. This problem is made clear by considering that what one is essentially doing in MR is "plugging in" the probe model phases with the native structure factor amplitudes as follows:

$$\rho(x,y,z) = 1/V \sum_h \sum_k \sum_l |F_{hkl,obs}| \exp(-2\pi i(hx + ky + lz) + i\alpha'_{hkl}^{model})$$

Indeed, it has been shown that phases are much more critical in the determination of $\rho(x,y,z)$ as incorrect structure factor amplitudes combined with correct phases may lead to an interpretable map, whereas the correct structure factor amplitudes combined with incorrect phases leads to an uninterpretable map (143). Another caution in using MR, which is underscored in this thesis, is that even a probe model which is identical, i.e., the same protein, may not work in MR if it is in a significantly different conformation, for example the open versus the closed form of the CBD of PurR. Judicious breaking of the probe model into smaller rigid pieces or subdomains can be helpful in such cases.

Besides a probe model, the only other requirement for structure determination by MR is a native data set containing reliable data in the 10 Å to 3 Å range. The general idea behind MR is to fit the probe structure into the unit cell of the unknown by carrying out two

three-dimensional searches, in which the first search is to find the three rotation angles, α , β and γ , and the second to determine the three x , y , z translation components. The rotation angles are determined first by the rotation function, $R(\alpha, \beta, \gamma)$. Although there are several specific methods that have been formulated to determine R , all originate from the Patterson function. In the determination of the rotation function one is looking at the angular relationship between the probe model and the unknown. Therefore, what is compared are the Patterson vectors which are found close to the origin and which lie within a radius equal to the dimensions of the probe model. These vectors are called the self-Patterson vectors and they correspond to intramolecular vectors or vectors between atoms within a molecule. The basis of the rotation function is that if the self-Patterson of the probe is rotated onto the self-Patterson of the unknown, maximum overlap will occur when they are superimposed. The rotation function, R , may therefore, be defined as (144):

$$R(\alpha, \beta, \gamma) = \int_U P(\mathbf{u})P_r(\mathbf{u}_r)\delta\mathbf{u}$$

U is the volume in the Patterson map in which the self vectors are found, $P(\mathbf{u})$ is the self-Patterson of the unrotated unknown molecule (\mathbf{u} represents u, v, w) and $P_r(\mathbf{u}_r)$ is the rotated self-Patterson of the probe. $P(\mathbf{u})$ and $P_r(\mathbf{u}_r)$ can be expanded as follows:

$$P(\mathbf{u}) = 1/V \sum_{\mathbf{h}} |F(\mathbf{h})|^2 \exp(-2\pi i \mathbf{h} \cdot \mathbf{u})$$

$$P_r(\mathbf{u}_r) = 1/V \sum_{\mathbf{h}'} |F(\mathbf{h}')|^2 \exp(-2\pi i \mathbf{h}' \cdot \mathbf{u}_r)$$

where \mathbf{h} or \mathbf{h}' is used to represent a given reflection, hkl . Applying a rotation, represented by matrix $[C]$ to Patterson positions \mathbf{u} , (u, v, w), results in \mathbf{u}_r , that is $\mathbf{u}_r = [C]\mathbf{u}$, therefore :

$$P_r(\mathbf{u}_r) = 1/V \sum_{\mathbf{h}'} |F(\mathbf{h}')|^2 \exp(-2\pi i \mathbf{h}' \cdot [C]\mathbf{u})$$

Because $\mathbf{h}'[C] = [C^{-1}] \mathbf{h}'$:

$$P_r(\mathbf{u}_r) = 1/V \sum_{\mathbf{h}'} |F(\mathbf{h}')|^2 \exp(-2\pi i [C^{-1}] \mathbf{h}' \cdot \mathbf{u})$$

\mathbf{h}'

$P(\mathbf{u})$ and $P_r(\mathbf{u}_r)$ can now be superimposed and for every position, \mathbf{u} , within U , $P(\mathbf{u})P_r(\mathbf{u}_r)$ can be calculated to obtain $R(\alpha, \beta, \gamma)$:

$$R(\alpha, \beta, \gamma) = \frac{1}{V^2} \sum_{\mathbf{h}} \sum_{\mathbf{h}'} |F(\mathbf{h})|^2 |F(\mathbf{h}')|^2 \int_U \exp(-2\pi i [\mathbf{h} + [\mathbf{C}^{-1}] \mathbf{h}'] \mathbf{u}) \delta \mathbf{u}$$

The integral in the above expression may be simplified by making the assumption that $U(\rho(x, y, z) = \rho)$ inside the body of U and $U(\rho(x, y, z) = 0)$ outside, where U is at the origin of the unit cell. The expression for R can then be simplified to:

$$R(\alpha, \beta, \gamma) = \frac{U}{V^3} \sum_{\mathbf{h}} \sum_{\mathbf{h}'} |F(\mathbf{h})|^2 |F(\mathbf{h}')|^2 G[-(\mathbf{h} + [\mathbf{C}^{-1}] \mathbf{h}')]$$

where

$$G = \frac{3(\sin 2\pi[(\mathbf{h} + [\mathbf{C}^{-1}] \mathbf{h}')r] - 2\pi[(\mathbf{h} + [\mathbf{C}^{-1}] \mathbf{h}')r] \cos 2\pi[(\mathbf{h} + [\mathbf{C}^{-1}] \mathbf{h}')r])}{(2\pi[(\mathbf{h} + [\mathbf{C}^{-1}] \mathbf{h}')r])^3}$$

and r is the probe molecular radius. This interference function, G , has a maximum value when $[\mathbf{C}^{-1}] \mathbf{h}' = -\mathbf{h}$ which limits the number of terms that need to be calculated.

Several conventions for the rotation angle directions, names, signs and origins exist. However, the usual system utilizes the Eulerian angles α , β and γ . As shown in Figure 16, in this system a rotation by the angle α around the z orthogonal axis is followed by a rotation around the new x axis by the angle β and finally, a rotation by γ about the new z axis. This describes the rotation function as implemented by Rossmann and Blow. One problem with this method of rotation is that for steps of a given angular increment, the fineness of the search is not constant. For example, when $\beta = 0$ or $\beta = 180^\circ$, the search consists of a closely packed set of orientations around α (or γ), but when $\beta = 90^\circ$, the difference in the orientation corresponding to each angular setting increases. Lattman modified the Rossmann and Blow method to "fix" this problem by defining the "distance" between two orientations and redefining the angular motions to ensure uniform sampling (145). In this method the Eulerian angles are redefined as θ_+ , θ_2 and θ_- where:

$$\theta_+ = \alpha + \gamma, \theta_2 = \beta \text{ and } \theta_- = \alpha - \gamma$$

The major problem with this method and the Rossmann and Blow method is the lengthy computational time. In 1971 Crowther outlined a new approach that is still based on the Patterson function however, in this method the Patterson functions are expanded as spherical harmonics, which casts the rotation function in the more amenable form of a spherical function (146). This function, which has been called the Crowther fast rotation function, is up to 100 times faster than the Lattman function and can be expressed as:

$$R(\Omega) = \sum_{lmm'} c_{lmm'} D_{m'n}^l(\Omega)$$

Where $c_{lmm'}$ refers to the Patterson coefficients. $D_{m'n}^l$ are the coefficients which contain the rotational aspect of the problem and refer to rotations of spherical harmonics. Calculation of the fast rotation function results in values for α and γ for each β . A best strategy in the calculation of the rotation function is to first obtain a rough set of values for the rotation angles using the Crowther fast function and then carry out a fine search around the main peaks using the Lattman function. This is the strategy that was employed in the MR determinations in this thesis.

After the probe has been correctly oriented using the rotation function, the translation function is employed. There are several different methods that have been developed to determine the translation function. The most straightforward method, again, involves Patterson functions. Specifically, cross-Patterson vectors, which are intermolecular vectors or vectors between two different molecules, are used. A translation function is calculated that gives the correlation between a set of model cross-Patterson vectors and the observed Patterson function (147). The cross-Patterson vectors are derived from vectors between atoms from two molecules related by a crystallographic operation. The final form of the translation function is a Fourier summation given as:

$$T(\mathbf{t}) = \sum_{\mathbf{h}} |F_{\text{obs}}(\mathbf{h})|^2 F_m(\mathbf{h}) F_m^*(\mathbf{h}[C]) \exp(-2\pi i[\mathbf{h}\mathbf{t}])$$

Where \mathbf{t} is the intermolecular vector, \mathbf{h} , again, represents hkl reflections, $|F_{\text{obs}}(\mathbf{h})|$ is the observed structure factor amplitude, $F_m(\mathbf{h})$ is the calculated structure factor from the model and $[C]$ is a crystallographic rotation. The goal is to determine the value of \mathbf{t} which gives the maximum overlap in the function.

Another method used for the determination of the translation function is the calculation of the R-factor, or correlation coefficient, as a function of the molecular position (148). In this, basically trial-and-error procedure, the probe is moved through the asymmetric unit and the corresponding structure factor amplitudes are calculated and compared with the observed structure factor amplitudes in the following manner:

$$R = \frac{\sum_{hkl} ||F_{obs}| - k |F_{calc}||}{\sum_{hkl} |F_{obs}|}$$

where k is a scale factor for intensities.

Another translation method is the packing function (148). This function attempts to determine allowed positions of the oriented molecule in the unknown cell by identifying translations that result in a minimum of intermolecular penetration. The basic idea is to translate systematically the oriented molecule throughout the unknown cell and calculate the number of close intermolecular contacts at each step. The utility of this function is that it delimits regions of space that the probe model can occupy, on the basis of the packing considerations of the given cell type. Thus, this method complements nicely other techniques, such as the R-factor correlation and the cross-Patterson translation function, in that it can be used to confirm a result obtained from one of these methods.

Once solutions for the rotation and translation functions have been obtained, they are applied to the probe model, producing a starting model. In this starting model, usually the relative positions of the atoms are quite good, however, their absolute positions (within the unit cell) will still be in error by a great deal. This occurs because the rotation and translation functions may have been in error by several degrees and, perhaps, an angstrom, respectively. Therefore, the first step in the refinement of a new model derived from MR should be rigid body refinement. In this technique the model is moved as a single rigid unit in order to optimize its placement within the unit cell. Refinement will be described in greater detail at the end of this section.

The technique utilized most often in the structure determination of a new macromolecule is multiple isomorphous replacement (MIR). The basis for this technique is that the incorporation of a few heavy atoms into the crystal leads to measureable intensity differences, which can then be used to obtain initial phase information. The requirement that the crystal remains isomorphous, meaning that there are no substantial changes in the underlying protein structure upon heavy atom incorporation, is critical.

MIR requires the following steps:

1. The preparation of several isomorphous heavy atom derivatives (at least two are required).
2. The collection of X-ray intensity data for the derivative crystals as well as the native crystal.
3. The determination of heavy atom positions through difference Patterson functions.
4. Refinement of heavy atom parameters and calculation of protein phase angles.
5. Calculation of electron density map.

The first step, the preparation of heavy atom derivatives, is basically empirical. Dozens of heavy atom reagents can be tried. Some indication of whether a given reagent may be successful can be gained by knowledge of whether heavy atom reactive residues, such as cysteines and methionine, are present in the protein which has been crystallized. Crystals of protein-DNA complexes present a potential advantage in this method as iodinated or brominated oligonucleotides may be substituted for the original oligonucleotide. If isomorphous crystals can be grown with these substituted oligonucleotides, not only does one have a heavy atom derivative of known occupancy but, because the location of the substitution will be known, the initial chain tracing of the DNA is aided substantially. In protein crystals, however, this is not an option and heavy atom preparation involves soaking the protein crystal in a solution of the heavy atom reagent and then looking for intensity differences, initially by comparing a "derivative" precession photograph with a native precession photograph taken of the same reciprocal lattice zone. When the soaking method is used, variable soaking times and heavy atom reagent concentrations are tried in order to optimize the derivatization.

Once a potential derivative has been identified by precession photography, intensity data can be collected. The data sets obtained from potential heavy atom derivatives must be carefully scaled to the native data set. In the work in this thesis, the program Loscale was used for this purpose. The potential usefulness of a heavy atom derivative can be estimated by calculating the derivative R-factor (Isomorphous R):

$$R_{\text{deriv}} = \frac{\sum_{hkl} ||F_{\text{PH}}(hkl)| - |F_{\text{p}}(hkl)||}{\sum_{hkl} |F_{\text{p}}(hkl)|}$$

where F_{PH} represents the derivative structure factor amplitude and F_{p} , the native structure factor amplitude. Successful derivatives generally have R_{deriv} in the range from 0.1 to 0.25. Once the derivative intensity data are scaled to the native intensity data, difference Patterson maps may be calculated by:

$$P(u,v,w) = 1/V \sum_{hkl} (|F_{PH}| - |F_P|)^2 \cos(2\pi(hu + kv + lw))$$

Because the diffraction contributions of all the electrons to a reflection are additive, the difference in amplitudes, $|F_{PH}| - |F_P|$, represents the contribution of the heavy atom alone. Therefore, the difference Patterson map is simplified to vectors between heavy atoms.

Once one derivative has been located from difference Pattersons, it is necessary to place all subsequent derivatives on the same origin. This can be accomplished by utilizing phases based on the first derivative, as will be described, and calculating a cross Fourier, which is one form of a difference Fourier, using:

$$\rho(x,y,z) = 1/V \sum_{\mathbf{h}} (|F_{PH(2)}| - |F_P|) \exp(-2\pi i(\mathbf{h}x + \mathbf{h}y + \mathbf{h}z) + i\alpha_P(\mathbf{h}))$$

In the above equation, $|F_{PH(2)}|$ are structure factor amplitudes for the second derivative and $\alpha_P(\mathbf{h})$ represents phases calculated from data based on the first derivative. After the positions of the heavy atom derivatives have all been located and placed on the same origin, the x, y, z coordinates and occupancy parameters of each are refined. Refinement by correlation of origin-removed Pattersons is one method of heavy atom refinement that is implemented in the program, Heavy, which was utilized in this thesis in the initial refinement. After heavy atom refinement has converged, phases can be calculated because, at this point, in addition to knowing the structure factor amplitudes for the derivative, $|F_{PH}|$, and protein, $|F_P|$, the heavy atom structure factors, F_H , can be calculated with:

$$F_H = \sum_{j=1}^n f_j \exp(-2\pi i(\mathbf{h}x_j + \mathbf{h}y_j + \mathbf{h}z_j))$$

Where f_j is the scattering factor for each heavy atom and the sum is taken for each heavy atom of a given type within the unit cell. With this information the protein phases can be obtained from the Harker construction as represented in Figure 17B (129). The principle of the Harker construction is as follows: First draw a circle with radius $|F_P|$. From the center of this circle draw vector $-F_H$, which as a vector has both length and direction. Next, draw a second circle of radius $|F_{PH}|$ with its center at the end of $-F_H$. The intersection of the two circles corresponds to two equally probable phase angles because for both phase angles a triangle of $F_{PH} = F_P + F_H$ closes exactly. Because of this ambiguity

of two equally probable phases for acentric reflections, a second derivative is required to resolve which phase angle is correct.

In the determination of phases, the treatment of errors is critical as, in practice, large errors arise from the inherent inaccuracies in intensity measurements and the lack of perfect isomorphism. As a result, the circles in the Harker construction for several derivatives do not normally intersect at clear points. The calculation of errors can be used to further refine heavy atom parameters and, ultimately, they are essential in calculating a weighting function that weights individual phases according to the probability that they are correct. The method that is used to minimize the errors resulting from inaccurate intensities and lack of perfect isomorphism, is called the "lack of closure" method (149). This method requires preliminary phase values. With these values the vector triangle $F_{PH} = F_P + F_H$ can be drawn. Generally it is found that $|F_{PH}|_{calc}$ will be either too short or too long to reach F_H given the initial phase (Fig 17A). This difference constitutes the "lack of closure error", ϵ . The goal of the refinement is to minimize this error. For each reflection, $|F_{PH}|_{calc}$ is obtained from the cosine rule as follows:

$$|F_{PH}|_{calc} = \{ |F_P|^2 + |F_H|^2 + 2|F_P||F_H| \cos(\alpha_H - \alpha_P) \}^{1/2}$$

For a given heavy atom derivative, j , ϵ_j is defined as:

$$\epsilon_j = \{ k_j (|F_{PH}|_{obs})_j - (|F_{PH}|_{calc})_j \}$$

Where k_j is a scaling factor. The actual function which is minimized is:

$$E_j = \sum_{hkl} m_{hkl} \epsilon_j (hkl)^2$$

Where m_{hkl} is the weighting function mentioned previously and is called the figure of merit. The values which are optimized in this minimization are the heavy atom coordinates and occupancy. These values determine the length and direction of F_H and therefore, its endpoint. Thus, this represents a further refinement of the heavy atom parameters. After a few cycles of such a refinement scheme, new and improved phase angles can be determined by the Harker construction.

As noted, phase angles must be determined for all the reflections. For phase angles in which ϵ_j is small, there is a higher probability that the phase is correct. A

Gaussian distribution is assumed for ϵ_j and for every reflection of each derivative as follows:

$$P(\alpha) = P(\epsilon) = N \exp[-\epsilon^2(\alpha)/2E^2]$$

Where N is a normalization factor originating from the fact that each phase is between 0 and 2π . Given a single heavy atom derivative P(α) can be calculated as a function of α . The result is a curve with two equally high peaks, as expected using a single derivative. The total probability of each reflection can be obtained by multiplying separate derivative probabilities, indicated by j, as follows:

$$P(\alpha) = \prod_{j=1}^n P_j(\alpha) = N' \exp[-\sum_j \epsilon_j^2(\alpha)/2E_j^2]$$

Therefore, electron density calculated with the phase angles $\alpha(hkl)$ is equal to $P_{hkl}(\alpha)$. However, this map is not necessarily the best map, which would be defined as the map with the minimum mean square error due to errors in the phase angle. The reason is that the function $P_{hkl}(\alpha)$ does not always have a single maximum. The best estimate of the structure factor, $F_{hkl}(\text{best})$ is given by the least squares criterion:

$$Q = \int_{\alpha} [P_{hkl}(\alpha) |F_{hkl}| \exp[i\alpha] - F_{hkl}(\text{best})]^2 \delta\alpha$$

in which Q should be a minimum or $\delta Q / \delta F_{hkl}(\text{best}) = 0$. Taking this partial derivative results in:

$$F_{hkl}(\text{best}) = \int_{\alpha} [P_{hkl}(\alpha) |F_{hkl}| \exp[i\alpha]] \delta\alpha = |F_{hkl}| m$$

where

$$m = \int_{\alpha} [P_{hkl}(\alpha) \exp[i\alpha]] \delta\alpha \quad 0 \leq |m| \leq 1$$

This indicates that the best value of F_{hkl} can be obtained by taking the weighted average over the possible range of F_{hkl} s. $F_{hkl}(\text{best})$ points to the center of gravity of the probability distribution and can be calculated as follows:

$$F_{hkl}(\text{best}) = |F_{hkl}| m \exp[i\alpha(\text{best})]$$

where $m = |F_{hkl}(\text{best})|/|F_{hkl}|$ and is called the figure of merit. This is the weighting function mentioned previously. It can be shown that the figure of merit is the weighted mean of the cosine of error in the phase angle and thus, a figure of merit of 1 indicates 0 error in a given phase angle.

With $F_{hkl}(\text{best})$ calculated, an initial electron density map can be calculated. The next step is the interpretation of this map or "tracing the chain". This is usually carried out using an interactive computer graphics program such as FRODO (150). If a major part of the macromolecule can be identified and built into the density, refinement of the structure can proceed. Often, however, an initial MIR map is of insufficient quality to allow for unambiguous chain tracing. In such cases, phase improvement is often necessary. There are several methods to improve phases, including solvent flattening, averaging and histogram matching. The method which was instrumental in the structure determination of the PurR-hypoxanthine-*purF* operator complex was solvent flattening. This method is a fairly simple one that takes advantage of the large amount of solvent in macromolecular crystals. However, in order for this method to be utilized, the solvent fraction of the given crystal must be determined either by calculation, using V_m or, more accurately by crystal density measurements with xylene/carbon tetrachloride solutions or Ficoll. The general principle behind this method is that the electron density in the solvent region in macromolecular crystals generally has a very low and constant value due to its dynamic nature.

The procedure followed in solvent flattening (151) is to first calculate an electron density map and set all nonpositive values to zero. After Fourier transformation, the modified coefficients are used to calculate a new map in a back transform. If the solvent fraction is known, a level of density in the map can be chosen so that a certain percentage of density points are below it. This serves to define the molecular envelope. After the density points in the solvent region are set to zero and the map is transformed, new coefficients are once again calculated. The process continues until convergence is reached.

Another method to improve phases, also used in this thesis, is phase combination. This method is applicable when a partial model has been obtained, usually by MIR. To help locate the remaining structural elements, the phase information from the known part of

the model (partial model) can be combined with the MIR phases. To do this, a general and convenient way of combining phases would be useful. Such a method was proposed by Hendrickson and Lattman (152) in the form of:

$$P_s(\alpha) = N_s \exp[k_s + A_s \cos\alpha + B_s \sin\alpha + C_s \cos 2\alpha + D_s \sin 2\alpha]$$

Where N_s is a constant, $P_s(\alpha)$ is the probability for phase angle α from source s , k_s and the coefficients A_s , B_s , C_s and D_s , contain structure factor amplitudes but not phase angles. The calculation of the overall probability, $P_s(\alpha)$, including information from MIR and the partial model, is now much simplified to an addition of all k_s and the coefficients A_s , B_s , C_s and D_s by:

$$P(\alpha) = \prod_s P_s(\alpha) = N' \exp[\sum_s k_s + (\sum_s A_s) \cos\alpha + (\sum_s B_s) \sin\alpha + (\sum_s C_s) \cos 2\alpha + (\sum_s D_s) \sin 2\alpha]$$

If phase improvement is successful, the result is a "new" initial model which contains most or all of the structural elements. Such an initial model, however, usually represents a poor model in terms of its fit to the data. Although not as extreme, this is also true for models generated by MR. The "fit" of the model to the data is described in terms of the agreement index (R-factor) between the calculated and observed structure factors as:

$$R = \frac{\sum_{hkl} |F_{obs}| - |F_{calc}|}{\sum_{hkl} |F_{obs}|} \times 100$$

An R-factor of 59% would result for a random acentric structure. However, it is not uncommon for a starting MIR model to have an R-factor of 50% and still be correct in terms of the fold. In this case, it is generally the exact placement of the atoms which are in error.

Refinement is the last step in structure determination, whereby the model is adjusted to maximize its agreement with the data. The refinement strategy will depend on the errors in the model, which ultimately depends on how the model was obtained. As noted, when a structure is determined by MR, the relative atomic positions are good but their absolute positions may be in error by a great deal. The refinement strategy followed in such a case is to begin with a few cycles of rigid body refinement in which the molecule is rotated and translated as a single unit. This may necessitate breaking the molecule into

subdomains after the whole molecule has been optimally positioned. Rigid body refinement is usually carried out with fairly low resolution data, for example 10 Å to 3 Å. This is because the maximization is for the overall molecule, which depends on gross secondary and tertiary and possibly, quaternary structural features. These features are best reflected in data from this range. Once convergence is reached with rigid body refinement, positional or x, y, z and temperature factor or B refinement, can begin. Also, the data may, at this point, be extended to the resolution limit.

A model resulting from MIR, unlike one generated with MR, will not require rotational or translational adjustments, but will have many misplaced atoms. To fix such errors, large shifts in atomic position will be required. The best initial range of data to use in this case is, again, low resolution data. A good starting strategy is to do a round of loosening and then tightening of the geometry which will allow the atoms to find their preferred positions. This usually leads to large improvements in the R-factor, which can serve as a guide as to how the refinement is proceeding. Once the R-factor has converged, a new electron density map is calculated and another attempt is made to fix the model in real space via interactive computing. This is an iterative process and the data are added gradually until the resolution limit is reached.

The refinement techniques in macromolecular crystallography are based on the principle of least squares. Least squares is an iterative process by which the observations have fixed values and certain parameters are varied so that the calculated values approach the observations as closely as possible. Usually many cycles are needed before refinement converges to the final parameter set. The range of convergence is the maximum distance for the atoms to move to their final positions and thus, if they are too far away, there is a good chance that they may become trapped in a local minimum that is not the true minimum. The theoretically derived radius of convergence is $d_{\min}/4$ where d_{\min} is the lattice plane spacing of the reflection of highest resolution used in refinement. Therefore, inclusion of data from higher resolution, although providing more information, decreases the radius of convergence which is why the initial refinement of an MIR model should start with moderate resolution data.

In practice, refinement is carried out in cycles in which real space intervention is carried out to fix the model followed by reciprocal space refinement. The name reciprocal space is given to such techniques because they attempt to maximize the agreement between the data and the model in the terms of structure factor amplitudes and not electron density. Reciprocal space is where the least squares fitting takes place and the function minimized is:

$$Q = \sum_{hkl} w(hkl) (|F_{\text{obs}}| - |F_{\text{calc}}|)^2$$

where $w(hkl)$ is the weighting factor, usually $1/\sigma^2$, where σ is the standard deviation from multiple measurements of $|F_{obs}|$. Taking the partial derivatives of the right side of the above equation with respect to each of the variables and setting them equal to zero results in a set of normal equations summarized in matrix form by $[A] \times [\epsilon] = [b]$. $[A]$ is the element varied, $[b]$ is the known gradient vector and $[\epsilon]$ is the unknown vector containing parameter shifts necessary for minimization. The shifts can be determined by $[\epsilon] = [A^{-1}] \times [b]$. The parameters of $|F_{calc}|$ which are varied can be noted by expanding $|F_{calc}|$ as follows:

$$F_{calc} = G \sum n_j f_j \exp 2\pi i (hx_j + ky_j + lz_j) \times \exp -[B_j (\sin \theta) / \lambda]^2$$

Here G is a scale factor used to place all F_{calc} s on a common numerical scale, n_j is the occupancy of atom j , f_j is its scattering factor, x_j , y_j and $l z_j$ are its coordinates and B_j is the temperature factor of the j^{th} atom.

To obtain a higher ratio of observations to unknown parameters, additional "observations" are incorporated into the refinement process. In the TNT refinement program (153), several stereochemical terms are added to the crystallographic terms. These restraints are derived from stereochemical data from very high resolution (therefore accurate), small molecule structures and can be included with the crystallographic terms as follows:

$$Q = \sum_{hkl} w(hkl) [|F_{obs}| - |F_{calc}|]^2 \quad (1)$$

$$+ \sum_{\text{dist}, j} w_D(j) (d_j^{\text{ideal}} - d_j^{\text{model}})^2 \quad (2)$$

$$+ \sum_{\text{planes}, k} \sum_{\text{coplanarity}, i} w_P(i, k) (m_k r_{i, k} - d_k)^2 \quad (3)$$

$$+ \sum_{\text{chiral centers}, l} w_C(l) (V_l^{\text{ideal}} - V_l^{\text{model}})^2 \quad (4)$$

$$+ \sum_{\text{nonbond. contacts}, m} w_N(m) (d_m^{\text{min}} - d_m^{\text{model}})^4 \quad (5)$$

$$+ \sum_{\text{torsion angles}, t} w_T(t) (X_t^{\text{ideal}} - X_t^{\text{model}})^2 \quad (6)$$

Term (1) is the crystallographic term and the remaining terms are stereochemical restraints. By using restraints these terms are allowed to vary somewhat around standard values. Term (2) restrains the distance between atoms, defining bond lengths, bond angles and dihedral angles. Term (3) imposes the planarity of aromatic rings. Term (4) restrains the configuration to the correct enantiomer. Term (5) introduces restraints for nonbonded or van der Waals contacts, preventing close approach of atoms not connected by chemical bond. Finally, term (6) restrains torsion angles.

The last stages of refinement consist of alternating cycles of computed reciprocal space refinement and map fitting. At this stage, the model may be checked by omit maps, in which a region of the model is omitted from the refinement, several rounds of refinement carried out and a new map calculated based on the refined structure with the omitted regions. This removes bias from this region as it no longer contributes to the phases which are calculated. Omit maps were calculated for all structures described in this thesis to check their accuracy. In the final stages of refinement, ordered water molecules may also be located and added to the model.

ACKNOWLEDGMENTS:

The PurR and CBD protein used in this thesis were provided by Kang-Yell Choi, Fu Lu and Howard Zalkin. Richard Jones and his lab provided the cross-linked hemoglobins. Melinda Dixon determined the structure of the carbonmonoxy $\alpha_2\beta^1S^{82}\beta$ hemoglobin. CD studies (Manuscript #1) were the joint work of Maria A. Schumacher (myself) and Randy MacDonald from Hans Peter Bachinger's lab. All other work was done by Maria A. Schumacher.

REFERENCES

1. Matthews, C.K. and van Holde, K.E. (1990). *Biochemistry*. Benjamin/Cummings Publishing Company, Inc.
2. von Hippel, P.H., Yager, T.D., and Gill, S.C. (1992). Quantitative Aspects of the Transcription Cycle in *Escherichia coli*. In: *Transcriptional Regulation*. Vol. 1 (Eds. S. Mcknight and K. Yamamoto) pp.179-201, Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press.
3. Gross, C. Lonetto, M., and Losick, R. (1992). Bacterial Sigma Factors. In: *Transcriptional Regulation*. Vol. 1 (Eds. S. Mcknight and K. Yamamoto) pp.129-176, Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press.
4. Busby, S. and Ebright, R.H. (1994). Promoter Structure, Promoter Recognition, and Transcription Activation in Prokaryotes. *Cell* **79**, 743-746.
5. Darnell, J., Lodish, H., and Baltimore, D. (1990). *Molecular Cell Biology*. Scientific American Books, W.H. Freeman and Company, New York.
6. Jacob, F. (1966). Genetics of the bacterial cell. *Science* **152**, 1470-1478.
7. Monod, J. (1966). From enzymatic adaptation to allosteric transitions. *Science* **154**, 475-482.
8. Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**, 318-356.
9. Gilbert, W. and Müller-Hill, B. (1966). Isolation of the *lac* repressor. *Proc. Natl. Acad. Sci. USA* **56**, 1891-1898.
10. Gralla, J.D. (1992). *lac* Repressor. In: *Transcriptional Regulation*. Volume 2 (Eds.S. Mcknight and K. Yamamoto) pp.629-642, Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press.
11. Zubay, G., Schwartz, D., and Beckwith, J. (1970). Mechanism of activation of catabolite-sensitive genes: a positive control system. *Proc. Natl. Acad. Sci. USA* **66**, 104-110.
12. Weickert, M.J. and Adhya, S. (1992). A Family of Bacterial Regulators Homologous to Gal and Lac Repressors. *J. Biol. Chem.* **267**, 15869-15874.
13. Quijcho, F.A. (1991). Structures and function of periplasmic receptors for active transport and chemotaxis. *Curr. Opin. Struct. Biol.* **1**, 922-933.
14. Zalkin, H. and Nygaard, P. (1994?). Bioynthesis of Purine Nucleotides. In press
15. Zalkin, H. and Dixon, J.E. (1992). *De novo* Purine Nucleotide Biosynthesis. *Prog. Nucl. Acid Research* **42**, 259-286.

16. Gots, J.S., Benson, C.E., Jochimsen, B., and Koduri, K.R. (1977). Microbial models and regulatory elements in the control of purine metabolism. *Ciba Found. Symp.* **48**, 23-41.
17. Kilstrup, M., Meng, L.M., Neuhard, J., and Nygaard, P. (1989). Genetic Evidence for a Repressor of Synthesis of Cytosine Deaminase and Purine Biosynthesis Enzymes in *Escherichia coli*. *J. Bacteriol.* **171**, 2124-2127.
18. Rolfes, R. and Zalkin, H. (1988). Regulation of *Escherichia coli purF*. *J. Biol. Chem.* **263**, 19649-19652.
19. Rolfes, R. and Zalkin, H. (1988). *Escherichia coli* Gene *purR* Encoding a Repressor for Purine Nucleotide Synthesis. *J. Biol. Chem.* **263**, 19653-19661.
20. Rolfes, R. and Zalkin, H. (1990). Purification of the *Escherichia coli* Purine Regulon Repressor and Identification of Corepressors. *J. Bacteriol.* **172**, 5637-5642.
21. Rolfes, R. and Zalkin, H. (1990). Autoregulation of *Escherichia coli purR* Requires Two Control Sites Downstream of the Promoter. *J. Bacteriol.* **172**, 5758-5766.
22. Houlberg, U. and Jensen, K.F. (1983). Role of hypoxanthine and guanine in regulation of *Salmonella typhimurium pur* gene expression. *J. Bacteriol.* **153**, 837-845.
23. Choi, K.Y. and Zalkin, H. (1992). Structural characterization and corepressor binding of the *Escherichia coli* purine repressor. *J. Bacteriol.* **174**, 6207-6214.
24. Choi, K.Y., Lu, F., and Zalkin, H. (1994?). Mutagenesis of amino acid residues required for binding of corepressors to the purine repressor. *J. Biol. Chem.* **269**, 24066-24072.
25. He, B. and Zalkin, H. (1992). Repression of *Escherichia coli purB* is by a Transcriptional Roadblock Mechanism. *J. Bacteriol.* **174**, 7121-7127.
26. He, B. and Zalkin, H. (1994). Regulation of *Escherichia coli purA* by Purine Repressor, One Component of a Dual Control Mechanism. *J. Bacteriol.* **176**, 1009-1013.
27. He, B., Smith, J.M. and Zalkin, H. (1992). *Escherichia coli purB* Gene: Cloning, Nucleotide Sequence, and Regulation by *purR*. *J. Bacteriol.* **174**, 130-136.
28. Meng, L.M., Kilstrup, M., and Nygaard, P. (1990). Autoregulation of PurR repressor synthesis and involvement of *PurR* in the regulation of *purB*, *purC*, *purL*, *purMN* and *guaBA* expression in *Escherichia coli*. *Eur. J. Biochem.* **187**, 373-379.
29. Steiert, J.G, Rolfes, R., Zalkin, H., and Stauffer, G.V. (1990). Regulation of the *Escherichia coli glyA* Gene by the *purR* Gene Product. *J. Bacteriol.* **172**, 3799-3803.

30. Choi, K.Y. and Zalkin, H. (1990). Regulation of *Escherichia coli pyrC* by Purine Regulon Repressor Protein. *J. Bacteriol.* **172**, 3201-3207.
31. Wilson, H. and Turnbough, C.L. (1990). Role of the purine repressor in the regulation of pyrimidine gene expression in *Escherichia coli* K-12. *J. Bacteriol.* **172**, 3208-3213.
32. He, B., Choi, K.Y., and Zalkin, H. (1993). Regulation of *Escherichia coli glnB*, *prsA*, and *speA* by the Purine Repressor. *J. Bacteriol.* **175**, 3598-3606.
33. Pabo, C.O. and Sauer, R.T. (1992). Transcription Factors: Structural Families and Principles of DNA Recognition. *Ann. Rev. Biochem.* **61**, 1053-1095.
34. Wolberger, C. (1993). Transcription factor structure and DNA binding. *Curr. Opin. Struct. Biol.* **3**, 3-10.
35. Harrison, S.C. (1991). A Structural Taxonomy of DNA-binding Domains. *Nature* **353**, 715-719.
36. Brennan, R.G. and Matthews, B.W. (1989). The Helix-Turn-Helix DNA Binding Motif. *J. Biol. Chem.* **264**, 1903-1906.
37. Wright, P.E. (1994). POU domains and homeodomains. *Curr. Opin. Struct. Biol.* **4**, 22-27.
38. Brennan, R.G. (1993). The Winged-Helix DNA-Binding Motif: Another Helix-Turn-Helix Takeoff. *Cell* **74**, 773-776.
39. Scmiedeskamp, M. and Kleivit, R.E. (1994). Zinc finger diversity. *Curr. Opin. Struct. Biol.* **4**, 28-35.
40. Kaptein, R. (1992). Zinc Finger Structures. *Curr. Opin. Struct. Biol.* **2**, 109-115.
41. Ellenberger, T. (1994). Getting a grip on DNA recognition: structures of the basic region leucine zipper, and the basic region helix-loop-helix DNA-binding domains. *Curr. Opin. Struct. Biol.* **4**, 12-21.
42. Raumann, B.E., Bronwen, M.B., and Sauer, R.T. (1994). Major groove recognition by β -sheets: the ribbon-helix-helix family of gene regulatory proteins. *Curr. Opin. Struct. Biol.* **4**, 36-43.
43. Gourisankar, G., Van Duyne, G., Ghosh, S., and Sigler, P.B. (1995). Structure of NF- κ B p50 homodimer bound to a κ B site. *Nature* **373**, 303-310.
44. Müller, C.W., Rey, F.A., Sodeoka, M., Verdine, G.L., and Harrison, S.C. (1995). Structure of the NF- κ B p50 homodimer bound to DNA. *Nature* **373**, 311-317.
45. Cho, Y., Svetlana, G., Jeffrey, P.D., and Pavletich, N.P. (1994). Crystal Structure of a p53 Tumor Suppressor-DNA Complex: Understanding Tumorigenic Mutations. *Science* **265**, 346-355.

59. Billeter, M., Qian, Y.Q., Otting, G., Müller, M., Gehring, W.J., and Wüthrich, K. (1993). Determination of the Nuclear Magnetic Resonance Solution Structure of an Antennapedia Homeodomain DNA Complex. *J.Mol. Biol.* **234**, 1084-1097.
60. Kissinger, C.R., Liu, B., Martin-Bianche, E., Kornberg, T.B, and Pabo, C.O. (1990). Crystal Structure of an Engrailed Homeodomain-DNA Complex at 2.8 Å Resolution: a Framework for Understanding Homeodomain-DNA Interactions. *Cell* **63**, 579-590.
61. Wolberger, C., Vershon, A.K., Liu, B. Johnson, A.D., and Pabo, C.O. (1991). Crystal Structure of a MAT α 2 Homeodomain-Operator Complex Suggests a General Model for Homeodomain-DNA Interactions. *Cell* **67**, 517-528.
62. Klemm, J.D., Rould, M.A., Aurora, R., Herr, W., and Pabo, C.O. (1994). Crystal Structure of the Oct-1 POU Domain Bound to an Octamer Site: Recognition with Tethered DNA-Binding Modules. *Cell* **77**, 21-32.
63. Letting, B., DeFrancesco, R., Tomei, L., Cortese, R., Otting, G., and Wüthrich, K. (1993). The Three-Dimensional NMR-Solution Structure of the Polypeptide Fragment 195-286 of the LFB1/HNF1 Transcription Factor from Rat Liver Comprises a Non-Classical Homeodomain. *EMBO J.* **12**, 1797-1803.
64. Ceska, T.A., Lamers, M., Monaci, P., Nicosia, A., Cortese, R., and Suck, D. (1993). X-ray structure of an Atypical Homeodomain Present in the Rat Liver Transcription Factor LFB1/HNF1 and Implications for DNA Binding. *EMBO J.* **12**, 1805-1810.
65. Assa-Munt, N., Mortishire-Smith, R.J., Aurora, R., Herr, W., and Wright, P.E. (1993). The Solution Structure of the Oct-1 POU-Specific Domain Reveals a Striking Similarity to the Bacteriophage λ Repressor DNA-Binding Domain. *Cell* **73**, 193-205.
66. Lai, E., Clark, K.L., Burley, S.K, and Darnell, J.E. (1993). HNF-3/forkhead or 'Winged Helix' Proteins: A New Family of Transcription Factors of Diverse Biological Function. *Proc. Natl. Acad. Sci. USA* **90**, 10421-10423.
67. Clark, K.L., Halay, E.D., Lai, E., and Burley, S.K. (1993). Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature* **364**, 412-420.
68. Ramakrishnan, V., Finch, J.J., Graziano, V., Lee, P.L., and Sweet, R.M. (1993). Crystal structure of Globular Domain of Histone H5 and its Implications for Nucleosome Binding. *Nature* **362**, 219-223.

69. Liang, H., Mao, X., Olejniczak, E.T., Nettesheim, D.G., Yu, L., Meadows, R.P., Thompson, C.B., and Fesik, S.W. (1994). Solution structure of the *ets* domain of *Fli-1* when bound to DNA. *Nature Struct. Biol.* **1**, 871-876.
70. Clubb, R.T, Omichinski, J.G., Savilahti, H., Mizuuchi, K., Gronenborn, A.M., Clore, G.M. (1994). A novel class of winged helix-turn-helix protein: the DNA-binding domain of Mu transposase. *Structure* **15**, 1041-1048.
71. Harrison, C.J., Bohm, A.A., and Nelson, H.C.M. (1994). Crystal structure of the DNA binding domain of the heat shock transcription factor. *Science* **263**, 224-227.
72. Wilson, K.P., Shewchuk, L.M., Brennan, R.G., Otsuka, A.J, and Matthews, B.W. (1992). *Escherichia coli* biotin holoenzyme synthetase/*bio* repressor crystal structure delineates the biotin- and DNA-binding domains. *Proc. Natl. Acad. Sci.* **89**, 9257-9261.
73. Ogata, K., Morikawa, S., Nakamura, H., Sekikawa, A., Inoue, T., Kanai, H., Sarai, A., Ishii, S., and Nishimura, Y. (1994). Solution Structure of a Specific DNA Complex of the Myb DNA-Binding Domain with Cooperative Recognition Helices. *Cell* **79**, 639-648.
74. Pavelitch, N.P. and Pabo, C.O. (1991). Zinc Finger-DNA Recognition: Crystal Structure of a Zif268-DNA complex at 2.1 Å. *Science* **252**, 809-817.
75. Theunissen, O., Rudt, F., Guddat, U., Mentzel, H., and Pieler, T. (1992). RNA and DNA Binding Zinc Fingers in Xenopus TFIIIA. *Cell* **71**, 679-690.
76. Liao, X., Clemens, K.R., Tennant, L., Wright, P.E., and Gottesfeld, J.M. (1992). Specific Interaction of the First Three Zinc Fingers of TFIIIA with the Internal Control Region of the Xenopus 5S RNA Gene. *J. Mol. Biol.* **223**, 857-871.
77. Gehring, U. (1987). Steroid hormone receptors: biochemistry, genetics and molecular biology. *Trends Biochem.* **12**, 399-402.
78. Reece, R.J. and Ptashne, M. (1993). Determinants of Binding-Site Specificity Among Yeast C₆ Zinc Cluster Proteins. *Science* **261**, 909-911.
Framework and Specificity Rules to Design Specific DNA Binding Proteins. *Proc. Natl. Acad. Sci. USA* **90**, 2256-2260.
79. Summers, M.F., South, T.L., Kim, B., and Hare, D.R. (1990). High-resolution structure of an HIV zinc fingerlike domain via a new NMR-based distance geometry approach. *Biochemistry* **29**, 329-340.

46. Hedge, R.S., Grossman, S.R., Laimins, L.A., and Sigler, P.B. (1992). The 1.7 Å Structure of the Bovine Papillomavirus-1 E2 DNA-binding Domain Bound to its DNA Target. *Nature* **359**, 505-512.
47. Seeman, N.C., Rosenberg, J.M., and Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci. USA* **73**, 804-809.
48. Otwinowski, Z., Schevitz, R.W., Zhang, R.G., Lawson, C.L., Joachimiak, A., Marmorstein, R.Q., Luisi, B.F., and Sigler, P.B. (1988). Crystal Structure of *trp* Repressor/Operator Complex at Atomic Resolution. *Nature* **335**, 321-329.
49. Brennan, R. G., Roderick, S.L., Takeda, Y., and Matthews, B.W. (1990). Protein-DNA conformational changes in the crystal structure of a λ Cro-operator complex. *Proc. Natl. Acad. Sci. USA* **87**, 8165-8169.
50. Schultz, S.C., Shields, G.C., and Steitz, T.A. (1991). Crystal Structure of a CAP-DNA Complex: the DNA is Bent by 90°. *Science* **253**, 1001-1007.
51. Jordan, S.R. and Pabo, C.O. (1988). Structure of the Lambda Complex at 2.5 Å Resolution: Details of the Repressor-Operator Interactions. *Science* **242**, 893-899.
52. Aggarwal, A.K, Rodgers, D.W, Drottar, M., Ptashne, M., and Harrison, S.C. (1988). Recognition of a DNA Operator by the Repressor of Phage 434: A View at High Resolution. *Science* **242**, 899-907.
53. Mondragón, A., Wolberger, C., and Harrison, S.C. (1989). Structure of phage 434 Cro protein at 2.35 Å resolution. *J. Mol. Biol.* **205**, 179-188.
54. Pabo, C.O. and Lewis, M. (1982). The operator-binding domain of λ repressor: structure and DNA recognition. *Nature* **298**, 443-447.
55. Anderson, W.F., Ohlendorf, D.H., Takeda, Y., and Matthews, B.W. (1981). Structure of the Cro repressor from bacteriophage λ and its interaction with DNA. *Nature* **290**, 754-758.
56. Mckay, D.B. and Steitz, T.A. (1981). Structure of catabolite gene activator protein at 2.9 Å resolution suggests binding to left-handed B-DNA. *Nature* **290**, 744-749.
57. Laughon, A. (1991). DNA Binding Specificity of Homeodomains. *Biochemistry* **30**, 11357-11367.
58. Otting, G. Qian, Y.Q., Billeter, M., Müller, M., Affolter, M., Gehring, W.J., and Wüthrich, K. (1990). Protein-DNA Contacts in the Structure of a Homeodomain-DNA Complex Determined by Nuclear Magnetic Resonance Spectroscopy. *EMBO J.* **9**, 3085-3092.

80. Omichinski, J.G., Trainor, C., Evans, T., Gronenborn, A.M., Clore, G.M, and Felsenfeld, G. (1993). A Small Single-'Finger' Peptide from the Erythroid Transcription Factor GATA-1 Binds Specifically to DNA as a Zinc or Iron Complex. *Proc. Natl. Acad. Sci. USA* **90**, 1676-1680.
81. Qian, X., Jeon, C., Yoon, H., Agarwal, K., and Weiss, M.A. (1993). Structure of a new nucleic-acid-binding motif in eukaryotic transcriptional elongation factor TFIIS. *Nature* **365**, 277-279.
82. Fourmy, D., Dardel, F., and Blanquet, S. (1993). Methionyl-tRNA Synthetase Zinc Binding Domain. *J. Mol. Biol.* **231**, 1078-1089.
83. Michelson, J.W., Schmeichel, K.L., Beckerle, M.C., and Winge, D. (1993). The LIM Motif Defines a Specific Zinc-Binding Protein Domain. *Proc. Natl. Acad. Sci. USA* **90**, 4404-4408.
84. Pérez-Alvarado, G.C., Miles, C., Michelson, J.W., Louis, H.A., Winge, D.R., Beckerle, M.C., and Summers, M.F. (1994). Structure of the C-terminal LIM domain from the cysteine rich protein CRP. *Nature Struct. Biol.* **1**, 388-398.
85. Barlow, P.N., Luisi, B., Milner, A., Elliot, M., and Everett, R.J. (1994). Structure of the C₃HC₄ Domain by ¹H-nuclear Magnetic Resonance Spectroscopy. *J. Mol. Biol.* **237**, 201-211.
86. Hommel, U., Zurini, M., and Luyten, M. (1994). Solution structure of a cysteine rich domain of rat protein kinase C. *Nature Struct. Biol.* **1**, 383-387.
87. Jacobs, G.H. (1992). Determination of the Base Recognition Positions of Zinc Fingers from Sequence Analysis. *EMBO J.* **11**, 4507-4517.
88. Fairwell, L., Schwabe, J.W.R., Chapman, L., Finch, J.T., and Rhodes, D. (1993). The crystal structure of a two zinc-finger peptide reveals an extension to the rules for zinc-finger/DNA recognition. *Nature* **366**, 483-487.
89. Pavletich, N.P., and Pabo, C.O. (1993). Crystal Structure of a Five-Finger GLI-DNA Complex: New Perspectives on Zinc Fingers. *Science* **261**, 1701-1707.
90. Ruppert, J.M., Vogelstein, B., and Kinzler, K.W. (1991). The Zinc Finger Protein GLI Transforms Primary Cells in Cooperation with Adenovirus E1A. *Mol. Cell Biol.* **11**, 1724-1729.
91. Klein-Hitpaß, L., Schorpp, M., Wagner, U., and Ryffel, G.U. (1986). An estrogen-responsive element derived from the 5' flanking region of the Xenopus vitellogenin A2 gene functions in transfected human cells. *Cell* **46**, 1053-1061.
93. Scheidereit, C., Westphal, H.M., Carlson, C., Bosshard, H., and Beato, M. (1986). Molecular mode of the interaction between the glucocorticoid receptor and the regulatory elements of inducible genes. *DNA* **5**, 383-391.

94. Luisi, B.L., Xu, W.X., Otwinowski, Z., Freedman, L.P., Yamamoto, K.R., And Sigler, P.B. (1991). Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. *Nature* **352**, 497-505.
95. Schwabe, J.W.R., Chapman, L., Finch, J.T., and Rhodes, D. (1993). The Crystal Structure of the Estrogen Receptor DNA-Binding Domain Bound to DNA: How Receptors Discriminate between Their Response Elements. *Cell* **75**, 567-578.
96. Lee, M.S., Kliewer, S.A., Provecal, J., Wright, P.E., and Evans, R.M. (1993). Structure of Retinoid X Receptor α DNA Binding Domain: A Helix Required for Homeo-domain DNA Binding. *Science* **260**, 1117-1121.
97. Knegtel, R.M.A., Katahira, M., Schilthuis, J.G., Borwin, A.M.J.J., Boelens, R., Eib, D., van der Saag, P.T., and Kaptein, R. (1993). The Solution Structure of the Human Retinoic Acid Receptor- β DNA Binding Domain. *J. Biomolec. NMR* **3**, 1-17.
98. Marmorstein, R., Carey, M., Ptashne, M., and Harrison, S. (1992). DNA recognition by GAL4: structure of a protein-DNA complex. *Nature* **356**, 408-414.
99. South, T.L., and Summers, M.F. (1993). Zinc- and sequence dependent binding to nucleic acids by the N-terminal zinc finger of the HIV-1 nucleocapsid protein: NMR structure of the complex with the Psi-site analog, dACGCC. *Protein Science* **2**, 3-19.
100. Omichinski, J.G., Clore, G.M, Schaad, O., Felsenfeld, G., Trainor, C., Appella, E., Stahl, S.J., and Gronenborn, A.M. (1993). NMR Structure of a Specific DNA Complex of Zn-Containing DNA Binding Domain of GATA-1. *Science* **261**, 438-446.
101. Struhl, K.(1992). Yeast GCN4 Transcriptional Activator Protein. In: *Transcriptional Regulation*. Volume 2 (Eds. S. Mcknight and K.Yamamoto) pp.833-859, Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press.
102. Konig, P. and Richmond, T.J. (1993). The X-ray Structure of GCN4-bZIP Bound to ATF/CREB Site DNA Shows the Complex Depends on DNA Flexibility. *J. Mol. Biol.* **233**, 139-154.
103. Ellenberger, T.E., Brandl, C.J., Struhl, K., and Harrison, S.C. (1992). The GCN4 Basic-region Leucine Zipper Binds DNA as a Dimer of Uninterrupted α -Helices: Crystal Structure of the Protein-DNA Complex. *Cell* **71**, 1223-1237.
104. Hu, J.C., Newell, N.E., Tidor, B., and Sauer, R.T. (1993). The Role of Surface Contacts in Leucine Zipper Stability. *Protein Science* **2**, 1072-1084.

105. Kouzarides, T. and Ziff, E. (1989). Leucine Zippers of fos, jun, and GCN4 Dictate Dimerization Specificity and Thereby Control DNA Binding. *Nature* **340**, 568-571.
106. Schuermann, M, Neuberg, M., Hunter, J.B., Jenuwein, T., Ryseck, R-P, Bravo, R., and Mueller, R. (1989). The Leucine Repeat Motif in Fos Protein Mediates Complex formation with Jun/AP-1 and is Required for Transformation. *Cell* **56**, 507-516.
107. Glover, M.J.N. and Harrison, S.C. (1995). Crystal structure of the heterodimer bZIP transcription factor c-Fos-c-Jun bound to DNA. *Nature* **373**, 257-261.
108. Murre, C. and Baltimore, D. (1992). The Helix-Loop-Helix Motif: Structure and Function. In: *Transcriptional Regulation*, Volume 2, (Eds. S. Mcknight and K. Yamamoto) pp. 861-879, Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press.
109. Ferre-D'Amare, A.R., Prendergast, G.C., Ziff, E.B., and Burley, S.K. (1993). Recognition by Max of its cognate DNA through a dimeric b/HLH/Z domain. *Nature* **363**, 38-44.
110. Ferre-D'Amare, A.R., Pognonec, P., Roeder, R.G., and Burley, S.K. (1994). Structure and function of the b/HLH/Z domain of USF. *EMBO J.* **13**, 180-189.
111. Ellenberger, T., Fass, D., Arnaud, M., and Harrison, S.C. (1994). Crystal structure of transcription factor E47: E-box recognition by a basic region helix-loop-helix dimer. *Genes and Develop.* **8**, 970-980.
112. Ma, P.C.M., Rould, M.A., Weintraub, H., and Pabo, C.O. (1994) Crystal Structure of MyoD bHLH Domain-DNA Complex: Perspectives on DNA Recognition and Implications for Transcriptional Activation. *Cell* **77**, 451-459.
113. Raumann, B.E., Rould, M.A., Pabo, C.O., and Sauer, R.T. (1994). DNA recognition by β -sheets in the Arc repressor-operator crystal structure. *Nature* **367**, 754-757.
114. Somers, W.S. and Phillips, S.E.V. (1992). Crystal structure of the *met* repressor-operator complex at 2.8 Å resolution reveals DNA recognition by β -strands. *Nature* **359**, 387-393.
115. Brown, B.M., Bowie, J.U., and Sauer, R.T. (1990). Arc Repressor is Tetrameric when Bound to Operator DNA. *Biochemistry* **29**, 11189-11195.
116. Brown, B.M. and Sauer, R.T. (1993). Assembly of the Arc Repressor-Operator Complex: Cooperative Interactions Between DNA-Bound Dimers. *Biochemistry* **32**, 1354-1363.

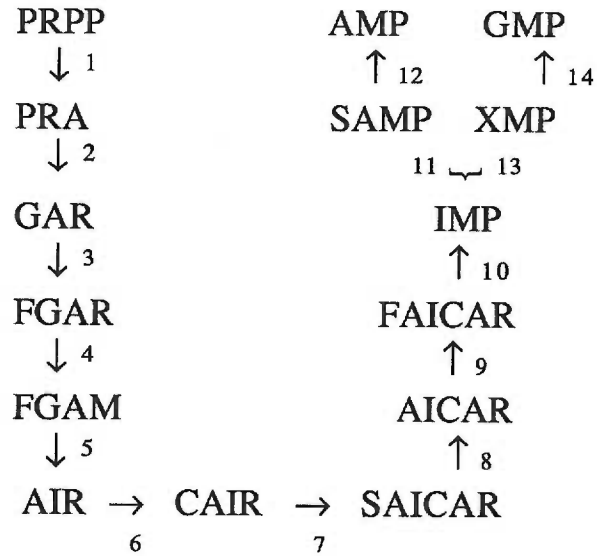
117. Phillips, S.E., Manfield, I., Parsons, I., Davidson, B.E., Rafferty, J.B., Somers, W.S., Margarita, D., Cohen, G.N., Saint, G.I., and Stockley, P.G. (1989). Cooperative Tandem Binding of Met Repressor of *Escherichia coli*. *Nature* **341**, 711-715.
118. Rafferty, J.B., Somers, W.S., Saint-Girons, I., and Phillips, S.E.V. (1989). Three-dimensional crystal structures of *Escherichia coli met* repressor with and without corepressor. *Nature* **341**, 705-710.
119. Phillips, K. and Phillips, S.E.V. (1994). Electrostatic activation of *Escherichia coli* methionine repressor. *Structure* **15**, 309-316.
120. Rigby, P.W.J. (1993). Three in One and One in Three: It all Depends on TBP. *Cell* **72**, 7-10.
121. Sentenac, A. (1985). Eukaryotic RNA Polymerases. *CRC Cur. Rev. Biochem.* **18**, 31-90.
122. Roeder, R.G. (1991). The Complexities of Eukaryotic Transcription Initiation: Regulation of Preinitiation Complex Assembly. *Trends Biochem. Sci.* **16**, 402-408.
123. Zawel, L. and Reinberg, D. (1992). Advances in RNA Polymerase II Transcription. *Curr. Opin. Cell Biol.* **4**, 488-495.
124. Dynlacht, B.D., Hoey, T., and Tijan, R. (1991). Isolation of Coactivators Associated with the TATA-Binding Protein That Mediate Transcriptional Activation. *Cell* **66**, 563-576.
125. Kim, Y., Geiger, J.H., Hahn, S., and Sigler, P.B. (1993). Crystal Structure of a Yeast TBP/TATA-box Complex. *Nature* **365**, 512-520.
126. Kim, J.L., Nikolov, D.B., and Burley, S.K. (1993). Co-Crystal Structure of TBP Recognizing the Minor Groove of a TATA Element. *Nature* **365**, 520-527.
127. Traver, A.A. (1991). DNA bending and kinking-sequence dependence and function. *Curr. Opin. Struct. Biol.* **1**, 114-122.
128. King, C-Y and Weiss, M.A. (1993). The SRY High-Mobility-Group Box Recognizes DNA by partial Intercalation in the Minor Groove: A Topological Mechanism of Sequence Specificity. *Proc. Natl. Acad. Sci. USA* **90**, 11990-11994.
129. Blundell, and Johnson, (1976). *Protein Crystallography*. Academic Press, Inc., San Diego, CA.
130. Stout, G.H. and Jenson, L.H. (1989). *X-Ray Structure Determination*. John Wiley & Sons, Inc., New York.

131. Drenth, J. (1994). *Principles of Protein X-ray Crystallography*. Springer-Verlag, New York, Inc.
132. McRee, D.E. (1993). *Practical Protein Crystallography*. Academic Press, Inc., San Diego, CA.
133. Rhodes, G.(1993). *Crystallography Made Crystal Clear: A Guide for Users of Macromolecular Models*. Academic Press, Inc., San Diego, CA.
134. Bragg, W.L. (1922). The diffraction of x-rays by crystals. Nobel Lectures: *Physics* 1901-21, 370-382.
135. Ewald, P.P. (1913) *Physikalische Zeit.* **14**, 465-472.
136. Churchill, R.V. (1963). *Fourier Series and Boundary Value Problems* McGraw-Hill, New York.
137. Franklin, P. (1958). *An Introduction to Fourier Methods and the Laplace Transform*. Dover, New York.
138. McPherson, A. (1982). *Preparation and Analysis of Protein Crystals*. John Wiley & Sons, Inc., New York.
139. Carter, C.W. and Carter, C.W., Jr. (1979). Protein Crystallization Using Incomplete Factorial Experiments. *J. Biol. Chem.* **254**, 12219-12223.
140. Matthews, B.W. (1968). Solvent content of protein crystals. *J. Mol. Biol.***33**, 491-497.
141. Hamlin, R.(1985). Multiwire Area X-Ray Diffractometers. In: *Methods of Enzymology*. Vol. 114 (Eds. Wycoff, H.W., Hirs, C.H.W., and Timasheff, S.N.) pp.416-452, Academic Press, Inc., San Diego, CA.
142. Howard, A.J., Nielson, C. and Xuong, N.G., (1985). Software for a Diffractometer with Multiwire Area Detector. In: *Methods of Enzymology*. Vol.114 (Eds.Wycoff, H.W., Hirs, C.H.W., and Timasheff, S.N.) pp.452-472, Academic Press, Inc., San Diego, CA.
143. Fraser, R.D.B. and MacRae, T.P. (1969) *Physical Principles and Techniques of Protein Chemistry*, part A (ed. S.J. Leach.) New York: Academic Press.
144. Rossmann, M.G. and Blow, D.M. (1962). The Detection of Sub-Units Within the Crystallographic Asymmetric Unit. In:*The Molecular Replacement Method* (Ed. M.G. Rossmann) pp. 45-57, Gordon and Breach Science Publishers.
145. Lattman, E.E. (1971). Optimal Sampling of the Rotation Function. In: *The Molecular Replacement Method* (Ed. M.G. Rossmann) pp.179-186, Gordon and Breach Science Publishers.

146. Crowther, R.A. (1971). The Fast Rotation Function. In: *The Molecular Replacement Method* (Ed. M.G. Rossmann) pp.174-178, Gordon and Breach Science Publishers.
147. Crowther, R.A. and Blow, D.M. (1967). A Method of Positioning a Known Molecule in an Unknown Crystal. In: *The Molecular Replacement Method* (Ed. M.G. Rossmann) pp.85-98, Gordon and Breach Science Publishers.
148. Fitzgerald, P.M.D. (1988). MERLOT, an integrated package of computer programs for the determination of crystal structures by molecular replacement. *J. Appl. Cryst.* **21**, 273-278.
149. Blow, D.M. and Crick, F.H.C. (1959). The treatment of errors in the isomorphous replacement method. *Acta Cryst.* **12**, 794-802.
150. Jones, A. (1985). Interactive Computer Graphics: FRODO. In: *Methods of Enzymology*. Vol.115 (Eds. Wycoff, H.W., Hirs, C.H.W., and Timasheff, S.N.) pp.157-171, Academic Press, Inc., San Diego, CA.
151. Wang, B.C. (1985). Resolution of Phase Ambiguity in Macromolecular Crystallography. In: *Methods of Enzymology*. Vol.115 (Eds. Wycoff, H.W., Hirs, C.H.W., and Timasheff, S.N.) pp.90-112, Academic Press, Inc., San Diego, CA.
152. Hendrickson, W.A. and Lattman, E.E. (1970). Representation of Phase Probability Distributions for Simplified Combination of Independent Phase Information. *Acta Cryst.* **B26**, 136-143.
153. Tronrud, D.E., Ten Eyck, L.F. & Matthews, B.W. (1987). An Efficient General-Purpose Least-Squares Refinement Program for Macromolecular Structures. *Acta crystallogr.* **A43**, 489-501.

Figure 1. Alignment of the proteins of the LacI family and the D-ribose binding protein. Pairwise comparisons made with Gap and Bestfit programs in Release 6.2 of the Genetics Computer Group Sequence Analysis Software Package, BLAST and the Fasta program were used to construct this multiple alignment with the SequenceEditingAligner. The proteins shown are the Galactose repressor, GalR; Galactose isorepressor, GalS; Lactose repressor from *E. coli*, LacI; Lactose repressor from *K. pneumoniae*, LacI *K.p.*; Amylase repressor from *B. subtilis*, CcpA *B.s.*; Cytidine repressor, CytR; evolved β -galactosidase repressor, EbgR; divergent endoglucanase reading frame from *B. polymyxa*, EndR *B.p.*; Fructose repressor from *E. coli*, FruR; Fructose repressor from *S. typhimurium*, FruR *S.t.*; Maltose repressor, MalI; Purine repressor, PurR; Raffinose repressor, RafR; Ribitol repressor from *K. aerogenes*, RbtR *K.a.*; Sucrose repressor from *K. aerogenes*, ScrR *K.a.*; Sucrose repressor from *S. typhimurium*, ScrR *S.t.*; Sucrose repressor fragment from *V. alginolyticus*, ScrR *V.a.*; Opine utilization repressor from *A. rhizogenes*, OpnR *A.r.*; potential Amylase repressor from *S. violaceus* (identified as opening reading frame), B77 *S.v.*; D-ribose binding protein from *E. coli*, RbsB; and the D-ribose binding protein from *S. typhimurium*, RbsB *S.t.* The signal sequence of the D-ribose binding proteins is not included in the figure. The purported helix-turn-helix DNA binding region and the amino acids believed to be involved in inducer binding (I; shaded), dimerization (D; open box) and the tetramerization (T; cross-hatched) of LacI are indicated above the corresponding regions. Residues highlighted in green; Ser, Thr, Pro, Ala, Gly: yellow; Met, Ile, Leu, Val: light blue; Asn, Asp, Glu, Gln: red; Cys, His, Arg, Lys: orange; Phe, Tyr, Trp. Highly conserved residues are indicated in the same color by a box. Reproduced, with permission from Weickert, M.J. and Adhya, S. (1992). *J. Biol. Chem.* **267**, 15869-15874.

Figure 2. Outline of the *de novo* purine biosynthetic pathway. Given are the common substrate abbreviations and *E. coli* gene designations and map positions (in minutes) for the genes encoding the *de novo* purine biosynthetic enzymes.



| <u>Enzymes</u> | <u><i>E. coli</i> Genes</u> | <u>Map Position</u> | <u>Substrate Abbreviations</u> |
|------------------------------------|-----------------------------|---------------------|---|
| 1. glutamine PRPP amidotransferase | purF | 50.0 | PRPP, 5'-phosphoribosyl-pyrophosphate |
| 2. GAR synthetase | purD | 90.3 | PRA, 5'-phosphoribosylamine |
| 3. GAR transformylase | purN | 53.5 | GAR, 5'-phosphoribosyl glycinamide |
| 4. FGAM synthetase | purL | 55.2 | FGAR, 5'-phosphoribosyl-N-formylglycinamide |
| 5. AIR synthetase | purM | 53.5 | FGAM, 5'-phosphoribosyl-N-formylglycinamide |
| 6. AIR carboxylase | purEK | 12.2 | AIR, 5'-phosphoribosyl-aminoimidazole |
| 7. SAICAR synthetase | purC | 53.3 | CAIR, 5'-phosphoribosyl-aminoimidazole carboxylate |
| 8. adenylosuccinate lyase | purB | 25.2 | SAICAR, 5'-phosphoribosyl-(4(N-succinocarboxamide)-5-aminoimidazole |
| 9. AICAR transformylase | purH | 90.3 | SAICAR |
| 10. IMP cyclohydrolase | purH | 90.3 | AICAR, 5'-phosphoribosyl-4-carboxamide-5-aminoimidazole |
| 11. adenylosuccinate synthetase | purA | 95.0 | FAICAR, 5'-phosphoribosyl-4-carboxamide-5-formyl-aminoimidazole |
| 12. adenylosuccinate lyase | purB | 25.2 | IMP, inosine monophosphate |
| 13. IMP dehydrogenase | guaB | 54.0 | SAMP, adenylosuccinate |
| 14. GMP synthetase | guaA | 54.0 | XMP, xanthosine monophosphate |

Figure 3. Schematic diagram of the PurR regulated *pur* regulon. The PurR regulatory circuit involves genes encoding enzymes for *de novo* purine biosynthesis, *de novo* pyrimidine biosynthesis, pyrimidine salvage, the production of one carbon units (*glyA* encoding serine hydroxymethyltransferase), polyamine synthesis (*speA* encoding arginine decarboxylase), PRPP synthesis (*prsA* encoding PRPP synthetase). PurR is also autoregulated.

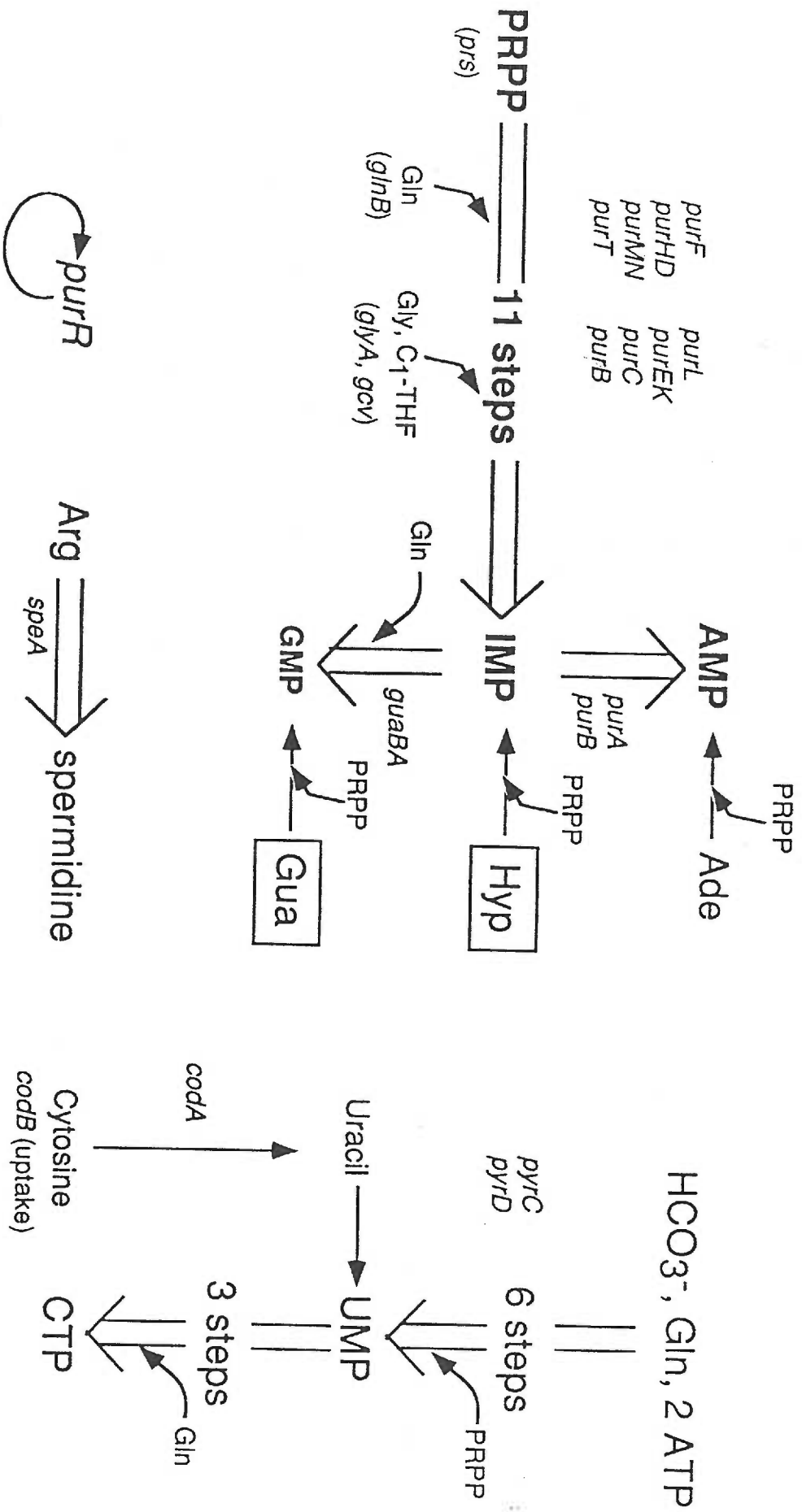
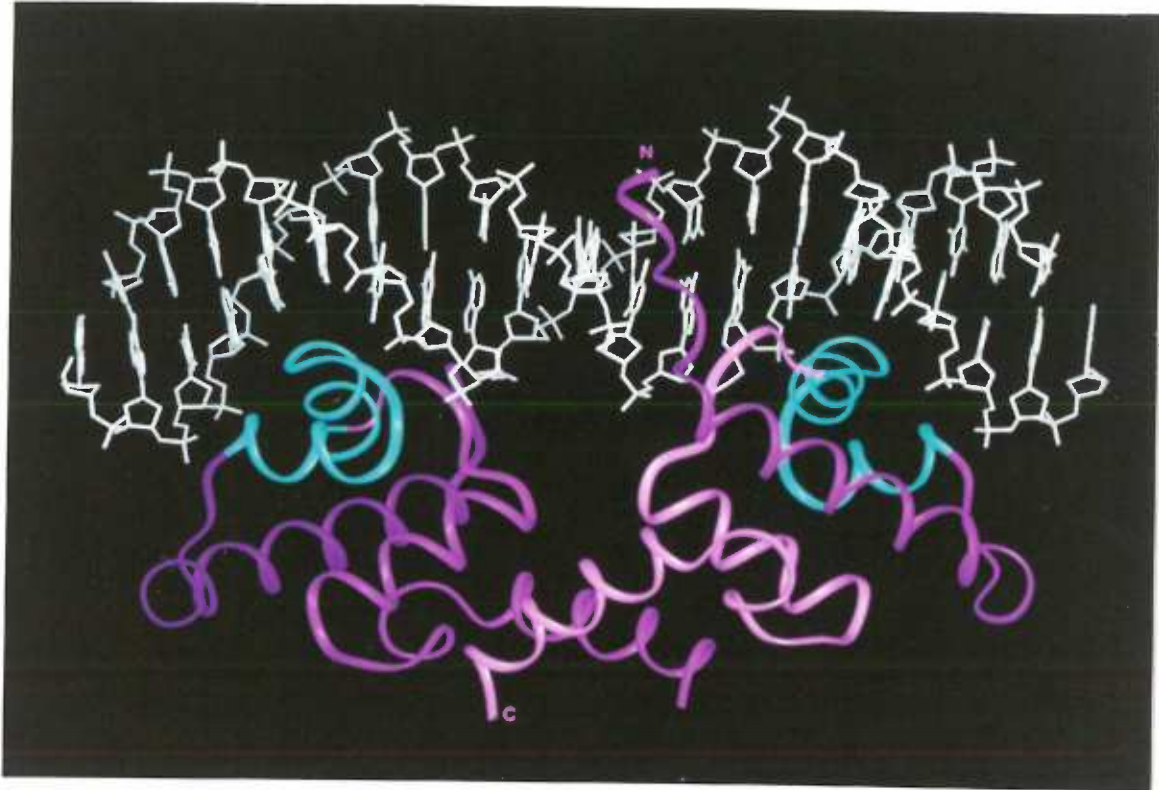
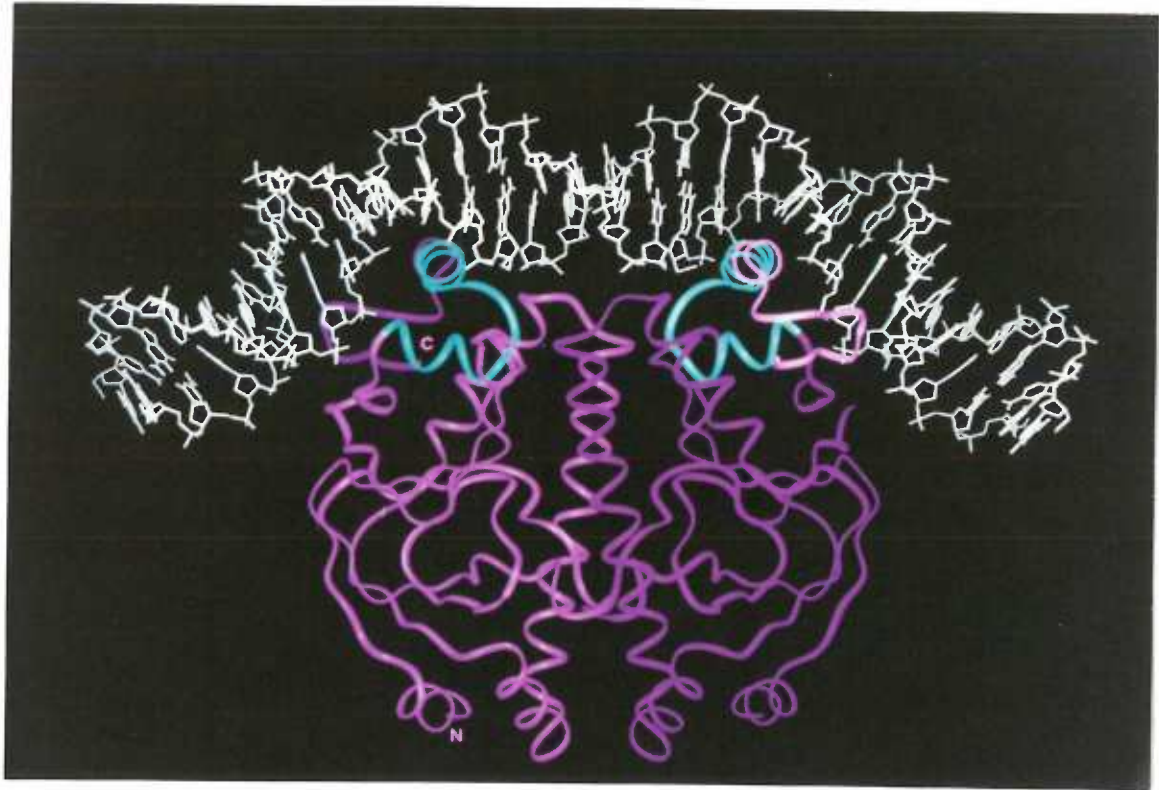


Figure 4. Examples of helix-turn-helix (HTH) DNA binding proteins. Each protein is represented as a ribbon diagram and the DNA is shown in white sticks. Each HTH element is colored blue and the N and C termini are labelled. A) The dimeric DNA binding domain of the λ repressor-DNA complex. B) The dimeric CAP DNA complex. C) The POU domain of Oct-1 bound to DNA. The N-terminal POU-specific domain and the COOH-terminal POU homeodomain are distinguished by labels.

A



B



C

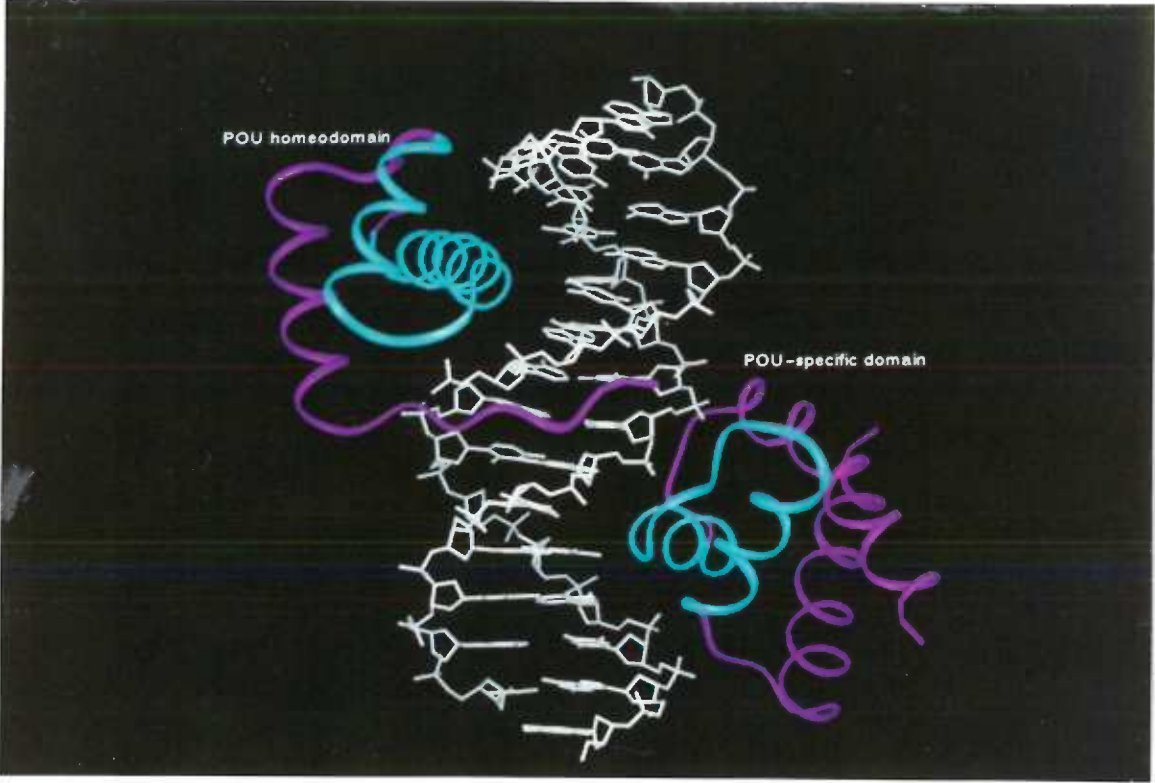


Figure 5. Stereo diagram of a canonical helix-turn-helix (HTH) motif. The positions subject to stereochemical constraints are shaded and numbered. The first helix of the HTH element is labelled "Helix 1" and the second helix, the recognition helix, is labelled "Helix 2". The HTH from the PurR-hypoxanthine-DNA complex was used to generate the figure.

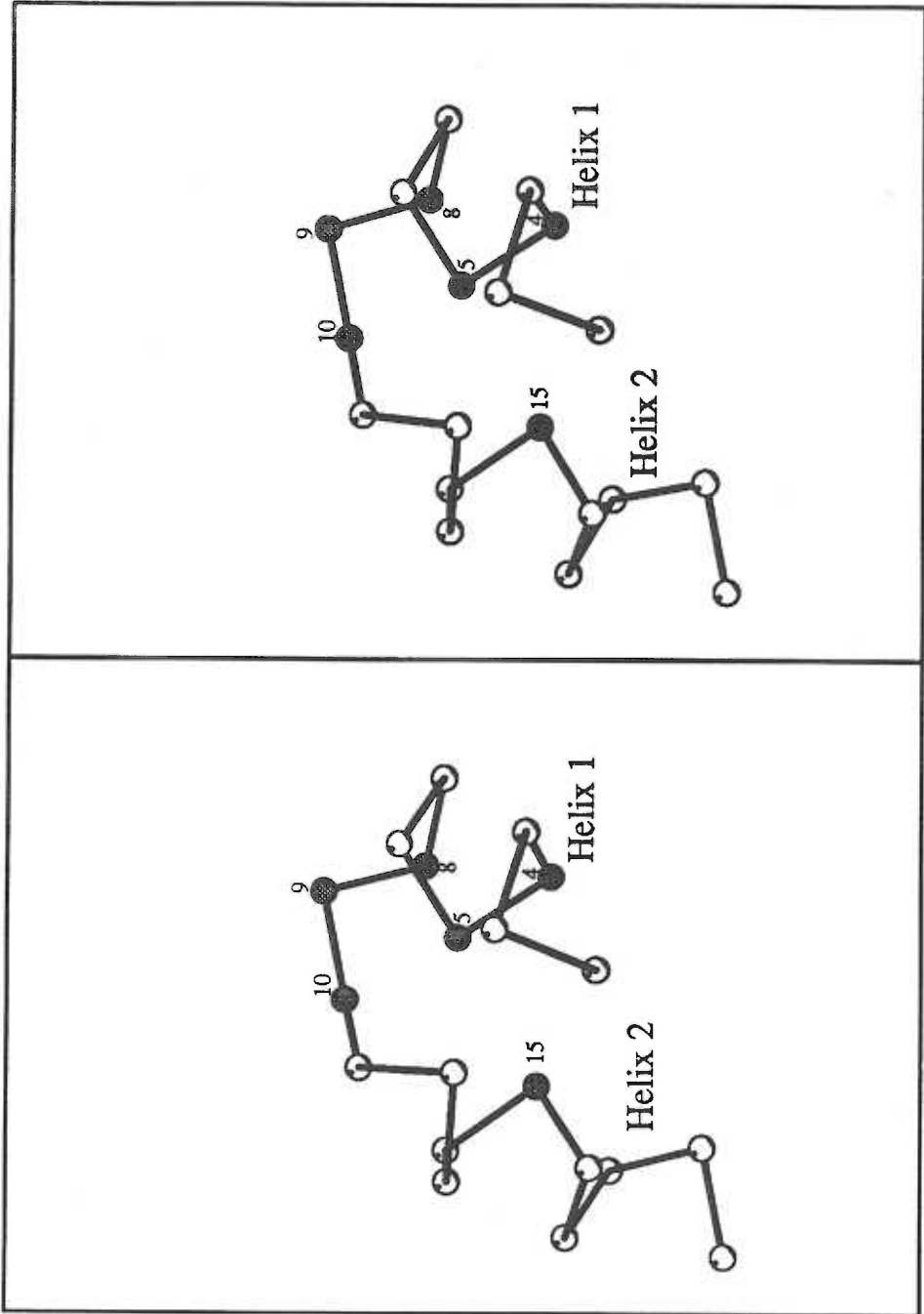


Figure 6. λ repressor DNA contacts. The HTH is colored blue as in Figure 4. Contacts critical for DNA binding specificity are shown in stick form. Important hydrogen bonds are also shown.

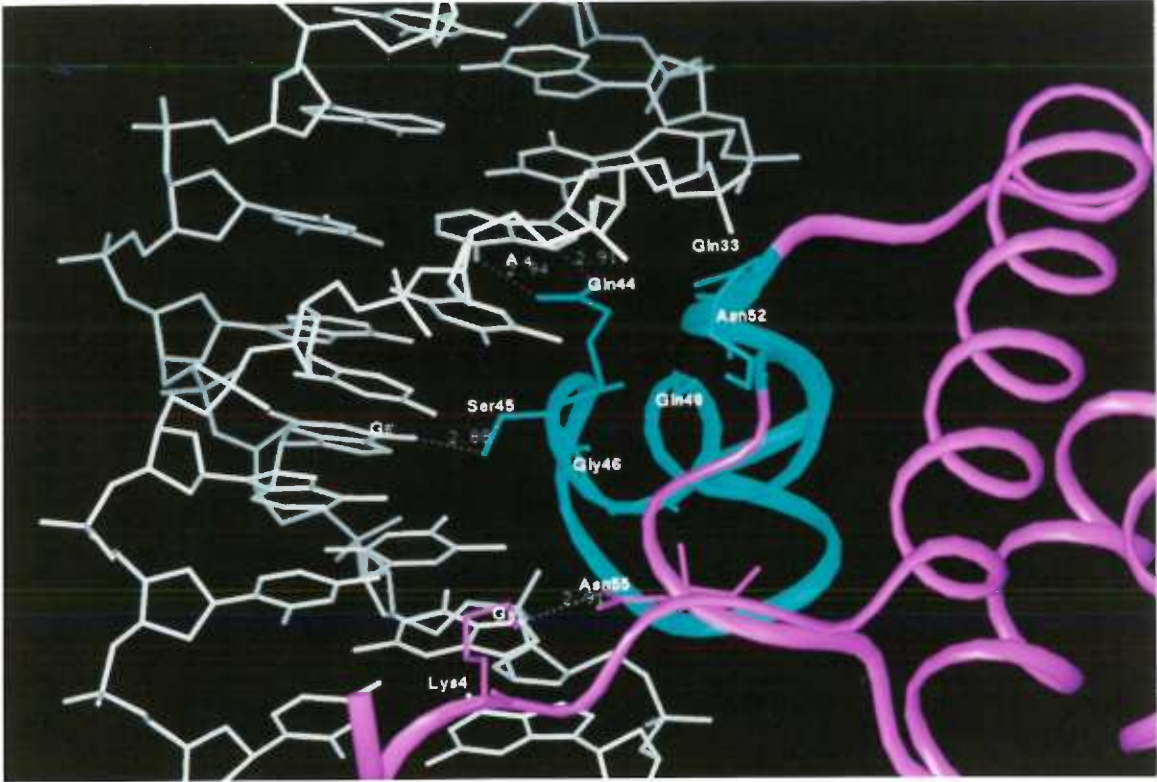
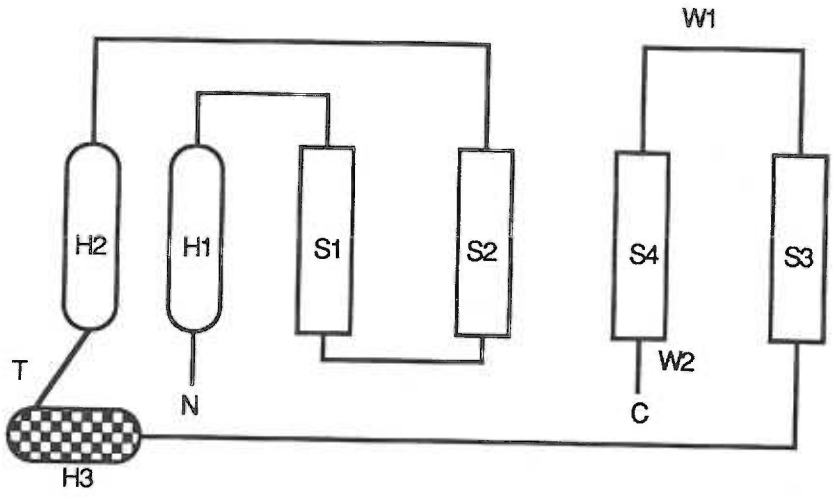


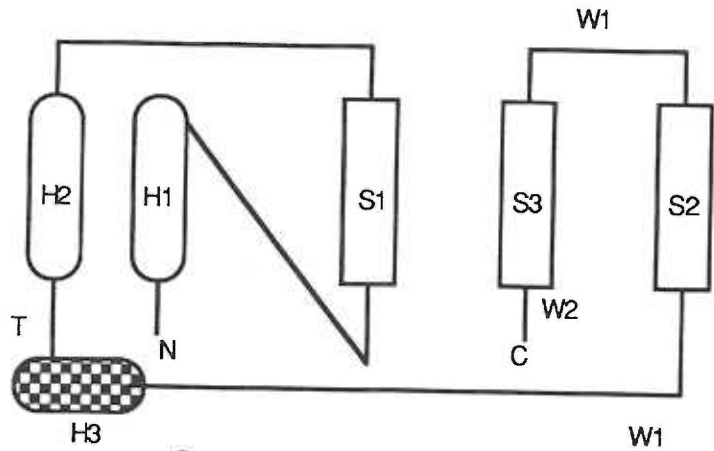
Figure 7. Schematic drawings illustrating the known subclasses of structural winged-helix DNA-binding domains. α -helices are represented by elongated ellipses and labelled H and β -strands by rectangles and labelled S. The recognition helices are hatched. The first subfamily, which includes CAP and HNF-3/fork head, contains a H1-S1-S2-H2-T-H3-S3-W1-S4-W2 topology. The second subfamily contains a H1-S1-H2-T-H3-S2-W1-S3-W2 topology. MuA represents a divergent winged-helix with the topology S1-H1-T-H2-S2-W1-S3-H3.

**Structural
Winged-Helix
Subfamilies:**

Subfamily 1



Subfamily 2



MuA

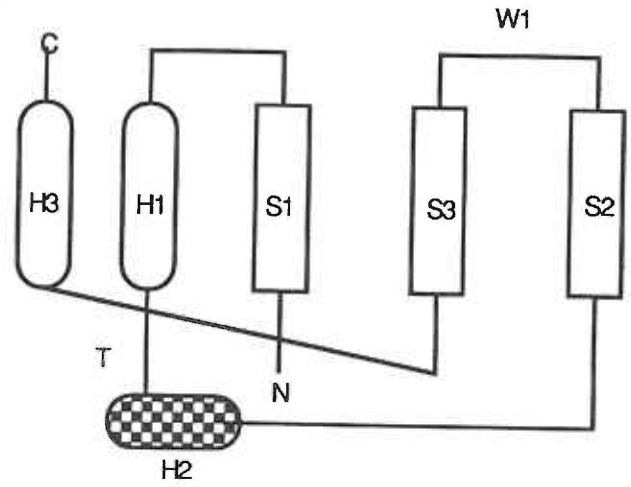
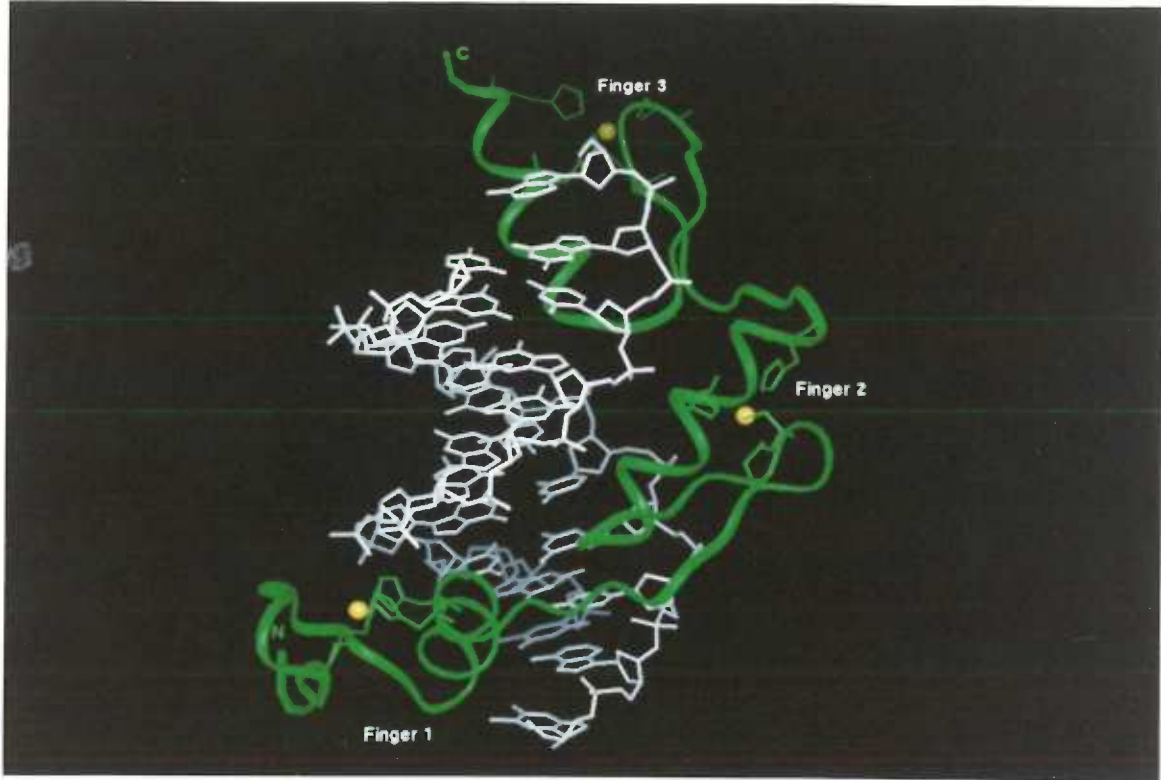
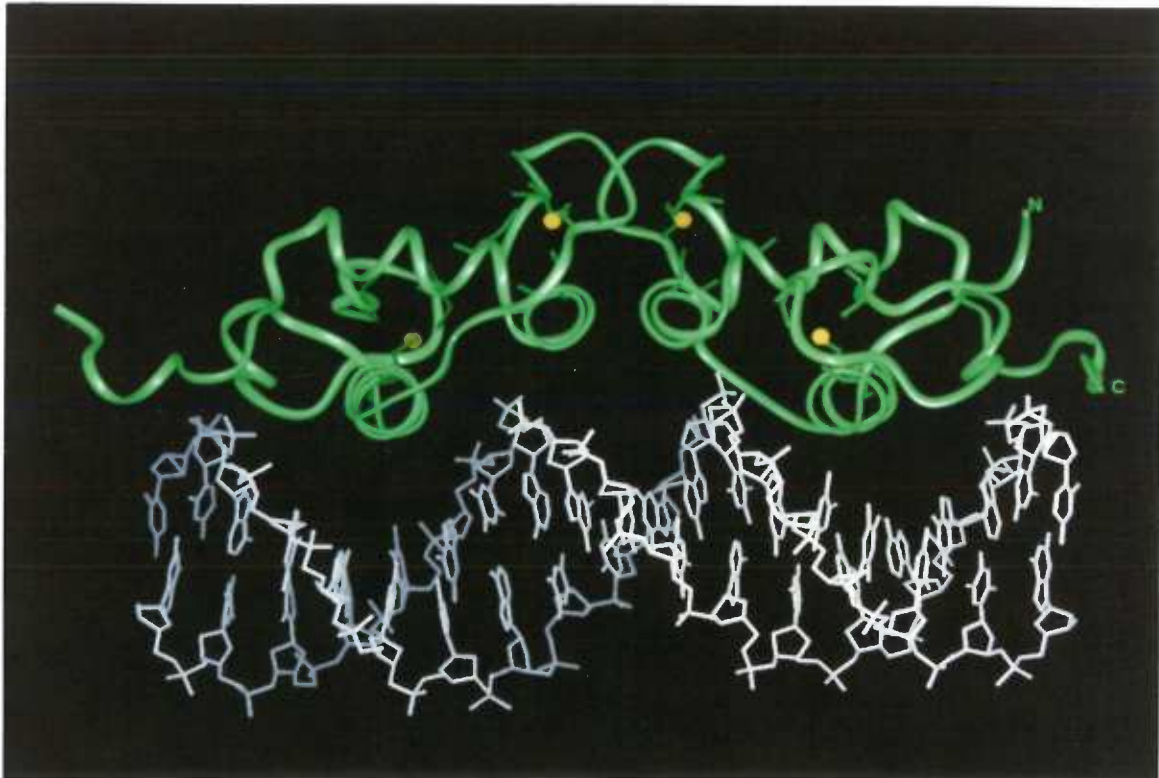


Figure 8. Examples of zinc domain DNA binding proteins. The zinc containing DNA binding domains of each are represented by ribbon diagrams and colored green. The zinc atoms are represented as balls and colored yellow and the DNA is represented as white sticks. The N and C termini are labelled. A) Zif268 bound to DNA. Each of the three fingers are labelled. B) Glucocorticoid bound to DNA. C) GAL4 bound to DNA. D) Zinc finger of the HIV (human immunodeficiency virus) nucleocapsid protein, Zn(HIV-F1) bound to DNA. E.) GATA-1 bound to DNA.

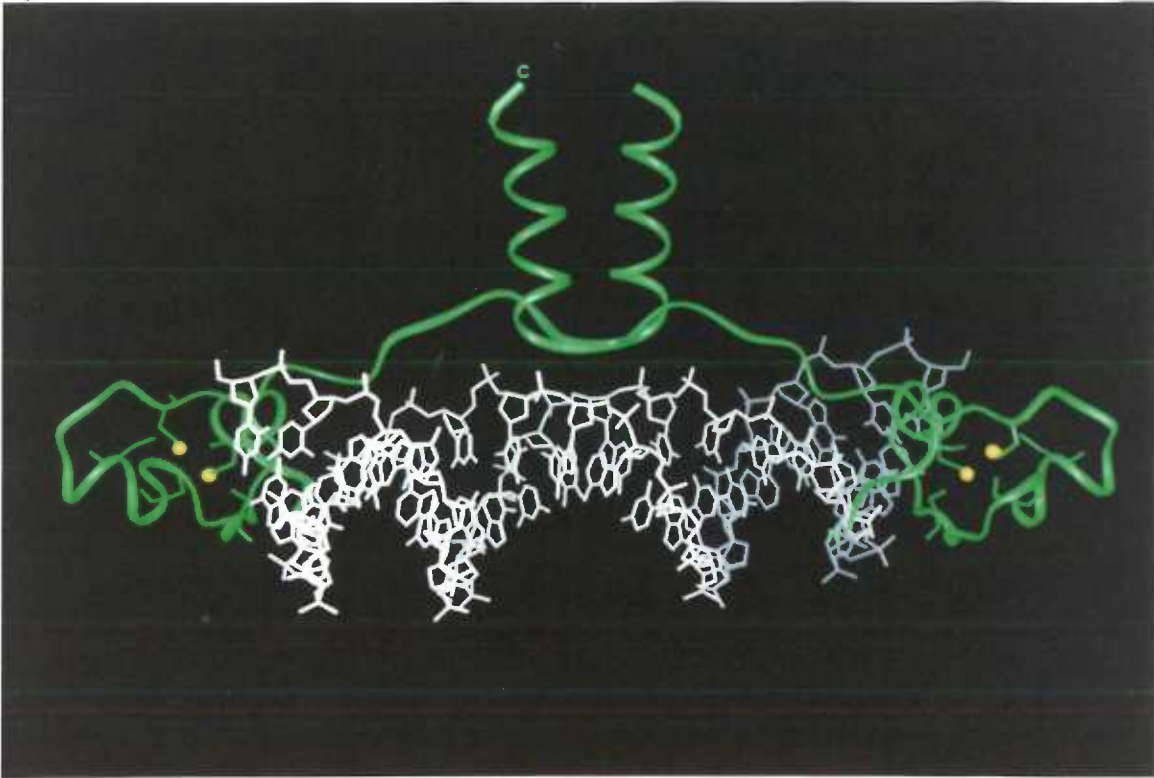
A



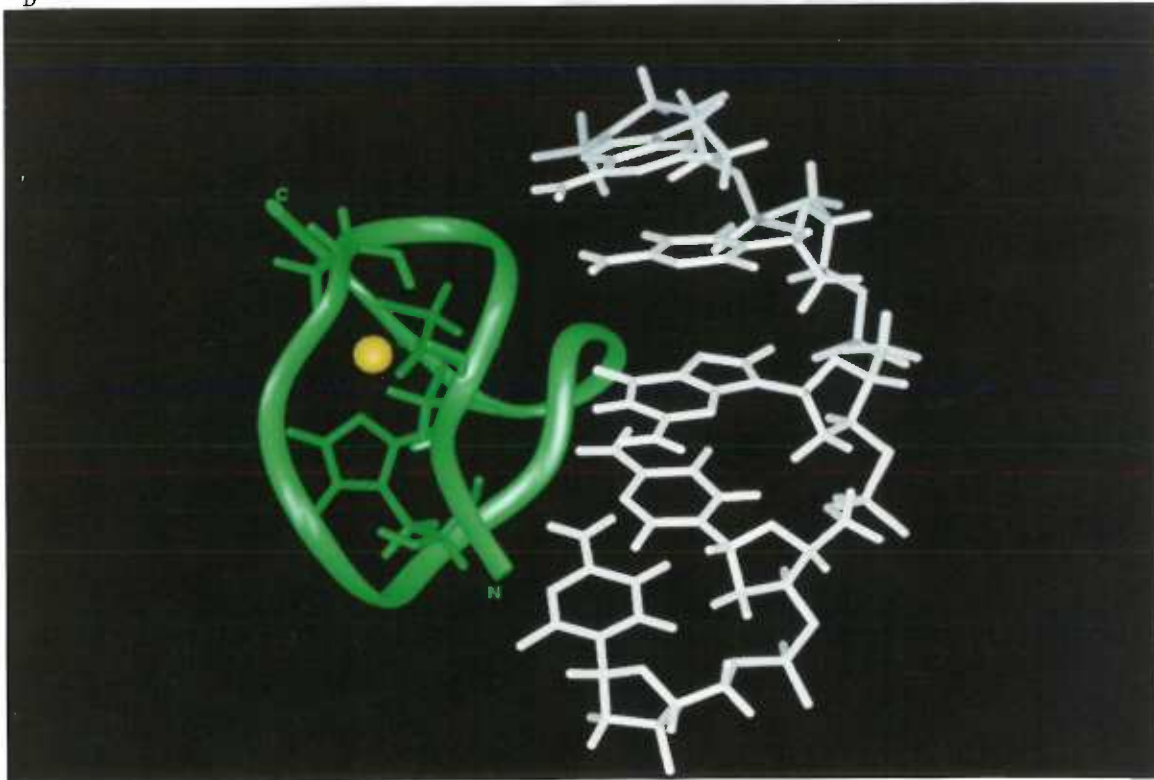
B



C



D



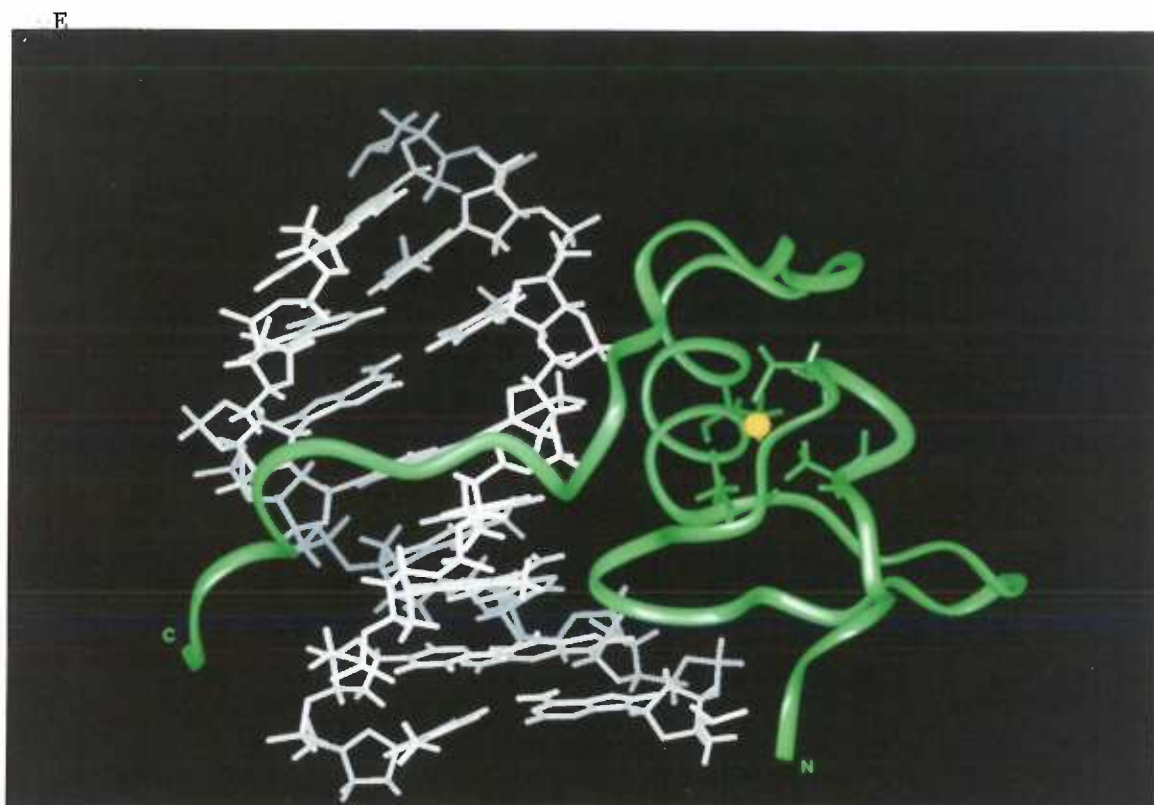


Figure 9. Zif268 DNA contacts. The protein is colored as in Figure 7. Residues making contacts from each finger are colored yellow. The conserved Arg, which is located one residue NH₂-terminal to the helix, and contacts guanine in each case, is indicated by an asterisk. The 5' and 3' ends of the DNA and the N and C termini of the protein are labelled, underscoring the opposite polarity of binding in which the protein binds from N to C termini in the 3' to 5' direction.

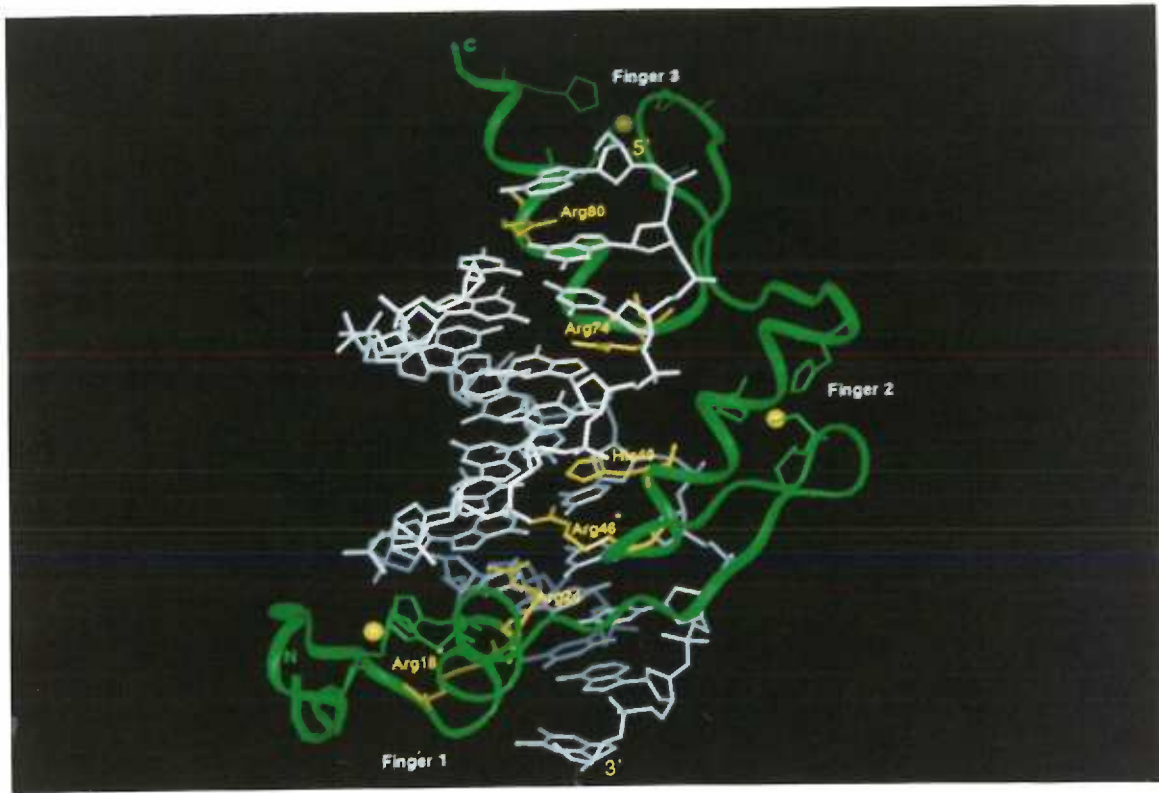
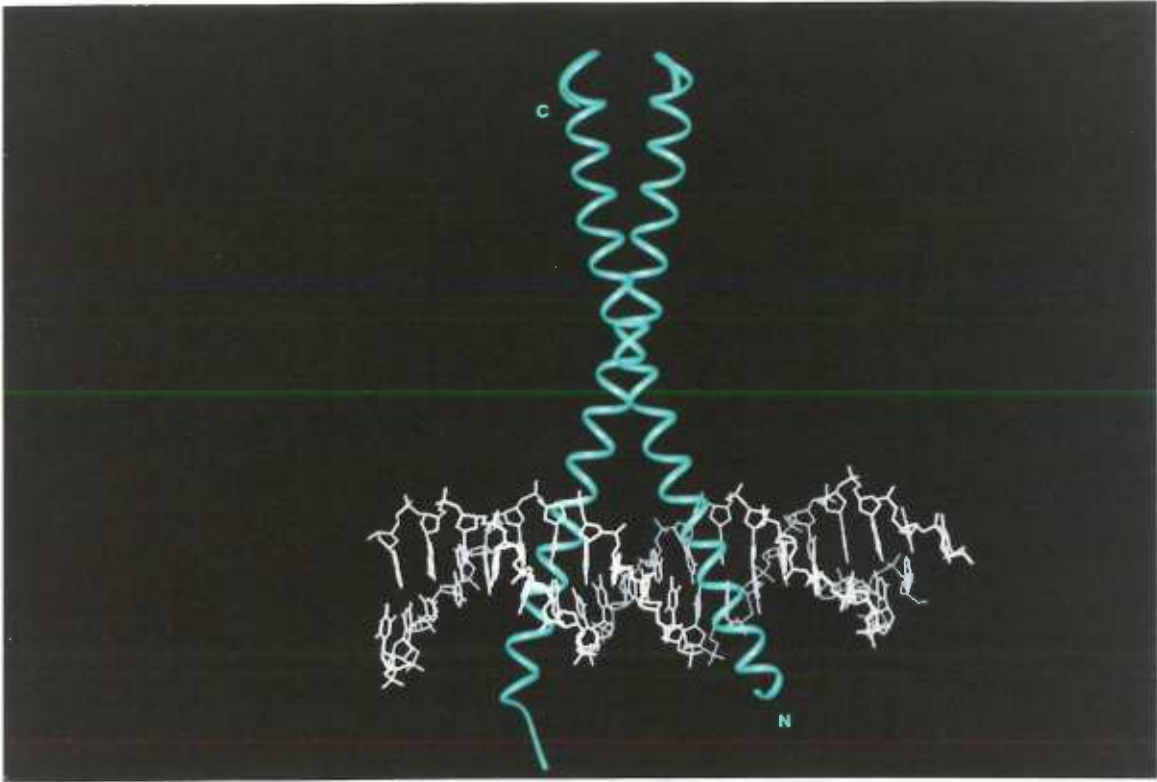


Figure 10. Examples of b/ZIP and b/HLH DNA binding domains. The proteins are shown as ribbon diagrams and are colored blue-green. The DNA is represented as white sticks. The N and C termini are labelled. A) the dimeric b/ZIP region of GCN4 bound to an AP-1 site. B) the dimeric b/HLH of MyoD bound to DNA.

A



B

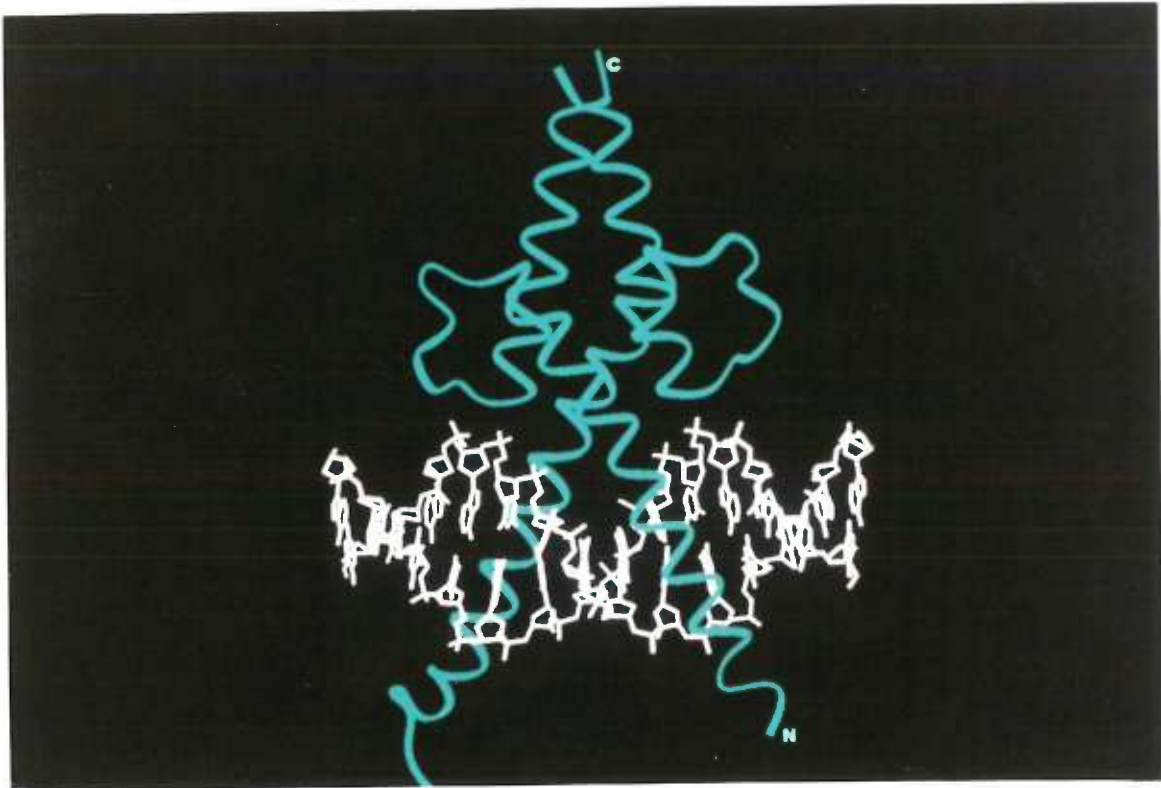


Figure 11. b/ZIP and b/HLH contacts compared. A) Looking down the basic region helix of one GCN4 dimer bound to the asymmetric AP-1 site. Hydrogen bonds are provided by Asn235, which is conserved in nearly all b/ZIP proteins. Hydrophobic contacts are provided by Ala238, Ala239 and Ser242. Arg232 makes key phosphate contacts. Asymmetric contacts between the monomers are made by Arg243. In this half-site, Arg243 contacts the phosphate backbone. In the other half-site (not shown) Arg243 contacts a guanine. B) Looking down the basic region helix of one MyoD monomer. Glu118, like the Asn of the b/ZIP proteins, is highly conserved among the b/HLH and b/HLH/ZIP proteins and makes important hydrogen bonds to the DNA. The Glu side chain also makes key van der Waals interactions with thymine methyl groups. Arg121 contacts the phosphate backbone and forms a salt bridge with Glu118, stabilizing its interactions with DNA bases. Thr115 makes van der Waals contact to a thymine and Arg111 contacts a guanine. The conformation of this arginine appears to differ between the b/HLH proteins solved thus far.

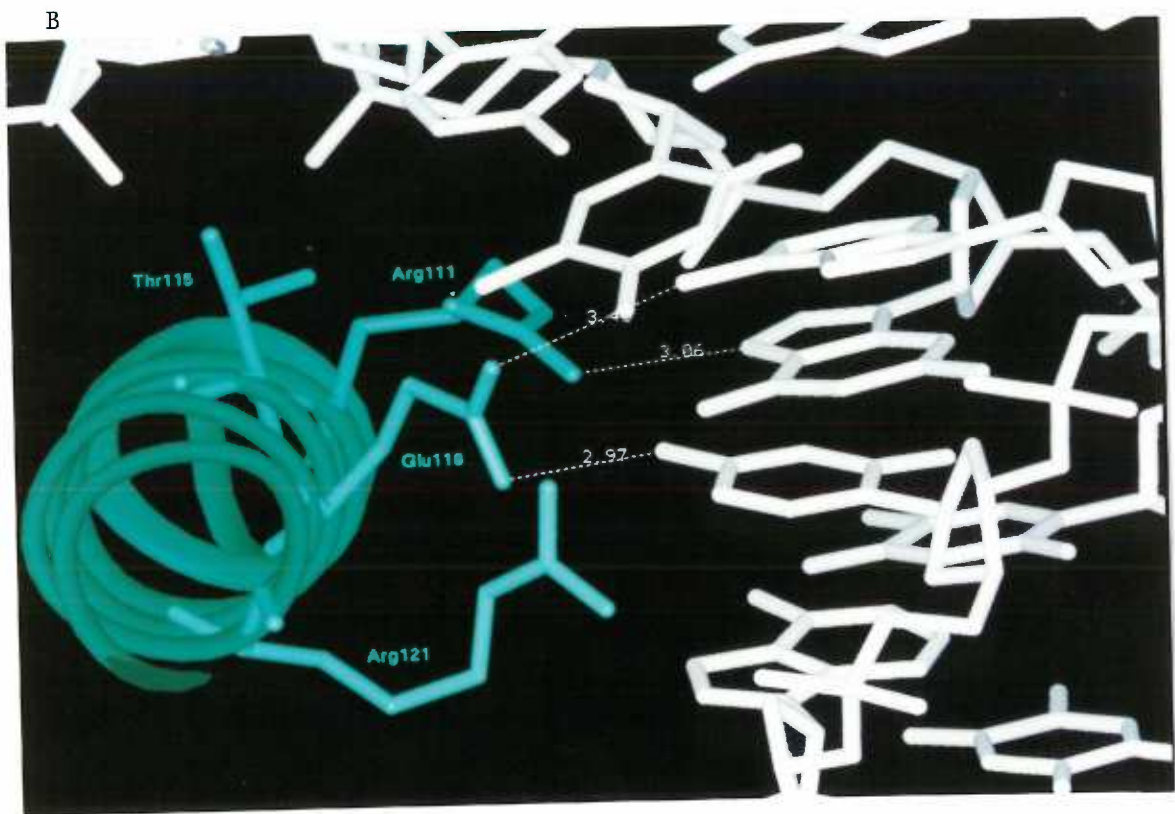
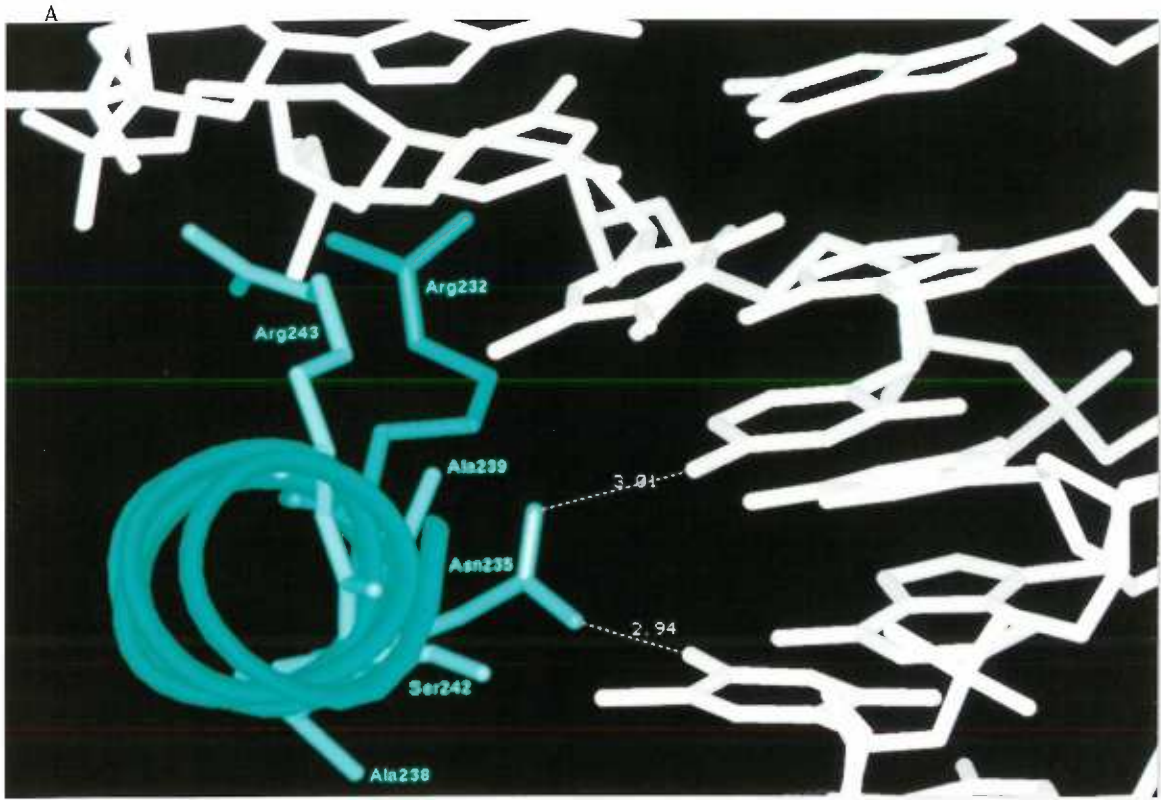


Figure 12. Examples of β -sheet DNA binding proteins. The proteins are represented as ribbons, the DNA by white sticks. The NH₂- and COOH-termini of one monomer subunit is, in each case, labelled. Colors used to represent monomers are red, purple, green and blue. A) Dimer of MetJ bound to DNA. The other crystallographically related dimer, not shown, combines with the dimer shown to bind DNA as a tetramer with each dimer making identical contacts, as implied by the crystallographic symmetry. B) The crystallographic tetramer of Arc bound to DNA. Each dimer in Arc makes different contacts to each DNA half-site.

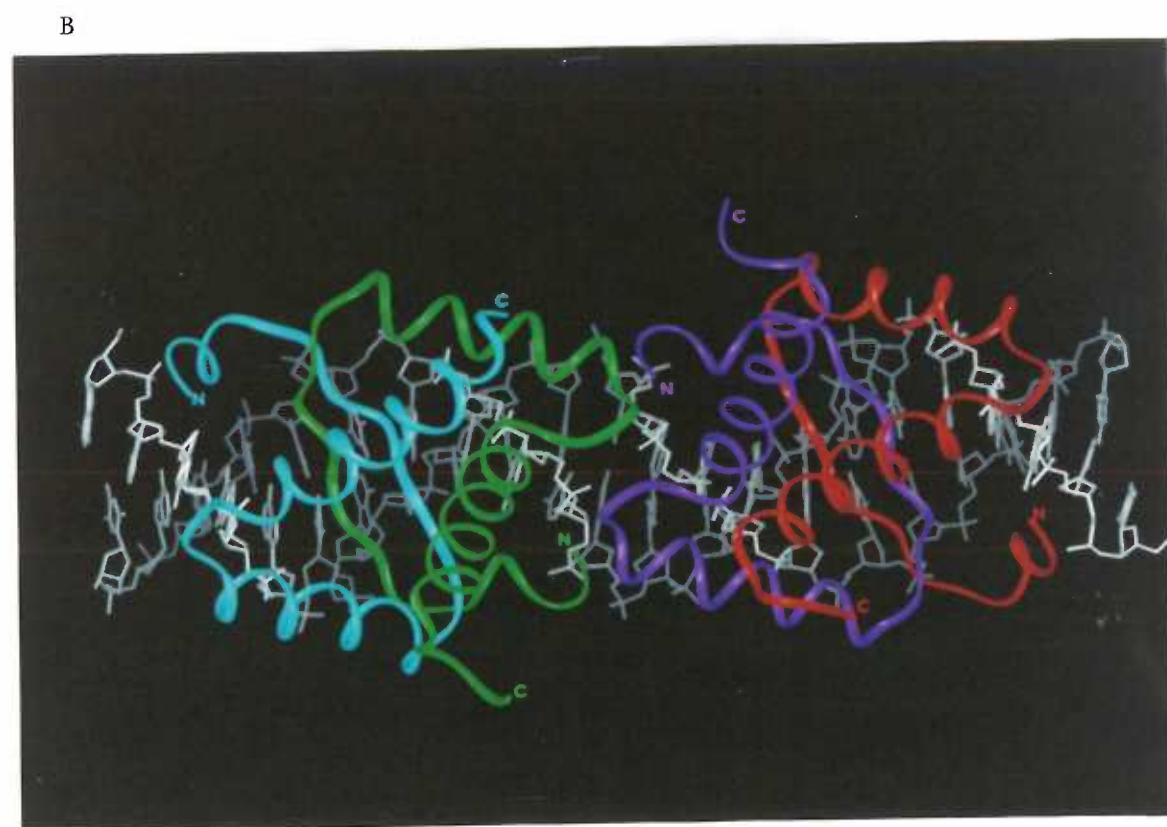
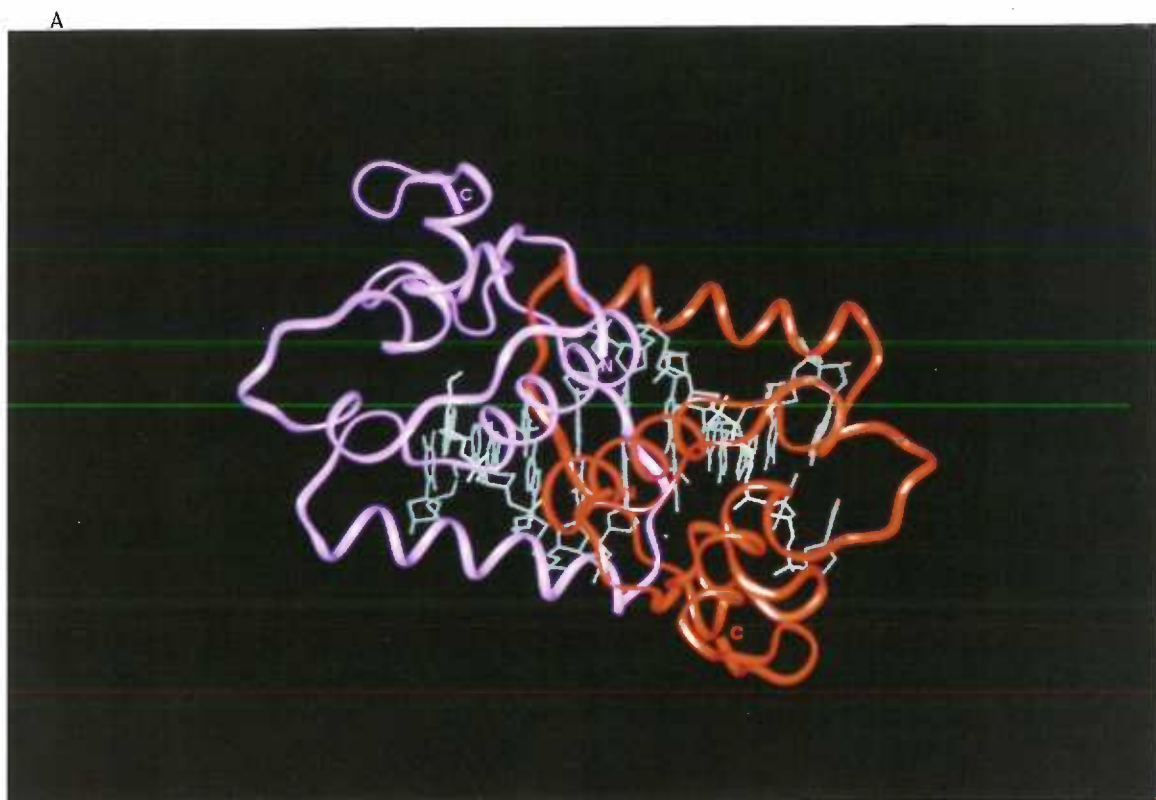
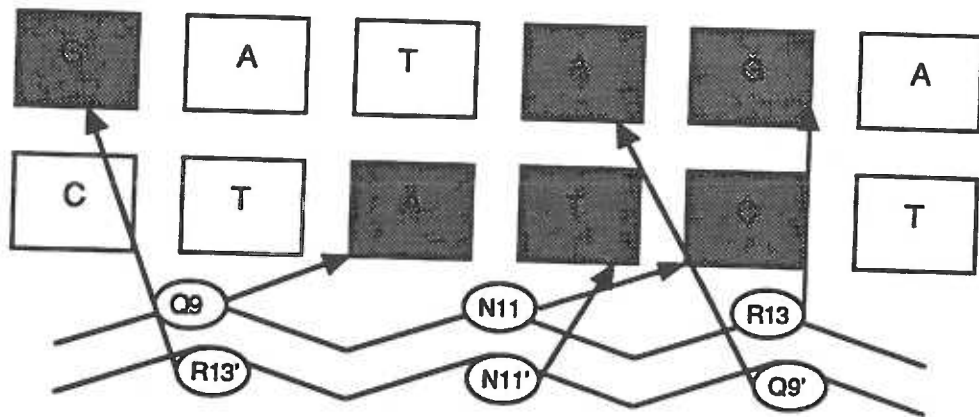
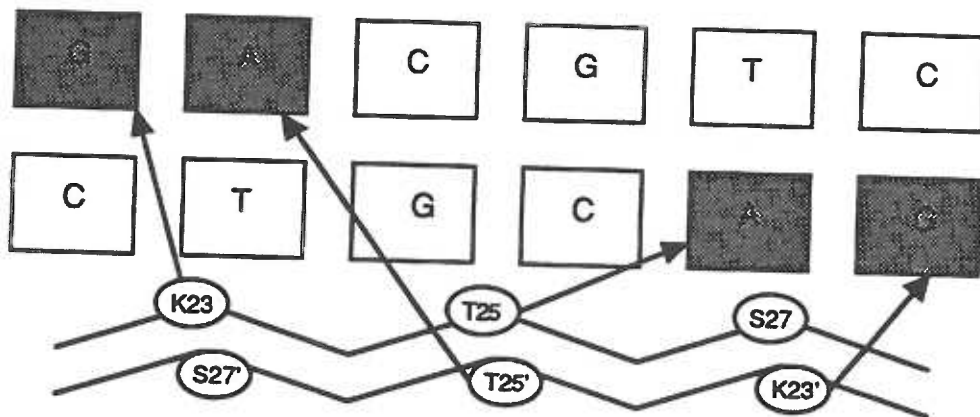


Figure 13. Schematic diagram comparing MetJ and Arc DNA contacts.

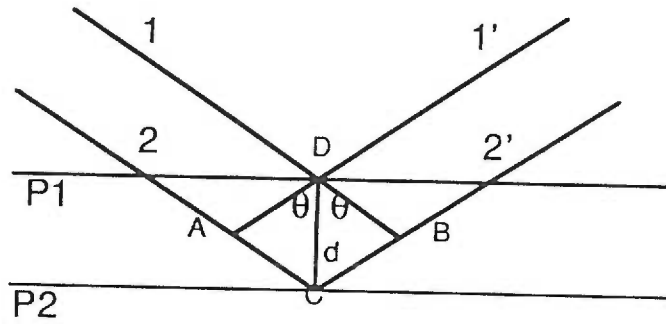


Arc

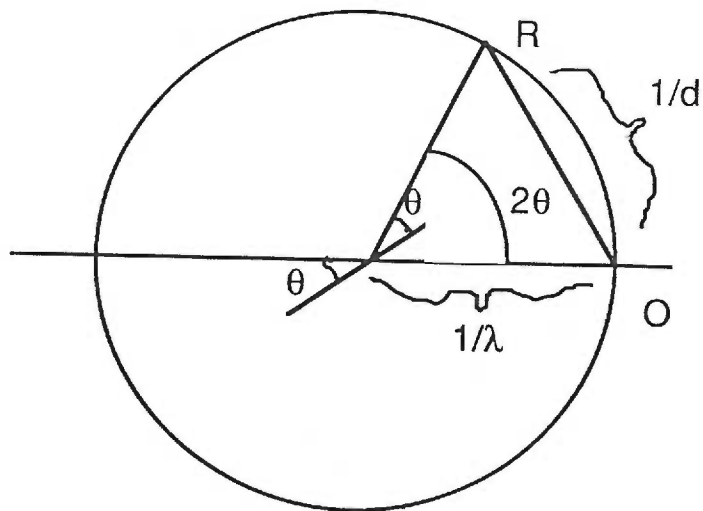


MetJ

Figure 14. A) Bragg's law. B) Ewald's sphere.

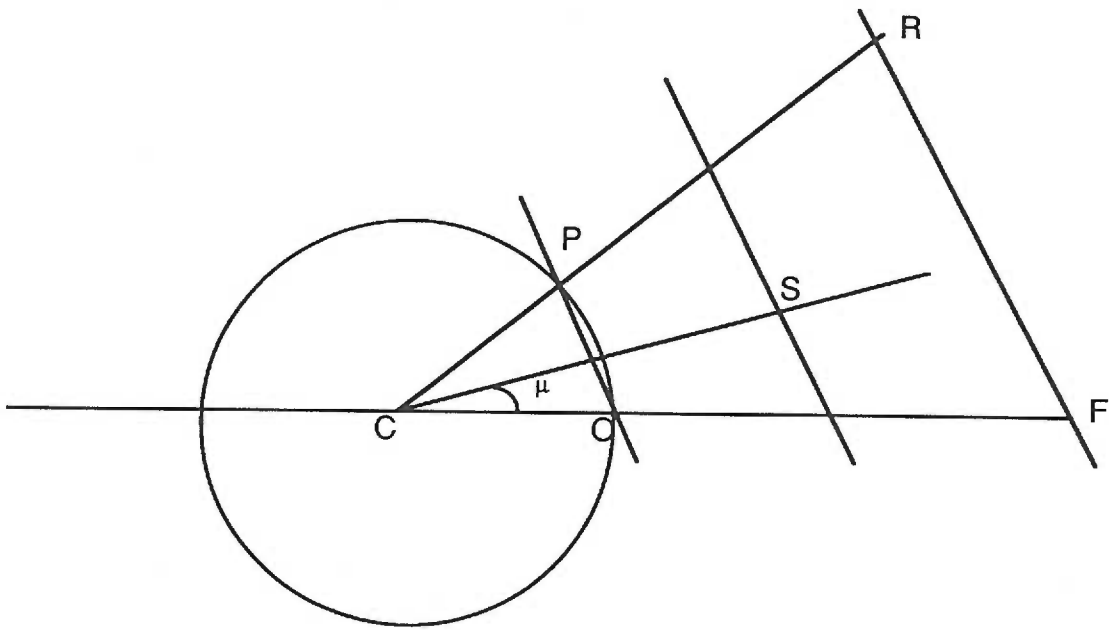


A. Bragg's Law :
 $2d\sin\theta = n\lambda$



B. Ewald's Sphere

Figure 15. Reciprocal lattice separations by similar triangles.



$$RF/CF = PO/CO$$

Figure 16. Eulerian angles α, β, γ used in the rotation function to relate the rotated axes X_1', X_2', X_3' to the original unrotated axes, X_1, X_2, X_3 .

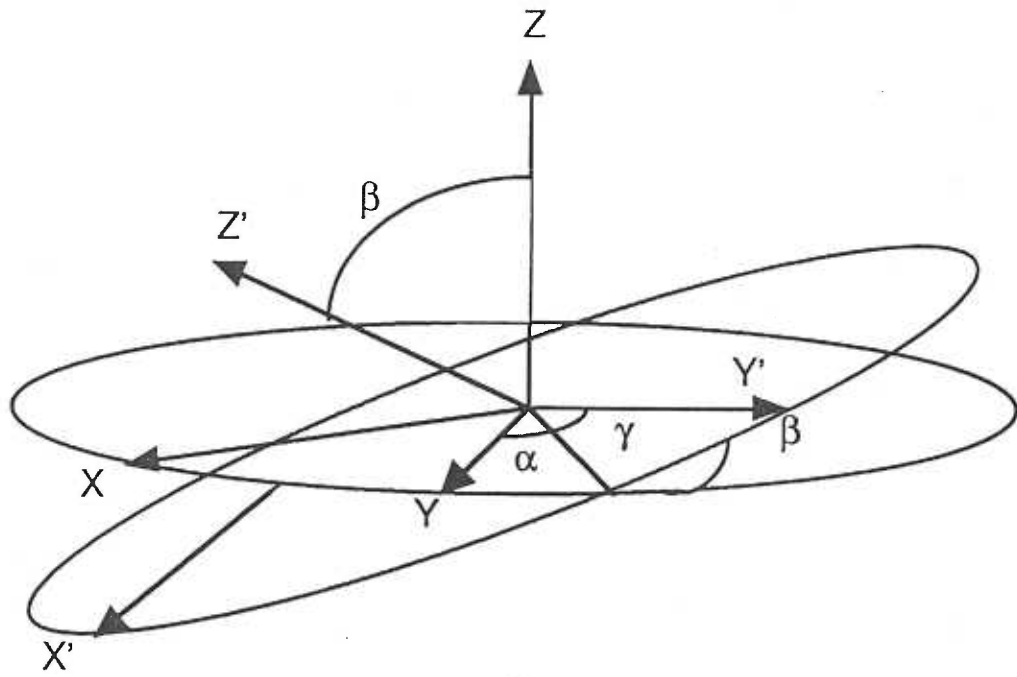
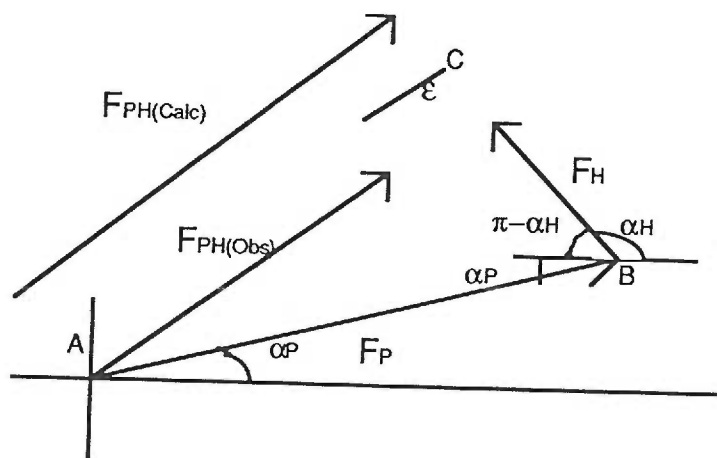
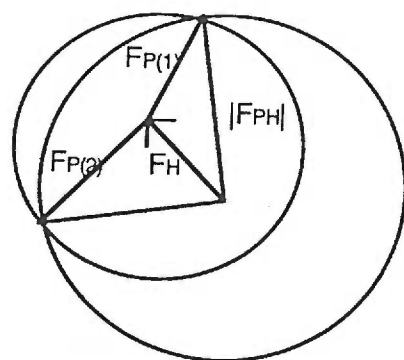


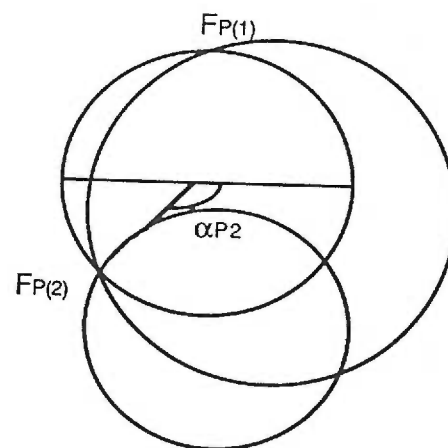
Figure 17A) Diagram showing the vector nature of F_H , F_{PH} and F_P and the lack of closure error. 17B) Harker construction showing resulting phase ambiguity from SIR (Single Isomorphous Replacement) and its resolution by MIR (Multiple Isomorphous Replacement).



A. "Lack of Closure"



B.
Harker construction
showing phase
ambiguity



Harker construction
showing how phase
ambiguity is resolved by
second derivative giving
phase angle = α_{P2}

Manuscript #1**STRUCTURAL ANALYSIS OF THE PURINE REPRESSOR,
AN *Escherichia coli* DNA-BINDING PROTEIN**

Maria A. Schumacher¹, J. Randy MacDonald¹, Joakim Björkman²,
Sherry L. Mowbray², and Richard G. Brennan^{1*}

¹Department of Biochemistry and Molecular Biology
Oregon Health Sciences University
Portland, Oregon 97201-3098

²Department of Molecular Biology
Swedish University of Agricultural Sciences
Uppsala Biomedical Center, Box 590, Uppsala, Sweden

*To Whom Correspondence Should Be Addressed.

ABSTRACT

The purine repressor, PurR is a member of a superfamily of *E. coli* DNA-binding proteins, including LacI, the lac repressor, which are purported to bind DNA by an NH₂-terminal helix-turn-helix domain. The members of this family also display strong sequence homology within their larger COOH-terminal effector binding/oligomerization domains. Analysis of the PurR amino acid sequence and secondary structure reveals that this protein is highly homologous to another group of *E. coli* binding proteins, namely the periplasmic binding proteins, especially to RBP (the ribose binding protein) and GGBP (the D-glucose/D-galactose binding protein). The high-resolution X-ray structure of RBP allows this protein to serve as a template with which to model the secondary structure of the corepressor binding domain of PurR. Similarly, PurR's NH₂-terminal DNA-binding domain can be modelled using the structure of the corresponding NMR-determined region (residues 1-59) from LacI as a template. Combining the two, results in a complete description of the secondary structure topology of PurR and implicates residues important for corepressor binding and dimerization. CD spectroscopic studies on PurR, its corepressor binding domain and RBP result in secondary structure estimates nearly identical with those obtained by sequence analyses, thereby providing further corroborating physical evidence for this topological assignment.

INTRODUCTION

In *E. coli*, the genes encoding enzymes required for purine biosynthesis are scattered throughout the chromosome in the form of polycistronic and monocistronic operons. Ten of these genes are regulated by a 38 kDa DNA-binding protein, the purine repressor or PurR. In addition, PurR also participates in the regulation of five other genes involved in pyrimidine biosynthesis, pyrimidine salvage and the generation of one carbon units (Zalkin & Dixon, 1992). Finally, PurR is itself autoregulated (Rolfes & Zalkin, 1990a). PurR is activated to bind its cognate DNA by binding of corepressors, hypoxanthine or guanine. This leads to repression of the Pur regulon (Rolfes & Zalkin, 1990b).

PurR is a member of the LacI family, which includes the cytidine repressor protein (Valentin-Hansen et al., 1986), galactose repressor protein (von Wilcken-Bergmann and Müller-Hill, 1982), ribose repressor protein (Mauzy & Hermodson, 1992a), maltose repressor protein (Reidl et al., 1989), raffinose repressor protein (Aslanidis & Schmitt, 1990), evolved β -galactosidase repressor protein (Stokes & Hall, 1985), fructose repressor protein (Henkin et al., 1991) and fructose repressor protein (Jahreis et al., 1991; Vartak et al., 1991). Proteins in this family contain two separate functional domains, a smaller NH₂-terminal HTH DNA-binding domain (Brennan & Matthews, 1989) and a larger COOH-terminal effector binding/oligomerization domain. LacI is unique in this family in that it exists as a tetramer (Riggs & Bourgeois, 1968; Barkley et al., 1975; Culard & Maurizot, 1981; Whitson & Matthews, 1986) whereas the other members appear to exist only as dimers. Specific proteolytic cleavage of PurR after Arg52 provides support for this two domain organization within PurR. This cleavage produces an NH₂-terminal DNA-binding domain (residues 1-52) and a core corepressor binding domain (residues 53-341), CBD or Clipped PurR (Choi & Zalkan, 1992).

Homology between this family of DNA-binding proteins and another group of *E. coli* binding proteins, the periplasmic proteins, has been noted (Müller-Hill, 1983; Vartak et al., 1991; Mauzy & Hermodson, 1992b; Weickert & Adhya, 1992). This homology exists despite lack of significant sequence identity. The periplasmic proteins (PBPs) are also structurally organized into two domains, consisting of a small NH₂-terminal signal-sequence, which targets these proteins to the bacterial inner membrane and is subsequently cleaved, and a larger COOH-terminal metabolite-binding domain. Importantly, the structures of several of these proteins have been solved to high resolution by X-ray crystallography, showing them to have a conserved a/b type structure (Quijcho, 1991). Strong homology between the periplasmic proteins and the LacI superfamily is found between the effector binding domains of the LacI members and the metabolite-binding

domains of the periplasmic proteins. Because the structure of several of the periplasmic proteins is known, they can potentially be used as scaffolds upon which to model the structures of the effector binding domains of LacI members. To further our structural understanding of PurR, we have carried out computer aided sequence homology studies, secondary structure analyses and CD spectroscopic studies on PurR, its corepressor binding domain and RBP. A complete secondary structure topology of PurR is presented and its functional ramifications discussed.

EXPERIMENTAL PROCEDURES

Sequence Homology - A sequence identity comparison program (Ohlendorf et al., 1983; Brennan et al., 1984) was used to analyze the primary sequences of GGBP, RBP, PurR, LacI and CytR. The secondary structure prediction of PurR was carried out using the method of Wilmont and Thorton (Wilmont and Thorton, 1988) and sequence alignments were made based on the algorithm of Feng and Doolittle (Feng and Doolittle, 1990).

The final "knowledge-based" alignment of PurR was carried out in five steps. First, as a control, the secondary structures of GGBP and RBP were predicted using the above described methods and compared to the known secondary structures from X-ray data. The secondary structure of PurR was similarly predicted by this method. The three sequences were subsequently aligned employing the algorithm of progressive alignments of Feng and Doolittle. The final alignment of PurR was refined and confirmed by calculating the significance of the alignment via direct amino acid comparisons (DAAC) and minimum base change per codon (MBC/C) comparisons between eight variable-length stretches of RBP and CBD (Ohlendorf *et al.*, 1983; Brennan *et al.*, 1986) and by using the recent sequence alignment of RBP against the *E. coli* and *Salmonella typhimurium* GBP (Mowbray, 1992). To complete the topological analysis, the NMR-determined secondary structure of the LacI headpiece region (Kaptein *et al.*, 1985) was used to model the secondary structure of the NH₂-terminal region of PurR.

CD Spectroscopy - CD spectra of CBD, PurR and RBP were taken on a JASCO J-500A spectrophotometer. Measurements were made using a 0.1 mm path length cell (Helma) thermostatted cell at room temperature. The instrument was calibrated by using (+)-10-camphorsulfonic acid ($\Delta\epsilon = +2.37 \text{ M}^{-1}\text{cm}^{-1}$ at 290.5 nm and -4.95 at 192.5 nm). Data were collected on an IBM/PC-XT using the IF-500 interface and software provided by JASCO. Spectra and buffer base lines were the average of four to eight scans each recorded at 0.1-nm intervals, using a scanning rate of 5nm/min and a 4-s time constant taken at room temperature. The buffer used for the spectral measurements of CBD and RBP was 10 mM potassium phosphate, pH 7.5, and because of the decreased solubility at low ionic strength, 100 mM potassium phosphate, pH 7.5, for PurR. All proteins were purified as described previously (Choi and Zalkin, 1992; Mowbray and Cole, 1992). The protein concentrations were determined by amino acid analysis and were ~0.5-1.0 mg/ml. Before spectral deconvolution for secondary structure analysis, the buffer base line was subtracted, and the resulting spectrum was smoothed using the smoothing program

provided by JASCO. The CD spectra for each protein were deconvoluted for secondary structure content using the singular-value and variable selection methods described in detail elsewhere (Compton *et al.*, 1987). All secondary structure values resulting from each combination which met these criteria were averaged to give the final secondary structure values for each experimental spectra.

RESULTS AND DISCUSSION

Corepressor Binding Domain - The analysis of PurR was undertaken to establish the extent of this protein's homology to the PBP and to determine which PBP family it most closely resembles. The periplasmic binding proteins contain several hallmark features that are highly conserved among members (Spurlino *et al.*, 1991). All mature PBP, *i.e.*, after signal sequence cleavage, are divided into two structural domains, an N terminus-containing domain and a COOH-terminus containing domain. Especially striking is the finding that in all periplasmic binding proteins, the x-ray structures of which have been determined, the first six secondary structural elements, which comprise approximately 100 residues, have the identical topological arrangement, $\beta_A\alpha_I\beta_B\alpha_{II}\beta_C\alpha_{III}$ (Quioco, 1991; Mowbray and Cole, 1992). The N- and C- terminal domains are connected by two to three peptide crossovers, which although not sequentially close, are structurally close. These peptide crossovers have been used in to delineate members of the PBPs into two families and indicate the secondary structures being connected. The first family, the ABP family, includes RBP, ABP, GBP-S and GBP-E and exhibits $\beta \rightarrow \alpha$ crossovers for the first two crossovers and a $\beta \rightarrow \beta$ crossover for the third. The second family, the SBP family, includes SBP and maltose-binding protein and exhibits $\beta \rightarrow \beta$ crossovers for the first two crossovers and an $\alpha \rightarrow \alpha$ crossover for the last.

The structural analysis of PurR commenced by initially focusing on the corepressor binding domain. Primary and secondary structure analyses and comparisons were carried out against selected PBP, the high resolution x-ray structures of which are known, and followed by DAAC and MBC/C significance analyses (Fig. 1 and Table 1). Interestingly, the secondary structure predicted for the CBD is that of the repeating β/α structure, with similarly located and sized helices and β -strands as seen in the PBP (Fig. 1). On the basis of all independent and corroborating methods used, it became evident that the first six secondary structural elements of the CBD follow the $\beta_A\alpha_I\beta_B\alpha_{II}\beta_C\alpha_{III}$ topology of the PBP. These elements begin with residue 61 of PurR and end at residue 137. Thus, they are contained within 100 residues as observed in the PBP. Secondary structural analysis reveals a second similar COOH-terminal motif which spans $\beta_F\alpha_V\beta_G\alpha_{VI}\beta_H\alpha_{VII}$. An analogous β/α motif is found in the PBP as well. Interestingly, self-alignment of CBD sequences 61-183 against 184-318 reveals little sequence identity ($\sim 12\%$). This is consistent with similar internal comparisons made between domains within the PBP in which sequence identities are only $\sim 14\%$. Indeed, sequence identity was shown to be greater between the corresponding domains of RBP and GBP-E ($\sim 24\%$) (Mowbray, 1992) and RBP and CBD ($\sim 24\%$) (this study).

To determine which family of periplasmic binding proteins PurR most closely resembles, the potential crossover regions need to be identified. This presented a problem as the secondary structure prediction data were ambiguous in the regions in which the crossovers were likely to occur. This problem was first addressed by studying the alignment data. Here it was noticed that, whereas the sequence of PurR aligns well with SBP family members within the NH₂-terminal subdomain (data not shown), it aligns well with ABP family members, especially GBP-E and RBP, throughout its entire sequence, including the crossover regions (Fig. 1). For example, the helix in the first $\beta \rightarrow \alpha$ crossover in RBP and GBP-E contains several highly conserved residues. In particular, residues Gly¹⁰⁹, Gly¹¹³ and Ile¹¹⁶ of RBP correspond to Gly¹¹⁶, Gly¹²⁰ and Ile¹²³ of GBP-E. PurR has the corresponding glycines at positions 166 and 170 and a conservative change at position 173, a leucine for isoleucine (Fig. 1). Further comparison of PurR with RBP reveals even more striking identity within this region in that Gly¹⁶⁵, Ala¹⁶⁹ and Tyr¹⁷² of PurR correspond directly to Gly¹⁰⁸, Ala¹¹² and Tyr¹¹⁵ of RBP. Additionally, several other conservative substitutions within the RBP, GBP-E and PurR sequences are found within this region (Fig. 1). Together, these identities strongly implicate this region as being the first crossover helix in PurR. Further corroborating this assignment is the secondary structure analysis which strongly predicts residues 162-176 of PurR to be helical (Fig. 1). On the basis of this assignment, PurR belongs to, or is closely related to, members of the ABP family. Therefore, a second $\beta \rightarrow \alpha$ crossover should be located near residue 292 in PurR after β -strand J (Fig. 1). In accordance with this supposition CBD residues 300-311 are predicted to be helical.

The analysis of PurR's third crossover was aided greatly by the recent x-ray structure of RBP (Mowbray and Cole, 1992). In that structure, the third crossover is $\beta \rightarrow \beta$, from β -strand K (residues 260-263) to β -strand L (residues 266-268) (Fig. 1). Corresponding PurR residues 318-321 and 324-326 are predicted to be β -strands by alignment and sequence identity data (Fig. 1). Additional evidence implicating PurR residues 318-326 as the third crossover region is the near identity between the sequences of residues 324-326 of PurR 266-268 of RBP, respectively. In RBP, these residues are Lys-Leu-Val whereas in PurR the corresponding residues are Arg-Leu-Ile. This β -strand, L, in RBP forms an antiparallel β -sheet with β -strand J. β -strand J is also highly similar between the two proteins; Ala-Thr-Ile-Ala in RBP and Thr-Thr-Ile-His in PurR. The finding that PurR exhibits such similarity to RBP in these two distant regions, which form the only antiparallel β -sheet in RBP, suggests strongly that it contains the same secondary structure.

Further support for CBD's striking similarity to RBP and other ABP family members comes from the recent structural comparison of the high resolution structures of RBP,

GBP-S and GBP-E (Mowbray, 1992). This study identified several conserved residues as being key structural elements, whereby the regions of highest sequence identity are those involved in forming the hydrophobic cores of these proteins. Remarkably, these same regions are also the most conserved between RBP and PurR (Fig. 1). For example, PurR residues Ile⁶², Leu⁹², Leu¹²⁰, Met²³⁰ and Leu²³⁴ correspond directly to RBP residue Ile⁴, Leu³⁴, Leu⁶², Met¹⁷³ and Leu¹⁷⁷, respectively. Importantly, these identities span both the N- and COOH-terminal domains. Conserved residues in RBP and GBP-E were also noted as forming unusual and highly specific interactions within these proteins. For example, Asp¹⁹¹ in RBP and Asp²¹² in GBP-E, located within helices adjacent to a ligand binding site residue, are buried within the protein. Critical to the stabilization of this buried charge are hydrogen bonds to main chain amide nitrogens from a nearby loop and to the side chain of Thr²³² in RBP and Thr²⁵³ in GBP-E. The presence of this interaction in PurR is strongly implicated by the presence of the corresponding pair Asp²⁴⁸ and Thr²⁸⁹. Furthermore, Asp²⁴⁸ is predicted to be within a helix (Fig. 1). An unusual interaction in RBP is a 1-3 hydrogen bonding interaction involving residues 88-90, which contains a central aspartate residue, Asp⁸⁹, that is also involved in sugar binding. The equivalent position in PurR is Asp¹⁴⁶, pointing out again the extraordinary conservation between RBP and PurR in residues known to be structurally important in RBP, despite the lack of significant global sequence identity between these two proteins. The main differences between RBP and GGBP have been noted as occurring in regions where GGBP binds Ca²⁺ (Mowbray et al., 1990; Vyas et al., 1991). Since RBP does not bind Ca²⁺ the corresponding sequences in RBP contain deletions within the alignment (Mowbray and Cole, 1992). Gaps also occur in these regions for PurR, indicating, as expected, that there are no Ca²⁺ binding sites within PurR (Fig. 1).

To provide statistical verification for the remarkable similarity between RBP and CBD, DAAC and MBC/C analyses were carried out by comparing various stretches of RBP against the entire sequence of CBD (Ohlendorf *et al.*, 1983; Brennan *et al.*, 1986). The results of DAAC and MBC/C comparisons between eight variable length segments of RBP and the entire CBD are shown in Table 1 and reveal overall 27.3% amino acid sequence identity. These segments were chosen around the few gaps and insertions that were necessary to align the two sequences optimally. Overall, in both DAAC and MBC/C analyses, the better scores are those obtained from comparisons between the NH₂-terminal regions, with the lower scores corresponding to the COOH-terminal region. The significance scores (Table 1, fourth column) for the DAAC analysis for the eight aligned pairs of sequences of RBP versus CBD are 4.21, 4.10, 4.56, 3.96, 5.07, 5.23, 3.86 and 3.9 and underscore the very strong homologies between these segments. Remarkably, the

best scores were found in regions in which the predicted secondary structure of PurR matches exactly the known secondary structure of RBP. The values of the given MBC/C also support this homology with significance values of 4.72, 3.62, 3.78, 4.21, 3.00, 2.66, 3.39 and 2.88 (Table 1, sixth column). However, in the MBC/C analyses, two segments of RBP (residues 191-230 and 251-258) score better with regions of PurR other than the aligned stretches shown in Fig.1. The former segment, which contains the most divergent residues between RBP and PurR, residues 218-230 and 275-287, respectively, corresponds to a region in LacI and CytR which has been implicated in dimerization (Daly and Matthews, 1986; Chakerian and Matthews, 1991; Barbier and Short, 1992; Weickert and Adhya, 1992) (Fig. 2). This being the case, one might expect this region to diverge structurally between PurR and the monomeric PBP. In accordance with this supposition, PurR residues 275-287 are predicted to be aperiodic unlike the corresponding regions in RBP and GBP-E which form helix VIII.

The above analyses have been combined to generate the alignment seen in Fig.1. It should be emphasized that this is the best alignment achievable based on predictive and comparative methods, and therefore the beginning and ends of each predicted structural element are the most subject to error. Only with the x-ray structure in hand will the exact secondary, tertiary and quaternary structures be known. However, it can be seen that the predicted secondary structural elements of CBD align very well with those observed in the x-ray structures of RBP and to a significant, but lesser extent, with those of GBP-E, thereby providing strong support that these proteins are structurally similar.

Ligand Binding Residues - A remarkable feature displayed by RBP and GBP-E, as well as other PBP, is the conservation of those residues involved in ligand binding, despite the differences in ligand specificity (Vyas *et al.*, 1991; Mowbray and Cole, 1992). Specificity is attained almost entirely by direct protein-ligand hydrogen bonding. However, the RBP and GBP-E ligand complexes are further stabilized by stacking interactions (Vyas, 1991; Mowbray and Cole, 1992). This type of interaction is expected to be even more significant in PurR's interaction with its planar, aromatic ligands, hypoxanthine and guanine. The residues of the PBP involved in ligand binding are widely dispersed throughout their primary structures making it more difficult to predict the functionally analogous residues in PurR. However, the alignment of PurR with RBP and GBP-E clearly suggests residues that may form, in part, the ligand binding pocket. That is to say, although specific PurR-ligand contacts cannot be predicted, by virtue of locating probable loop locations, the site of PBP-ligand interaction, certain PurR residues can be implicated in ligand binding.

Specific examples include the three polar groups that have been shown to be critical in protein-ligand hydrogen bonding interactions for GBP-E and RBP (Vyas, 1991; Mowbray

and Cole, 1992). In RBP these are Asp⁸⁹, Arg¹⁴¹ and Asp²¹⁵, and Asn⁹¹, Arg¹⁵⁸ and Asp²³⁶ in GBP-E (Table II). The corresponding PurR residues are Asp¹⁴⁶, Arg¹⁹⁶ and Asp²⁷⁵. In RBP and GBP-E these amino acids make extensive cooperative and bidentate hydrogen bonds with other protein residues as well as the ligand. Stacking interactions have also been shown to be important in RBP and GBP-E, acting to "sandwich" the sugar ligands. For the purine repressor with its planar, aromatic ligands, hypoxanthine and guanine, such stacking interactions will likely be of greater importance in the formation of a stable protein-ligand complex. Indeed, residues Tyr⁷³ and Phe²²¹ clearly correspond to stacking residues Phe¹⁶ and Trp¹⁸³ in GBP-E and Phe¹⁵ and Phe¹⁶⁴ in RBP (Fig. 1). Phe¹⁶ of RBP is also involved in stacking and an analogous stacking interaction could be made by the homologous PurR residue, Phe⁷⁴. Another important PBP-ligand interaction is the hydrogen bond formed between Arg⁹⁰ (RBP) and Lys⁹² (GBP-E) and their respective carbohydrates. Interestingly, the corresponding residue in PurR is Trp¹⁴⁷. It is possible that in PurR this Trp provides an additional stacking interaction or is involved in ligand hydrogen bonding through its indole ring nitrogen.

Other residues in PurR, which correspond to ligand binding residues in GBP-E and RBP, include Ala⁷¹ and Gly²⁴⁷. In GBP-E and RBP, the corresponding residues are polar and planar, either aspartates or asparagines, and function in hydrogen bonding and van der Waals interactions. It is possible that in PurR, these small, nonpolar residues exist in regions within the binding cleft where small residues are needed to accommodate the larger, more hydrophobic, planar purines. A summary of possible ligand binding residues is presented in Table II.

DNA Binding Domain - Upon corepressor binding to the effector binding domain, the DNA binding domain of PurR is activated to bind to its cognate DNA (Rolfes and Zalkin, 1990b). As described above, members of this family display a high degree of sequence homology throughout their sequences with their DNA binding domains particularly well conserved (Weickert and Adhya, 1992). The proposed DNA-binding domain of these proteins is one such region. Sequence alignment of PurR with LacI reveals almost 50% sequence identity within this region (residues 1-60) (Fig. 2). NMR studies on the LacI DNA binding domain (residues 1-51) confirm the presence of a helix-turn-helix within residues 6-25 and also identify a third helix in residues 34 to 45 after an extended loop (Kaptein et al., 1984). The corresponding residues of the DNA binding domain of PurR, residues 4-23 and 32-43, respectively, display greater than 55% sequence identity and 68% similarity, strongly suggesting the presence of a similar helix-turn-helix-loop-helix motif. An additional 9-residue stretch in PurR, residues 52-60, is predicted to contain a small

helix. However, no structural data have been presented for this region in any LacI family member.

Subunit Interactions - PurR, like most members of the LacI family and unlike members of the periplasmic proteins, exists as a dimer (Choi and Zalkin, 1992). Those residues involved in the subunit interface can be surmised, in part, by previous mutagenesis on LacI and CytR, which implicated residues 269-291 and residues 288-310, respectively, as critical in subunit interaction (Daly and Matthews, 1986; Chakerian and Matthews, 1991; Barbier and Short, 1992). The sequences in these regions diverge the most from the PBP and are predicted to be loops instead of a helix as observed in the PBP (Vyas *et al.*, 1991; Mowbray, 1992). However, this region does display significant sequence identity among all LacI members, over which PurR, LacI and CytR are ~60% identical (Fig. 2). The sequence homology and predicted structural differences between the PBP and LacI proteins are consistent with this region's involvement in oligomerization.

Support for this assertion is provided by the observation that one residue within this region, Tyr²⁸² in LacI and Cys²⁸⁹ in CytR are *essential* for dimerization in these proteins (Chakerian and Matthews, 1991; Barbier and Short, 1992). Mutation of of LacI Tyr²⁸² to any amino acid other than phenylalanine or leucine abolishes dimerization, implying that a large hydrophobic residue is required in this position for effective subunit interaction (Chakerian and Matthews, 1991). In PurR, the corresponding residue, Phe²⁸³, is also a large hydrophobic residue and presumably plays a like role. Similarly, substitutions for Cys²⁸⁹ in CytR results in the production of only monomeric protein (Barbier and Short, 1992). In LacI, Cys²⁸¹ has also been implicated as being important in subunit interaction. Substitution of of Cys²⁸¹ with virtually any amino acid does not affect dimerization but does influence inducer affinity and cooperativity (Chakerian and Matthews, 1991). It is interesting to speculate that the corresponding residue in PurR and CytR, Tyr²⁸² and Phe²⁸⁸, respectively, play analogous roles in ligand binding and subunit cooperativity and that their aromatic nature reflects their aromatic ligands.

CD Analysis - The above described primary and secondary structure analyses strongly suggest that the corepressor binding domain of PurR has a α/β type structure and fold very similar to that of RBP. When combined with NMR studies on the LacI DNA binding domain, these analyses also clearly indicate that PurR's NH₂-terminal DNA binding domain has a closely related helix-turn-helix-loop-helix structure. However, they provide no direct physical evidence. Perhaps one of the best physical methods available for the examination of protein structure, exclusive of x-ray crystallography and NMR, is circular dichroism spectroscopy. By measuring primarily the amide chromophore, CD spectroscopy is exquisitely sensitive to a protein's secondary structure (Johnson, 1990).

If, indeed, the CBD and RBP have similar secondary structures, their CD spectra should reveal this. Additionally, the intact form of PurR should also produce a similar spectrum, albeit with a greater helical content than the CBD and a slightly lower helical content than RBP. A greater helical content is reflected by a more positive absorbance at 192 nm and a more negative absorbance at 220 nm, two wavelengths that are particularly sensitive to a protein's helical content. Therefore, CD spectroscopic studies were undertaken on RBP and both the intact form and the CBD of PurR.

The CD spectra of RBP, CBD and PurR are shown in Fig. 3. The known or calculated percentages of the various secondary structural elements present in all three proteins, taken from either the x-ray crystallographically determined structure (RBP) or the optimum alignment (CBD and PurR) are presented in Table III as are those values for each protein calculated after deconvolution of its CD spectrum. These results correlate extremely well with the predicted values for the CBD given that the correlation coefficients for the proteins of CD versus x-ray are 0.97 for α -helix, 0.76 for β -sheet, 0.49 for β -turn and 0.86 for other (Johnson, 1990). The close match between the amount of known secondary structure of RBP as determined from the x-ray structure and that calculated by CD analysis indicates that the CD spectral analysis is a very reliable method for determining the secondary structure content of proteins with similar PBP folds and bolsters the sequence analyses of CBD and PurR. Further support for the PurR structural prediction was provided by CD data for the intact PurR protein, which shows 40% helix compared with the predicted value of 39%. As discussed above, PurR should show greater helix content than the CBD because of the four additional helices that are predicted to lie within the NH₂-terminal DNA binding domain.

Although the amount of α -helix predicted for PurR agrees well with that determined by CD analysis, the match for β -sheets and aperiodic structures is weaker. A plausible explanation for this latter discrepancy, aside from the fact that these structures are not predicted as reliably with CD, can be offered in light of the nature of PurR's tertiary and quaternary structures. It is possible that the tertiary contacts between the DNA binding domain and corepressor binding domain as well as quaternary contacts between subunits of the intact PurR dimer lead to a more ordered structure in the region encompassing the domain-domain or monomer-monomer contact points than is predicted by the algorithm employed in this study, which only examines the linear sequence.

CONCLUSIONS

A diagram describing the purine repressor protein's deduced secondary structure topology is presented in Fig. 4. Both predictive and physical methods indicate that the structure of PurR's corepressor binding domain is very similar to the β/α structure observed for the metabolite binding domains of the PBP, especially RBP, and that PurR's DNA binding domain assumes the structure of the helix-turn-helix-loop-helix DNA binding domain of LacI. A similarly conserved topology is anticipated for all members of the LacI family. Significantly, although the intact LacI protein (Pace *et al.*, 1990) and the LacI core (Steitz *et al.*, 1980) have been crystallized, no inducer binding domain of the LacI superfamily has been solved. The high resolution x-ray structure determination of the corepressor binding domain of PurR (in progress) could serve as a potential model for the other inducer binding domains of the LacI proteins (Schumacher *et al.*, 1992). Furthermore, this structure will allow the comparison of the predicted secondary structure of PurR's corepressor binding domain from sequence analyses and the CD analysis with that found in the CBD's crystal structure. It is anticipated that CBD will have a structure very similar to CBP-E and RBP. However, certain regions, in particular the purported dimerization domain, will likely be different but conserved among the LacI family members. Knowledge of those subunit interactions that affect dimerization are essential to understanding not only how and where dimerization occurs, but should shed considerable light on the biochemistry and dynamics of cooperativity in ligand binding by PurR.

REFERENCES

- Aslanidis, C., and Schmitt, R. (1990) *J. Bacteriol.* **172**, 2178-2180.
- Barbier, C.S., and Short, S.A. (1992) *J. Bacteriol.* **174**, 2881-2890.
- Barkley, M.D., Riggs, A.D., Jobe, A., And Bourgeois, S. (1975) *Biochem.* **14**, 1700-1712.
- Bowie, J.U., LHthy, R., and Eisenberg, D. (1991) *Science* **253**, 164-170.
- Brennan, R.G., and Matthews, B.W. (1989) *J. Biol. Chem.* **264**, 1903-1906.
- Brennan, R.G., Weaver, L.H., and Matthews, B.W. (1986) *Chem. Scripta* **26B**, 251-255.
- Brenowitz, M., Mandal, N., Pickar, A., Jamison, E., and Adhya, S. (1991) *J. Biol. Chem.* **266**, 1281-1288.
- Chakerian, A.E., and Matthews, K.S. (1991) *J. Biol. Chem.* **266**, 22206-22214.
- Chen, J., and Matthews, K.S. (1992) *J. Biol. Chem.* **267**, 13843-13850.
- Choi, K.Y., and Zalkin, H. (1990). *J. Bacteriol.* **174**, 6207-6214.
- Compton, L.A., Mathews, C.K. and Johnson, W.C., Jr. (1987) *J. Biol. Chem.* **262**, 13039-13043.
- Daly, T.J., and Matthews, K.S. (1986) *Biochem.* **25**, 5474-5478.
- Feng, D.F., and Doolittle, R.F. (1990) *Methods Enzymol.* **183**, 375-387.
- Henkin, T.M., Grundy, F.J., Nicholson, W.L., and Chambliss, G.H. (1991) *Mol. Microbiol.* **5**, 575-584.
- Jahreis, K.P., Postma, W., and Lengeler, J.W. (1991) *Molec. Gen. Genet.* **226**, 332-336.
- Johnson, C.W. (1990) *Protein Struct. Funct. Genet.* **7**, 205-214.
- Kaptein, R., Zuiderweg, E.R., Scheek, R.M., Boelens, R., and van Gunsteren, W.F. (1985) *J. Mol. Biol.* **182**, 179-182.
- Matthews, K.S. (1979) *J. Biol. Chem.* **254**, 3348-3353.
- Mauzy, C.A., and Hermodson, M.A. (1992a) *Prot. Sci.* **1**, 831-842.
- Mauzy, C.A., and Hermodson, M.A. (1992b) *Prot. Sci.* **1**, 843-849.
- Mowbray, S.L. (1992) *J. Mol. Biol.* **227**, 418-440.
- Mowbray, S.L., and Cole, L.B. (1992) *J. Mol. Biol.* **225**, 155-175.
- Mowbray, S.L., Smith, R.D., and Cole, L.B. (1990) *Receptor* **1**, 41-54.
- Müller-Hill, B. (1983) *Nature* **302**, 163-164.
- Quioco, F.A. (1991) *Curr. Opin. Struct. Biol.* **1**, 922-933.
- Ohlendorf, D.H., Anderson, W.F., and Matthews, B. W. (1983) *J. Mol. Evol.* **19** 109.

- Pace, H.C., Lu, P., and Lewis, M. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 1870-1873.
- Reidl, J., Karin, R., Ehrmann, M., and Boss, W. (1989) *J. Bacteriol.* **171**, 4888-4899.
- Richardson, J.S., and Richardson, D.C. (1988) *Science* **240**, 1648-1652.
- Riggs, A.D., and Bourgeois, S. (1968) *J. Mol. Biol.* **34**, 361-364.
- Rolfes, R.J., and Zalkin, H. (1990a) *J. Bacteriol.* **172**, 5758-5766.
- Rolfes, R.J., and Zalkin, H. (1990b) *J. Bacteriol.* **172**, 5637-5642.
- Schumacher, M.A., Choi, K.Y., Zalkin, H., and Brennan, R.G. (1992) *J. Mol. Biol.* **225**, 1131-1133.
- Spurlino, J.C., Lu, G-Y., and Quioco, F.A. (1991) *J. Biol. Chem.* **266**, 5202-5219.
- Steitz, T.A., Stenkamp, R.E., Geisler, N., Weber, K., and Finch, J. (1980) *Biomol. Struct. Conform. Funct. & Evol.* **1**, 425-430.
- Stokes, H.W., and Hall, B.G. (1985) *Mol. Biol. Evol.* **2**, 478-483.
- Valentin-Hansen, P., Larsen, J.E.L., Hojrup, P., Short, S.A., and Barbier, C.S. (1986) *Nucl. Acids Res.* **14**, 2215-2228.
- Vartak, N.B., Reizer, J., Reizer, A., Gripp, J.T., Groisman, E.A., Wu, L-F., Tomich, J.M., and Saier, Jr., M.H. (1991) *Res. Microbiol.* **142**, 951-960.
- von Wilcken-Bergmann, B., and Müller-Hill, B. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 2429-2431.
- Vyas, N.K. (1991) *Curr. Opin. Struct. Biol.* **1**, 732-740.
- Vyas, N.K., Vayas, M.N., and Quioco, F.A. (1991) *J. Biol. Chem.* **266**, 5226-5237.
- Weickert, M.J., and Adhya, S. (1992) *J. Biol. Chem.* **267**, 15869-15874.
- Whitson, P.A., and Matthews, K.S. (1986) *Biochem.* **25**, 3845-3852.
- Whitson, P.A., Olson, J.S., and Matthews, K.S. (1986) *Biochem.* **25**, 3852-3858.
- Wilmot, C.M., and Thornton, J.M. (1988) *J. Mol. Biol.* **203**, 221-232.
- Zalkin, H., and Dixon, J.E. (1992) *Progr. in Nucleic Acid Res. and Mol. Biol.* **42**, 259-287.

Figure 1. Alignment of PurR against RBP and GBP-E. This alignment was carried out using the following steps: 1) secondary structure prediction of PurR; 2) DAAC and MBC/C analyses of PurR against PBP of known structure; and 3) alignment of PurR against the PBP. The final refined alignment is based on the above results of DAAC and MBC/C data stretches of RBP against the entire sequence of CBD. The predicted secondary structure of CBD is shown below the sequence alignment.

FIGURE 1.

Puer: 1 MATIKDVAKRANVSTTTVSHVINKTRFVAZEETRNAVWAAIKELHYSPSAVARSLKVNHTKSI^{(β)A}GLLA
RBP: 1 -----KDTIALVY^{(β)A}
OBP-E: 1 -----ADIRIGVTI^{(β)A}

Puer: 67 TSSZAAYFAEIIIEAVEKNCFL - GYLLILGNHANNLEKQRAYLLSMHAKRVDGLLVMCSEYPE^{(β)C}
RBP: 8 STLNPFVSLKDGAKKEADKL - GYNLVVLDSONNPAKELANVQDLTVRGTIKILLIMPTDSDA^{(β)C}
OBP-E: 10 KYDDNFMVVRKAIIEODAKAAPDVO^{(β)B}LLMNDSONDQSKONDQIDVLLAKGVKALAINLVDPAA^{(β)C}

Puer: 128 PLLAM-LEERYRHIPMVVMDWGEAK - AD - PT - - DAVI^{(β)D}DNAFE^{(β)E}GGYMA^{(β)IV}GRYLIERG - - - -
RBP: 71 VGNAYKMANQANIPYITLDRQAT - KG - - EV - - VSHIAS^{(β)E}D^{(β)E}VLGGK^{(β)IV}IAGDYIAKKA - - - -
OBP-E: 73 AGTVIEKARGQNVFVVEFNKPEPSEKALDSYDKAYVVGID^{(β)D}SKESGI^{(β)E}IQGDLIAKHWAANQOWD^{(β)IV}

Puer: 178 - HR - - RIGVIGPPLERNTGA - - GRLAGPKKAMEERAMIKVPESSNI^{(β)F}VQGD^{(β)G}PEPESGYRAMQOIL^{(β)I}
RBP: 121 - GE - - GAKVIEIQGIAGTSAARQGE^{(β)F}GF^{(β)F}QOAVAAH - - KFNVLASQPAD^{(β)F}DEIKGLNVKQNL^{(β)I}
OBP-E: 135 LNXDQIQIOPVLLXGEPQHPDAKERTYVYKELNDKGIKTEQLD^{(β)F}TAMWD^{(β)F}IAQAKDKMDAWL^{(β)I}

Puer: 235 SQP - - HRPTAVFCGGDIMAHGALCAADEMGLRVPQDVSLIGYDNVNRNARYFTPALITIH^{(β)H}
RBP: 178 TA - - RPDVQAVFAQND^{(β)H}EMALGAL^{(β)H}ALQTA^{(β)H}GKSDVMVVGFDGTRDGEKAVNDGKLAATIA^{(β)H}
OBP-E: 197 SGPANKIEVVIANN^{(β)H}DAMAMGAVEALKAHNKSSIPVEGVDALPEALALYKSGALAGIYL^{(β)H}

Puer: 282 QPKDSLQETAFPHLLDRIVNKRREP - Q - - - - SEVHP^{(β)H}ELI^{(β)H}ERRSVADGGPFRDYRR^{(β)H}
RBP: 235 QLEDQIGAKGVET - ADK - VLXGQKV - Q - - - - AKYFV^{(β)H}DL^{(β)H}KLVVXQ^{(β)H}
OBP-E: 256 NDA^{(β)H}NQAKATFDL - AKN - LADGKGAADGTINWKINDNKVVRVYVYGVGDKDNLAEP - - - - SKK^{(β)H}

Table 1. Results for DAAC (direct amino acid comparisons) and MBC/C (minimum base change per codon) of a search carried out using the entire sequence of the CBD of PurR with various stretches of RBP (column 1a). The CBD segments which matched best those in RBP column 1a (for DAAC) are indicated in column 1b, next to the corresponding residues in RBP.

TABLE I

Identity scores of contiguous segments of RBP and the CBD of PurR

Results of a search carried out on the entire sequence of the CBD of PurR with various stretches of RBP. Significance scores were calculated by the method of Ohlendorf *et al.* 1983.

| Amino acid segment | | Disagreement | | Disagreement | |
|--------------------|---------|---------------------|--------------|----------------------|--------------|
| ----- | | score | Significance | score | Significance |
| RBP | PurR | (DAAC) ^a | (DAAC) | (MBC/C) ^b | (MBC/C) |
| 2-41 | 60-100 | 0.77 | 4.21 | 0.98 | 4.72 |
| 41-62 | 99-120 | 0.73 | 4.10 | 1.00 | 3.62 |
| 83-126 | 140-183 | 0.77 | 4.56 | 1.11 | 3.78 |
| 126-159 | 183-216 | 0.76 | 3.96 | 1.00 | 4.21 |
| 160-190 | 217-247 | 0.71 | 5.07 | 1.16 | 3.00 |
| 191-230 | 248-287 | 0.73 | 5.23 | 1.20 | 2.66 |
| 231-247 | 288-304 | 0.71 | 3.86 | 0.88 | 3.39 |
| 251-258 | 310-317 | 0.63 | 3.99 | 0.88 | 2.88 |

^a The DAAC disagreement score for two random sequences is 0.94.

^b The MBC/C disagreement score for two random sequences is 1.44.

Figure 2. Alignment of PurR against *E. coli* DNA-binding proteins, LacI and CytR. The region of the helix-turn-helix-loop-helix (H-T-H-L-H) in LacI, as determined by NMR, and the corresponding regions in PurR and CytR are appropriately designated. Also designated is a fourth, predicted helix.

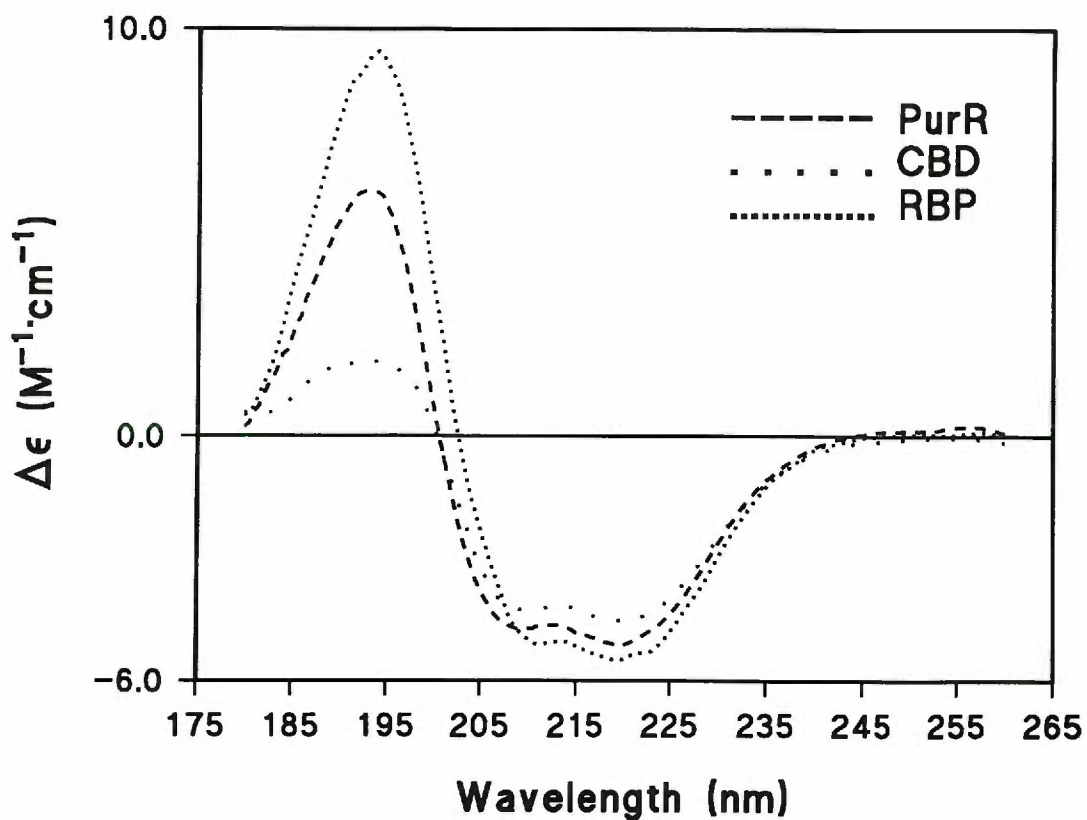
Table II. The observed ligand-binding residues of GBP-E and RBP and the corresponding predicted residues in PurR.

TABLE II

Observed ligand-binding residues of GBP-E and RBP and the corresponding predicted residues in PurR.

| GBP-E observed | Protein-ligand interaction | RBP observed | Protein-ligand interaction | PurR predicted | Proposed interaction |
|--------------------|-------------------------------|--------------------|-------------------------------|--------------------|--------------------------|
| Asp ¹⁴ | H-bonding | Asn ¹³ | H-bonding | Ala ⁷¹ | None or van der Waals |
| Phe ¹⁶ | Stacking | Phe ¹⁵ | Stacking | Tyr ⁷³ | Stacking |
| Met ¹⁷ | None | Phe ¹⁶ | Stacking | Phe ⁷⁴ | Stacking |
| Asn ⁹¹ | H-bonding | Asp ⁸⁹ | H-bonding | Asp ¹⁴⁶ | H-bonding |
| Lys ⁹² | H-bonding | Arg ⁹⁰ | H-bonding | Trp ¹⁴⁷ | Stacking or H-bonding |
| Arg ¹⁵⁸ | H-bonding | Arg ¹⁴¹ | H-bonding | Arg ¹⁹⁶ | H-bonding |
| Trp ¹⁸³ | Stacking | Phe ¹⁶⁴ | Stacking | Phe ²²¹ | Stacking |
| Asn ²¹¹ | H-bonding | Asn ¹⁹⁰ | H-bonding | Gly ²⁴⁷ | None or van der Waals |
| Asp ²³⁶ | H-bonding | Asp ²¹⁵ | H-bonding | Asp ²⁷⁵ | H-bonding |

Figure 3. CD Spectra of the E. coli purine repressor protein, its corepressor binding domain, and the E. coli ribose-binding protein and secondary structure analysis of the CD spectra. H, α -helix; B, β -sheet; T, β -turns; O, other; Total, total secondary structure content.

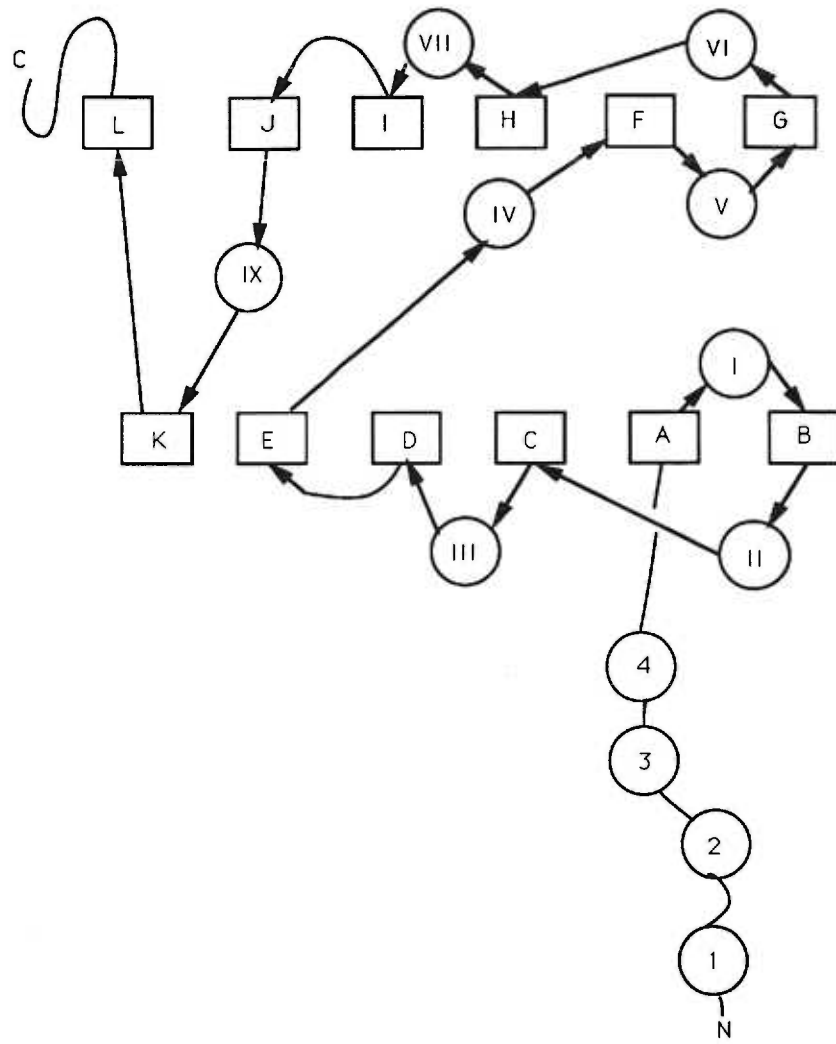


Secondary Structure Analysis of the CD Spectra

| <u>Protein</u> | <u>Method</u> | <u>H</u> | <u>B</u> | <u>T</u> | <u>O</u> | <u>Total</u> |
|----------------|---------------|----------|----------|----------|----------|--------------|
| PurR | CD | 0.40 | 0.27 | 0.16 | 0.20 | 1.03 |
| | Predicted | 0.39 | 0.15 | 0.19 | 0.27 | |
| CBD | CD | 0.33 | 0.16 | 0.21 | 0.32 | 1.02 |
| | Predicted | 0.33 | 0.17 | 0.21 | 0.29 | |
| RBP | CD | 0.46 | 0.19 | 0.11 | 0.23 | 0.99 |
| | X-ray | 0.48 | 0.20 | 0.11 | 0.21 | |

Abbreviations: H, α -helix; B, β -sheet; T, β -turns; O, other;
Total, total secondary structure content

Figure 4. Predicted secondary structure topology of the purine repressor protein. The boxes indicate b-strands and the circles a-helices. Domain crossovers are between $\beta_E \rightarrow \alpha_{IV}$ and $\beta_J \rightarrow \alpha_{IX}$ and $\beta_K \rightarrow \beta_L$.



Manuscript # 2

Crystallization and Preliminary X-ray Analysis of an *Escherichia coli* Purine

Maria A. Schumacher¹, Kang Yell Choi², Howard Zalkin² and Richard G. Brennan^{1*}

¹Department of Biochemistry and Molecular Biology
Oregon Health Sciences University
Portland, OR 97201-3098, U.S.A.

²Department of Biochemistry
Purdue University
West Lafayette, IN 47907, U.S.A.

* Author to whom correspondence should be addressed:

Department of Biochemistry and Molecular Biology
Oregon Health Sciences University
3181 S.W. Sam Jackson Park Road
Portland, OR 97201-3098, U.S.A.
Phone: 503-494-4427; E-Mail: BRENNANR@OHSU.EDU
FAX: 503-494-8393

Running title: Crystallization of a PurR-Hypoxanthine-DNA Complex

Keywords: PurR, protein-DNA crystallization, LacI family, transcription repressor.

ABSTRACT

The Purine Repressor, PurR, is a DNA-binding protein, which together with a purine corepressor, serves to regulate *de novo* purine and pyrimidine biosynthesis in *E. coli*. PurR belongs to the structurally homologous *lac* repressor family of transcription regulators. These proteins are functionally bipartite with an NH₂-terminal Helix-Turn-Helix DNA-binding domain and a larger COOH-terminal effector molecule binding domain which is structurally related to the bacterial periplasmic binding proteins and several mammalian ligand-binding proteins. To fully understand the structural basis of the DNA and corepressor-binding functions of PurR and the mechanism by which the environmental signal of corepressor binding is transduced to the DNA-binding domain, a PurR-hypoxanthine-DNA complex has been crystallized. The DNA encompasses the high-affinity *purF* operator site and is 16-base pairs long with 5'-deoxynucleoside overhangs on each complementary strand. The crystals diffract to better than 2.6Å and take the orthorhombic space group C222₁ with unit cell dimensions **a**= 175.9Å, **b**= 94.8Å and **c** = 81.8Å. The structure determination of this PurR-hypoxanthine-DNA complex will provide the first high resolution view of a LacI member-DNA complex.

The Purine Repressor, PurR, is a 38-kDa DNA-binding protein of 341 amino acid residues that functions in *E. coli* as the master regulatory protein for *de novo* purine biosynthesis and to a lesser extent, *de novo* pyrimidine biosynthesis (Rolfes & Zalkin, 1988; Zalkin & Dixon, 1992). In *E. coli* the genes encoding the enzymes utilized in *de novo* purine biosynthesis are arranged in ten polycistronic and monocistronic operons. PurR also regulates six genes encoding enzymes involved in *de novo* pyrimidine biosynthesis and salvage and the generation of one carbon units needed for *de novo* purine biosynthesis and 5'-phosphoribosyl-1'-pyrophosphate which is utilized for *de novo* and salvage synthesis of nucleotides. In addition, PurR is autoregulated (Rolfes & Zalkin, 1990a; Meng *et al.*, 1990). The DNA operator binding sites recognized by dimeric PurR have been characterized and are pseudo-palindromes, 16-base pairs long and fairly AT-rich (Rolfes & Zalkin, 1990a). One such high affinity site, the *purF* operator, is bound by PurR with a K_D of 3.4 nM (Rolfes & Zalkin, 1990b). To bind its DNA operator sites, PurR must first bind a purine corepressor, either hypoxanthine or guanine (Meng & Nygaard, 1990; Rolfes & Zalkin, 1990b), which bind with equilibrium dissociation constants of 9.3 μ M and 1.5 μ M, respectively (Choi & Zalkin, 1992).

PurR is a member of the LacI family of transcription regulators (Vartak *et al.*, 1991; Weickert & Adhya, 1992). As observed for all LacI members, the biochemical functions of PurR are structurally divided whereby the smaller NH₂-terminal domain, approximately the first 60 amino acid residues, binds DNA via a Helix-Turn-Helix motif and the larger COOH-terminal domain binds corepressor. That the NH₂-terminal domain binds cognate DNA sites via a Helix-Turn-Helix motif has been confirmed most recently by NMR and restrained molecular dynamics studies on residues 1-56 of the lac repressor bound to an 11-base pair *lac* operator half site (Chuprina *et al.*, 1993). However, unlike any other structurally characterized Helix-Turn-Helix DNA-binding protein, PurR is unusual in that the "invariant" glycine of the Turn is an asparagine residue (Brennan & Matthews, 1989). PurR as well as all other members of the lac family, diverges further from LacI as the former repressors do not have the COOH-terminal leucine zipper extension that is responsible for the tetrameric state of LacI (Chakerian *et al.*, 1991; Alberti *et al.*, 1993). Interestingly, the ligand binding domains of all LacI members appear to be structurally homologous to the bacterial periplasmic binding proteins (Müller-Hill *et al.*, 1983; Sams *et al.*, 1984; Bowie *et al.*, 1991; Mauzy & Hermodson, 1992; Schumacher *et al.*, 1993; Nichols *et al.*, 1993) which are also structurally related to lactoferrin (Anderson *et al.*, 1987). Very recently this structural relationship has been extended to the extracellular domain of the mammalian metabotropic glutamate receptors (O'Hara *et al.*, 1993) and parathyroid Ca²⁺-sensing receptor (Brown *et al.*, 1993; Conklin & Bourne, 1994).

However, structural confirmation of this homology awaits the crystal structure determinations of intact LacI or its core (Pace *et al.*, 1990; Steitz *et al.*, 1980) and the Corepressor Binding Domain of PurR (Schumacher *et al.*, 1992).

To gain insight into the mechanism by which the signal of corepressor binding is transduced to the DNA-binding domain of PurR, as well as other LacI family members, thereby inducing high-affinity binding to cognate DNA sites, the high resolution structure of an intact protein-corepressor-DNA complex is necessary. To accomplish this we undertook the crystallization of a PurR-corepressor-DNA operator complex.

All deoxyoligonucleotides used in these crystallization experiments encompassed the high affinity pseudo-palindromic, 16-base pair *purF* operator site, the sequence of which is

5'-ACGCAAACGTTTTCTT-3' and its complement. Initially very large tetragonal protein-corepressor-DNA crystals with dimensions of greater than 1.0 mm on a side were obtained with a 21 or 22-base pair operator site using ammonium sulfate solutions as the crystallization reagents. However, the limited resolution (5.5Å) of these crystals made them unsuitable for further crystallographic studies. A longer operator site, 26-base pairs, also yielded large but poorly diffracting crystals. Crystals of a PurR-hypoxanthine-*purF* operator complex that diffract to high resolution were obtained ultimately using a 16-base pair operator site with 5'-nucleoside overhangs (Figure 1). Interestingly, the identity of either 5'-overhang influenced neither crystallization nor diffraction quality. In a typical "hanging drop" experiment (McPherson, 1990), 1.0 µl hypoxanthine-saturated PurR (0.5 mM protein in 1.0 mM DTT buffered by 160 mM sodium-potassium phosphate, pH 7.4) was added to 1.2 µl *purF* oligonucleotide (0.6 mM DNA in 50 mM sodium cacodylate, pH 6.9), mixed with 2 µl of solution B (25% PEG 4000, 0.4 M ammonium sulfate, 50 mM cobalt hexammine, 0.1 M ammonium phosphate, pH 7.5) and equilibrated over a 1 ml reservoir of solution B at room temperature (20° C). Cobalt hexammine was critical to the crystallization process.

Solution B is unusual in that its components are immiscible and must be vigorously mixed immediately prior to addition to the protein-DNA drop. Phase separation is noticeable within both the drop and the reservoir in less than one hour (Figure 1). Crystallization also proceeds in an unusual manner. Two days after set-up thin two dimensional plates appear in the drops. These plates slowly dissolve and give rise to three dimensional crystals over a period of two weeks to two months. Often the crystals grow at the interface of the two phases. The crystals are rhombohedrally shaped and typically grow to dimensions of 0.4 mm X 0.3 mm X 0.2 mm but occasionally grow to approximately twice these dimensions. Native polyacrylamide gel electrophoresis confirmed the presence

of both PurR and DNA (data not shown). Increasing the drop volume multiplicatively improved the crystal volume.

The crystals diffract isotropically and typically to 2.6 Å but diffraction to 2.4 Å has been observed on X-ray "stills". Precession photography reveals that these crystals take the orthorhombic space group $C222_1$ with unit cell dimensions $a = 175.9$ Å, $b = 94.8$ Å and $c = 81.8$ Å. Assuming one PurR dimer-hypoxanthine-*purF* operator complex per asymmetric unit results in a calculated V_M value of 1.95 Å³/dalton. This value is atypically low for proteins and protein-DNA complexes (Matthews, 1968). Alternatively, a PurR monomer-hypoxanthine-*purF* operator half site per asymmetric unit would give a more reasonable V_M of 3.9 Å³/dalton. However, this would dictate the DNA to be statistically disordered as the *purF* operator site used in these crystallization experiments is only pseudo-palindromic. An initial data set has been collected using a San Diego Multiwire Systems area detector and Rigaku RU200-H X-ray generator set at 40 kV and 150 mA. The data were collected at room temperature (20° C) and were 97% complete to 2.7 Å with an R_{SYM} of 5.6% and an average $I/\sigma I > 9.4$ for all data and > 2.5 for data in the 2.8 Å to 2.7 Å resolution shell.

The search for heavy atom derivatives has commenced. A successful strategy for the generation of isomorphous heavy atom derivatives for protein-DNA complexes has been to substitute thymidine with 5-iodouridine and to a lesser extent cytidine with 5-iodocytidine (reviewed in Agarwaal, 1990). The former approach looks particularly promising for the *purR*-hypoxanthine-*purF* operator crystals as the *purF* operator contains ten possible substitution sites (Figure 1). Indeed, isomorphous crystals of this protein-corepressor-DNA complex substituted with 5-iodouridine, singly or multiply, have been grown recently to data-quality size. Intensity data collections for these derivatized crystals are now underway.

The X-ray structure determination of this PurR-hypoxanthine-*purF* operator complex will provide for the first time a high-resolution view of any intact LacI family member bound to DNA. Furthermore, this structure will provide considerable insight into the biochemical mechanisms by which small effector molecules transduce environmental signals into precise transcription regulation.

ACKNOWLEDGEMENTS

This work is supported by Public Health Service grant GM 24658 (H.Z.), by a National Defense Science and Engineering Graduate Fellowship in Biosciences (M.A.S.) and by Basic Research Grant No. 505 from the March of Dimes Birth Defects Foundation (R.G.B.) and the Oregon Community Foundation (R.G.B.)

REFERENCES

- Aggarwal, A. (1990). Crystallization of DNA binding proteins with oligonucleotides. *Methods: A Companion to Methods in Enzymology* **1**, 83-90.
- Alberti, S., Oehler, S., von Wilcken-Bergmann, B. & Müller-Hill, B. (1993). Genetic analysis of the leucine heptad repeats of Lac repressor: evidence for a 4-helical bundle. *EMBO J.* **12**, 3227-3236.
- Anderson, B. F., Baker, H. M., Norris, G. E., Rumball, S. V. & Baker, E. N. (1990). Apolactoferrin structure demonstrates ligand-induced conformational change in transferrins. *Nature* **344**, 784-787.
- Bowie, J. E., Lüthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164-170.
- Brennan, R. G. & Matthews, B. W. (1989). The helix-turn-helix DNA-binding motif. *J. Biol. Chem.* **264**, 1903-1906.
- Brown, E. M., Gamba, G., Riccardi, D., Lombardi, M., Butters, R., Kifor, O., Sun, A., Hediger, M. A., Lytton, J. & Hebert, S. C. (1993). Cloning and characterization of an extracellular Ca²⁺-sensing receptor from bovine parathyroid. *Nature* **366**, 575-580.
- Chakerian, A. E., Tesmer, V. M., Manly S. P., Lynch, M. J., Hoh, J. T. & Matthews, K. S. (1991). Evidence for leucine zipper motif in lactose repressor protein. *J. Biol. Chem.* **266**, 1371-1374.
- Choi, K. Y. & Zalkin, H. (1992). Structural characterization and corepressor binding of the *Escherichia coli* purine repressor. *J. Bacteriol.* **174**, 6207-6214.
- Chuprina, V. P., Rullmann, J. A. C., Lamerichs, R. M. J. N., van Boom, J. H. & Kaptein, R. (1993). Structure of the Complex of *lac* repressor headpiece and an 11 base-pair half-operator determined by nuclear magnetic resonance spectroscopy and restrained molecular dynamics. *J. Mol. Biol.* **234**, 446-462.

Conklin, B. R. & Bourne, H. R. (1994). Marriage of the flytrap and the serpent. *Nature* **367**, 22.

Matthews, B. W. (1968). Solvent content of protein crystals. *J. Mol. Biol.* **33**, 491-497.

McPherson, A. (1990). Current approaches to macromolecular crystallization. *Eur. J. Biochem.* **189**, 1-23.

Meng, L., M. & Nygaard, P. (1990). Identification of hypoxanthine and guanine as the corepressors for the purine regulon genes of *Escherichia coli*. *Mol. Microbiol.* **4**, 2187-2191.

Meng, L. M., Kilstrup, M. & Nygaard, P. (1990). Autoregulation of PurR repressor synthesis and involvement of *purR* in the regulation of *purB*, *purC*, *purL*, *PurMN* and *guaBA* expression in *Escherichia coli*. *Eur. J. Biochem.* **187**, 374-379.

Müller-Hill, B. (1983). Sequence homology between lac and gal repressors and three sugar-binding periplasmic proteins. *Nature* **302**, 163-164.

Nichols, J. C., Vyas, N. K., Quioco, F. A. & Matthews, K. S. (1993). Model of lactose repressor core based on alignment with sugar-binding proteins is concordant with genetic and chemical data. *J. Biol. Chem.* **268**, 17602-17612.

O'Hara, P. J., Sheppard, P. O., Thogersen, H., Venezia, D., Haldeman, B. A., Mcgrane, V., Houamed, K. M., Thomsen, C., Gilbert, T. L. & Mulvihill, E. R. (1993). The ligand-binding domain in metabotropic glutamate receptors is related to bacterial periplasmic binding proteins. *Neuron* **11**, 41-52.

Pace, H. C., Lu, P. and Lewis, M. (1990). *lac* repressor: crystallization of intact tetramer and its complexes with inducer and operator DNA. *Proc. Nat'l. Acad. Sci., U.S.A.* **87**, 1870-1873.

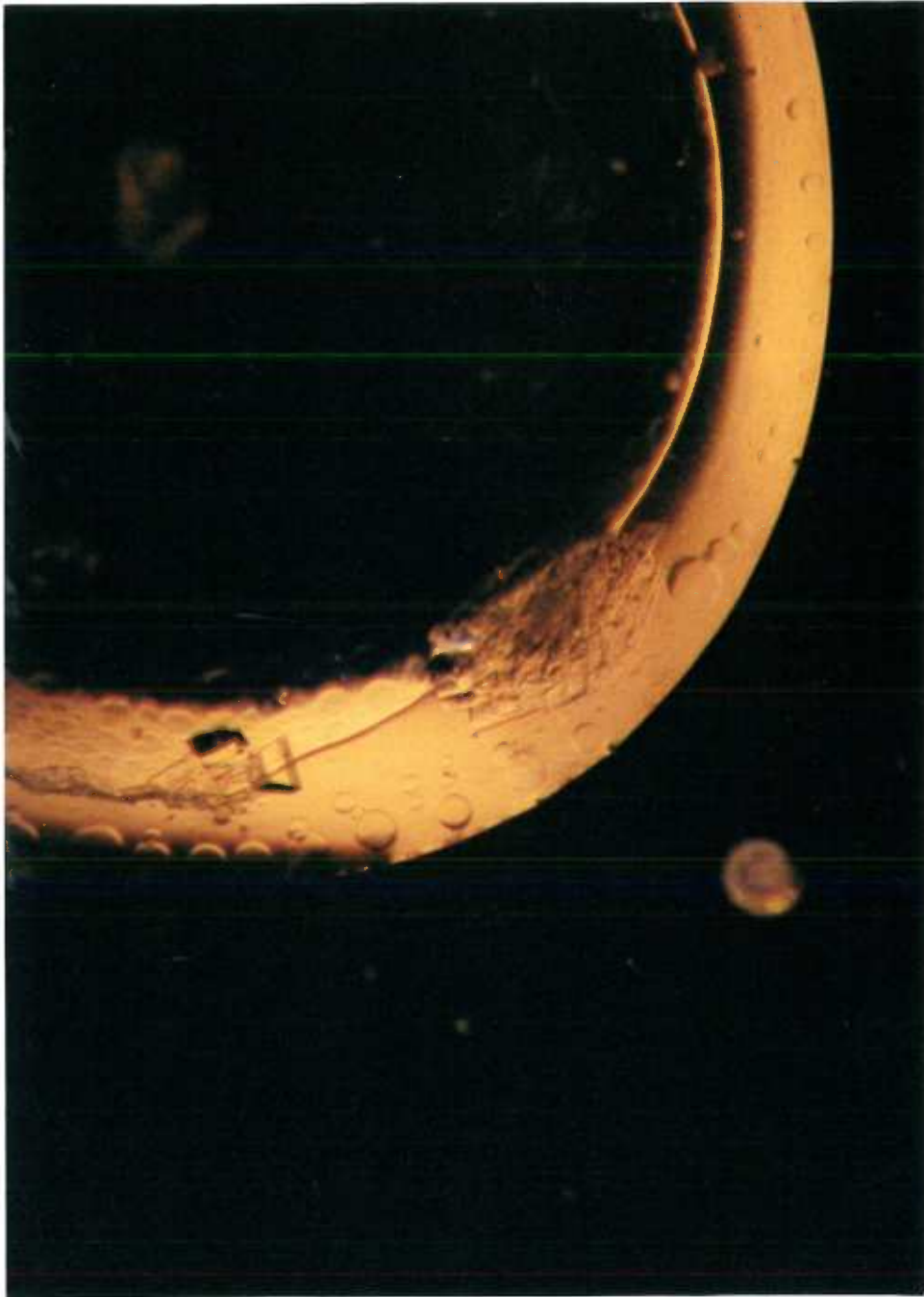
Rolfes, R. J. & Zalkin, H. (1988). *Escherichia coli* gene *purR* encoding a repressor protein for purine nucleotide synthesis. *J. Biol. Chem.* **263**, 19653-19661.

Rolfes, R. J. & Zalkin H. (1990a). Autoregulation of *Escherichia coli* *purR* requires two control sites down stream of the promoter. *J. Bacteriol.* **172**, 5758-5766.

- Rolfes R. J. & Zalkin, H. (1990b). Purification of the *Escherichia coli* purine regulon repressor and identification of corepressors. *J. Bacteriol.* **172**, 5202-5219.
- Sams, C. F., Vyas, N. K., Quioco, F. A. & Matthews, K. S. (1984). Predicted structure of the sugar-binding site of the *lac* repressor. *Nature* **310**, 429-430.
- Schumacher, M.A., Choi, K. Y., Zalkin, H. & Brennan, R. G. (1992). Crystallization and preliminary X-ray studies on the co-repressor binding domain of the *Escherichia coli* purine repressor. *J. Mol. Biol* **225**, 1131-1133.
- Schumacher, M. A., Macdonald, J.R., Bjorkman, J., Mowbray, S. L., & Brennan, R. G. (1993). Structural analysis of the purine repressor, an *Escherichia coli* DNA-binding protein. *J. Biol. Chem.* **268**, 12282-12288.
- Steitz, T. A., Stenkamp, R. E., Geisler, N., Weber, K. & Finch, J. (1980). X-ray and electron microscopic studies of crystals of core *lac* repressor protein. *Biomol. Struct. Conform. Funct. & Evol.* **1**, 425-430.
- Vartak, N. B., Reizer, J., Reizer, A., Gripp, J. T., Groisman, E. A., Wu, L.-F., Tomich, J. M., & Saier, Jr., M. H. (1991). Sequence and evolution of the FruR protein of *Salmonella typhimurium*: a pleiotropic transcriptional regulatory protein possessing both activator and repressor functions which is homologous to the periplasmic ribose binding protein. *Res. Microbiol.* **142**, 951-960.
- Weickert, M. J. & Adhya, S. (1992). A family of bacterial regulators homologous to gal and lac repressor. *J. Biol. Chem.* **267**, 15869-15874.
- Zalkin, H. & Dixon, J. E. (1992). *de novo* purine nucleotide biosynthesis. *Progr. Nucleic Acids Res. Mol. Biol.* **42**, 259-287.

Figure 1: Crystals of the PurR-hypoxanthine-*purF* operator complex. Notice the presence of both "oils" and thin plates, the latter of which are beginning to disintegrate. The rhombohedrally shaped crystal in this photograph has dimensions of approximately 0.22 mm X 0.17 X 0.12 mm. The sequence of the deoxyoligonucleotide cocrystallized in this complex is

5'-TACGCAAACG**TTTTCTT**-3' and its complement,
5'-AAAGAAAACG**TTTGCGT**-3', in which the *purF* operator site is shown in bold.



Manuscript # 3

**Crystal Structure of the LacI Member, PurR, Bound
to DNA: Minor Groove Binding by α -Helices**

Maria A. Schumacher¹, Kang Yell Choi², Howard Zalkin² and Richard G. Brennan^{1*}

¹Department of Biochemistry and Molecular Biology
Oregon Health Sciences University
Portland, Oregon 97201-3098
Phone: (503) 494-4427
FAX: (503) 494-8393

²Department of Biochemistry
Purdue University
West Lafayette, Indiana 47907

*Person to whom correspondence should be addressed.

ABSTRACT

The three-dimensional structure of a ternary complex of the purine repressor, PurR, bound to both its corepressor, hypoxanthine, and the 16-base pair *purF* operator site has been solved at 2.7 Å resolution by X-ray crystallography. The bipartite structure of PurR consists of an NH₂-terminal DNA-binding domain and a larger COOH-terminal corepressor binding and dimerization domain, which is strikingly similar to that of the bacterial periplasmic binding proteins. The DNA-binding domain contains a helix-turn-helix motif that makes base-specific contacts in the major groove of the DNA. Base contacts are also made by residues of symmetry-related α helices, the "hinge" helices, which bind deeply in the minor groove. Critical to hinge helix-minor groove binding is the intercalation of the side chains of Leu54 and its symmetry-related mate, Leu54', into the central CpG base pair step. These residues thereby act as "leucine levers" to pry open the minor groove and kink the *purF* operator by 45°.

The purine repressor, PurR, is a 341 amino acid DNA-binding protein that functions as the master regulator of *de novo* purine biosynthesis and, to a lesser extent, *de novo* pyrimidine biosynthesis in *Escherichia coli* (1). Specifically, PurR represses transcription from ten polycistronic and monocistronic operons, which encode the enzymes of *de novo* purine biosynthesis, and four genes encoding enzymes involved in *de novo* pyrimidine biosynthesis and salvage. In addition, PurR regulates the transcription of *glyA*, *gcv* and *prs* which encode enzymes for the synthesis of glycine, one-carbon units and 5-phosphoribosyl-1-pyrophosphate (PRPP), respectively, all of which are necessary for *de novo* purine biosynthesis. Finally, PurR is autoregulated (2).

PurR is a member of the LacI (lactose repressor) family of transcription regulators of which there are more than 21 members (3). These proteins show strong sequence similarity indicative of a structural relationship. Sequence identity is greatest in the amino-terminus and often exceeds 60 percent. Genetic and biochemical studies have shown that these proteins can be divided into two functional domains, an NH₂-terminal DNA-binding domain, approximately the first 60 residues, and a larger COOH-terminal domain, approximately 280 residues, which imparts the functions of effector binding and oligomerization (3). Whereas the lactose, fructose and raffinose repressors exist as tetramers (4) all other LacI members appear to be dimeric (3). These proteins function as repressors by binding to operator sites, typically 16 to 18 base pairs (bp) long, that also display significant sequence identity (3). For most LacI family members, operator affinity is highest for the unliganded state of the protein. However, for PurR, binding to operator DNA is dependent upon a corepressor. The corepressors for PurR are hypoxanthine and guanine (2), which bind cooperatively with equilibrium dissociation constants (K_d s) of 9.3 and 1.5 μ M, respectively, (5).

Several studies have suggested that the structures of the effector binding domains of LacI family members are similar to the bacterial periplasmic binding proteins (PBPs) (6), and three-dimensional models of the effector binding domains of LacI (7) and the galactose repressor, GalR, (8) have been constructed based on these studies. However, the only three-dimensional structural data available for the LacI proteins have been obtained from nuclear magnetic resonance studies of the LacI headpiece, both free (9) and bound to DNA (10). From these studies, the presence of a helix-turn-helix (HTH) motif (11) was confirmed and contacts between several side chains and bases, some of which had been implicated genetically (12), were determined. However, these studies left in question the structure of the effector binding and dimerization domain and the complete source of DNA sequence discrimination. Additionally, these studies could not address the question of how a signal in the form of a small effector molecule is transduced to effect gene regulation.

We now describe the crystal structure of PurR bound to both its corepressor, hypoxanthine, and a cognate DNA site containing the high affinity *purF* operator ($K_d=3.4 \times 10^{-9}$ M) (2). The structure of this LacI member confirms that the corepressor binding domain has a PBP-like fold and an NH₂-terminal HTH DNA-binding motif. However, the key to DNA binding specificity resides not only in base specific contacts made by the HTH in the major groove but also in DNA deformability and contacts made to the minor groove by the "hinge" helix. Symmetry-related residues Leu54 and Leu54' from each hinge helix act as levers to pry open the minor groove thereby unwinding and kinking the DNA toward the major groove. The strong sequence similarity between the LacI members and their cognate DNA sites suggests that this is also the means by which other LacI members interact with their operators.

PurR structure and corepressor binding

The structure of the PurR-hypoxanthine-*purF* operator complex was solved by multiple isomorphous replacement (Table 1). The asymmetric unit contains one PurR monomer-hypoxanthine-*purF* operator half site, which requires the statistical disorder of two base pairs and the 5'-nucleoside overhangs because each *purF* operator half-site is not identical (Fig 1A). However, difference Fourier maps of an isomorphous complex, in which the *purF* operator was replaced by a perfect 16 bp palindrome (Fig 1B), revealed no significant differences between these PurR-hypoxanthine-*purF* operator complexes (13). Our model, which includes residues 3 through 340, and 20 water molecules has a current R-factor of 19.5 percent based on all data from 10.0 Å to 2.7 Å (Table 1). The stereochemistry of the model is very good and only two violations of ϕ , ψ space are found (14). A typical section of the current $2(F_{\text{obs}} - F_{\text{calc}})$ electron density map and an "omit" map are shown in Fig 2. A topology diagram of the PurR monomer and a stereo view of the biologically relevant PurR-hypoxanthine-*purF* operator are shown in Fig 3.

The NH₂-terminal DNA-binding domain of PurR can be divided into two functionally important regions. The first contains the HTH motif that spans residues 4-23 (Fig 3). Following this structural element is a short loop, residues 24-29, helix 3, residues 30-43, and another short loop, residues 44-47. Helices 1 through 3 form a globular subdomain that is connected to the corepressor binding domain by the "hinge" or more properly, the "hinge" helix, residues 48-56, which is followed by four extended residues. The hinge helix, helix 4, constitutes the second DNA-binding element of PurR.

The corepressor binding domain (CBD), residues 61-340, has the shape of an oblate ellipsoid (axial ratio 2:1) and consists of two topologically similar subdomains, namely, the CBD NH₂-subdomain, which is directly attached to the DNA-binding domain, and the CBD COOH-subdomain, which contains the COOH-terminus of PurR. The CBD NH₂-subdomain is composed of a core of six parallel β strands, A to E and J, that are flanked on both sides by α helices, I, II, III, and IX, and follows the topology, $\beta_B\beta_A\beta_C\beta_D\beta_E\beta_J$ (Fig 3). The CBD COOH-subdomain is composed of a core of five parallel β strands, F to I and K, flanked by α helices IV, V, VI, VII, and VIII. The β sheet topology of this subdomain is $\beta_F\beta_G\beta_H\beta_I\beta_K$. Three crossover regions connect the two subdomains, the first two are from β strands to α helices (β_E to α_{IV} and β_I to α_{IX}) and the last from β strand to β strand (β_J to β_K). These crossover regions, like those of the PBPs, presumably act as a hinge to allow relative movements of the two subdomains upon ligand association and dissociation (15).

Several studies have predicted that the CBD is structurally similar to the PBPs (6), which are (i) monomeric, and (ii) constitute a large, structurally conserved family that participates in transport of metabolites across the bacterial membrane and, in some cases, chemotaxis (16). Comparisons of the CBD structure with several PBPs reveal that its highest similarity is to the ribose binding protein (RBP) (17). An overlay of the corresponding 144 alpha carbons (C α s) of the CBD and RBP results in a root mean squared deviation (rmsd) of 2.29 Å. A similar overlay with the glucose-galactose binding protein yields an rmsd of 2.48 Å. The corresponding CBD and RBP NH₂-subdomains are more similar than the CBD and RBP COOH-subdomains as their C α overlays reveal respective rmsds of 1.81 Å and 2.34 Å.

The CBD is solely responsible for binding corepressor (5) and the residues lining the interface of the CBD NH₂- and COOH-subdomains contribute to high affinity ligand binding through a combination of polar, nonpolar and aromatic interactions (Fig 3 and 4). Six direct and water-mediated protein-hypoxanthine hydrogen bonds are provided by residues Asp275, Thr192 and Arg190. Asp275 hydrogen bonds to N9 of hypoxanthine via its carboxylate O δ 1 atom (2.73 Å). This contact is anchored by Arg196, which donates hydrogen bonds from its NH2 and NH1 to the O δ 2 of Asp 275 (2.76 Å and 2.98 Å, respectively) (Fig 4). This interaction also neutralizes the charges of these side chains which are buried in the ligand binding pocket. The importance of both residues in corepressor binding is underscored by the greatly diminished corepressor binding observed upon substitution of either residue with alanine (18). Thr192, which forms a side chain hydrogen bond with the N7 of hypoxanthine (O γ -N7, 2.60 Å), also participates in van der Waals contacts to the hypoxanthine by way of its C γ methyl group (Fig 4). As observed in the structures of liganded-PBPs (16, 17), side chain-ligand stacking interactions are also prevalent in PurR-hypoxanthine binding (Fig 4). Aromatic residues Tyr73 and Phe221 form a "sandwich" interaction with the corepressor in which Phe221 is located centrally "above" the hypoxanthine base and Tyr73 interacts on the opposite face more with the purine imidazole ring. Phe74 completes the stack by contacting N1 and C6 (Fig 4).

Whereas the above described interactions are important for purine binding affinity, Arg190 is the key to corepressor specificity in that its side chain ϵ -NH and NH2 hydrogen bond to the O6 acceptor of hypoxanthine (ϵ -NH-O6, 3.10 Å and NH2-O6, 2.83 Å) (Fig 4). These interactions allow PurR to read the exocyclic atom at position 6 of a purine ring and consequently, discriminate against the NH2 donor group of adenine. Arg190 further contributes to hypoxanthine binding by making a bridging hydrogen bond

from its NH₂ nitrogen to a water molecule (2.95 Å), which is in turn hydrogen bonded to the N1 of the purine ring (3.03 Å).

Unlike the PBPs, PurR is a dimer (5), in which the dimerization interface of PurR excludes 2242 Å² of protein surface area from the solvent and is formed, for the most part, equally by both CBD subdomains (Fig 3). Subunit contacts between the two CBD NH₂-subdomains are provided by residues 68 through 115 and include part of the loop before α_I, α_I, β_B, α_{II} and the turns in between (Fig 3). The dimerization region between the two CBD COOH-subdomains is noncontiguous and spans residues 223-229, 249-267, 278-285 and 328-329 (Fig 3). This region includes residues from α_{VI}, the NH₂-terminus of α_{VII}, α_{VIII} and β_K. Two parallel three-helix bundles formed between α_{VI} and α_{VII} of one monomer and α_{VIII} of the other monomer comprise most of the CBD COOH-subdomain interface. Contacts made between residues 328-329 and residues of α_{VII} of the other subunit complete the dimerization interface.

Three cross-subunit contacts between the DNA-binding domain of one subunit and the CBD of the other are found between the main chain carbonyl oxygen of Gln113 and the backbone NH of Ala49', and the NH1 and NH2 of Arg115 and side chain O_γ and main chain CO of Ser46'. Linkage between the CBD NH₂-subdomain of one subunit to the CBD COOH-subdomain of the other is provided by two salt bridges between Arg278 and Glu70' and Glu70 and Arg278'.

DNA binding

PurR makes extensive contacts to the bases and phosphate backbone in the major groove of the *purF* operator through its HTH, the loop that follows and helix 3. The HTH of PurR is unusual in that unlike other structurally characterized HTH motifs, the invariant glycine of the turn is an asparagine. However, the side chain of this branched amino acid does not alter the HTH structure and likely stabilizes the turn by making a hydrogen bond to the peptide backbone CO of residue 10 thereby capping helix 1. Superposition of the 20 C_αs of the HTH of PurR, residues 4-23, on the corresponding C_αs of the λ repressor results in an rmsd of 0.55 Å. A similar superposition of the PurR and LacI HTH (10) yields an rmsd of 0.59 Å (Fig 5) that when extended to include residues 24 through 43, that is, the loop and helix 3, the rmsd is 1.24 Å. Exclusion of the loop residues results in an rmsd of 0.88 Å.

As observed in other prokaryotic HTH protein-DNA complexes, the recognition helix of PurR (helix 2) is positioned on the DNA such that the NH₂-terminus points into the major groove (Fig 6A). In this binding mode, helix 1 contributes one phosphate

contact from the amide NH of Ile4, the first residue in helix 1, to phosphate 8'. The Ile4 side chain also makes van der Waals contacts to the deoxyribose ring of Thy8'. Residues from the turn and helix 2 make six phosphate contacts. Ser14 makes side chain and main chain NH hydrogen bonds to phosphate 3. The side chain of Thr17 also hydrogen bonds to phosphate 3, whereas the side chain NH₂ of Asn23, the last residue of the recognition helix, and the O γ of Thr19 form hydrogen bonds to phosphate 7' (Fig 6A and B). The Ile4 and Asn23-phosphate contacts are analogous to the positioning contacts observed in other HTH protein-DNA complexes (20). Asn23 also makes a key hydrogen bond via its side chain CO to the side chain NH₂ of hinge helix residue Arg52 thus providing a critical link between the globular subdomain, helices 1 through 3, and the hinge region. The loop that follows helix 2 provides two phosphate contacts, one between the amide NH of Ala29 and phosphate 2, and a second, water-mediated contact, from the main chain CO of Phe27 to phosphate 2. Phe27 also stacks against the deoxyribose ring of Ade2. From helix 3 and the short loop which follows, two phosphate contacts are made that involve the side chain of Thr32 and phosphate 2, and the side chain of Tyr45, which hydrogen bonds to three oxygens of phosphate 8. Of the 13 protein-DNA backbone interactions only four phosphate groups are contacted via the major groove (Fig 6A and B).

Five residues from the three-helix globular subdomain participate in direct and water-mediated major groove base contacts (Fig 6A and 6B). Helix 1 contributes van der Waals contacts from the C γ methyl group of Ile4 to the methyl groups of Thy7' and Thy8'. The first residue of the recognition helix, Thr15, participates in the only water-mediated base contact, in which its O γ is hydrogen bonded to a water (2.54 Å) that is hydrogen bonded to the O4 atom of Thy7' (2.56 Å). The O γ of Thr16 bonds simultaneously to the N6 of Ade6 and the O4 of Thy6' (Fig 6A and B). Our structure analysis of the complex of PurR-hypoxanthine and the perfect palindrome operator (Fig 1B) reveals that when base pair 5 is C-G, the O γ of Thr16 makes an additional hydrogen bond to the N4 of Cyt5. The final major groove contact is directed by a van der Waals interaction between the imidazole ring of His20 and the side chain of Arg26, which is located in the loop that follows helix 2. This protein-protein contact positions the guanidino nitrogens to donate hydrogen bonds to the N7 (2.96 Å) and O6 (2.52 Å) acceptors of Gua4 (Fig 6A and 6B). Accordingly, Gua4 is the only guanine of the *purF* operator to display methylation interference sensitivity (20). Furthermore, this contact provides an explanation for the absolute conservation of guanine at position 4 in all known *pur* operators (1,21).

The hinge helix and minor groove binding

The most striking and unanticipated feature of the PurR-hypoxanthine-*purF* operator complex is the binding of the DNA minor groove by helix 4, the hinge helix, and its dyadic partner (Fig 6C and 6D). The hinge helices interact with each other through a series of van der Waals contacts between the side chains of residues Val50 and Val50' and Leu54 and Leu54' (Fig 6D). The accommodation of the hinge helices in the minor groove requires its expansion. This is accomplished by the side chains of residues Leu54 and Leu54', which intercalate into the central Cyt9pGua9' base pair step and thereby act as "leucine levers" to pry open the minor groove (Fig 2, 6C and 6D). As a result, two sets of van der Waals contacts are formed, one, between the C γ methyl group of Leu54 and the O2 oxygen of Cyt9 and the second between the Leu54 C δ 1 and the deoxyribose O4' atom of Gua9' on the same strand (Fig 6D). Operator site specificity is contributed to by hinge helix residue Lys55 which makes a hydrogen bond from its ϵ -NH₃ to the N3 of Ade8 (3.11 Å) and van der Waals contacts to C2. Additional DNA backbone contacts are made by the amide group of Ser48, which contacts phosphate 8', the side chain of Arg52, which hydrogen bonds to phosphate 7', and the C β methyl group of Ala51, which makes van der Waals contacts to the deoxyribose ring of Ade7.

That hinge helix residues are critical for *purF* operator binding by PurR is corroborated by mutagenesis data showing that substitution of Leu54 by lysine, serine, tryptophan, threonine or arginine, produces a PurR that cannot regulate *in vivo* transcription from a *purF-lacZ* fusion gene (22). However, replacement by methionine results in a near wild type phenotype, an indication that the size and hydrophobic nature of this position are crucial to proper function. Substitution of hinge helix residues Arg52 and Lys55 with either alanine or glutamic acid results in similar deregulation of the *purF-lacZ* fusion gene with the glutamic acid substitutions being particularly severe (22).

DNA structure and kinking

The most evident consequence of the hinge helices binding to the minor groove is a single 45° kink at the central Cyt9pGua9' base pair step, which bends the DNA toward the major groove and away from the protein (Figs 2, 6C and 6D). The Cyt9pGua9' step displays an unusually large roll angle of 45° and helical rise of 6.4 Å and is clearly unstacked. This CpG step is also unwound with a twist angle of 27°. Helix insertion splay open the minor groove which can be quantitated by the Cyt9-Gua9' cross-groove O4' to O4' distance of 9.2 Å. The A-T base pair directly 5' of the kink, which is contacted

by a hinge helix residue Lys55, displays severe unwinding, a slide toward the minor groove and an A-DNA like helical rise (Table 2). Analysis of the helical twist angles of the three central base-pair steps reveals that PurR locally unwinds the *purF* operator by 42°, a value similar to that reported for the unwinding of the *lac* operator by LacI (23).

The remainder of the *purF* operator half-site is also affected by PurR binding (Table 2). Measurement of the minor groove width reveals values that range from 7.8 Å to 10.3 Å (Table 2). Such expansion of the entire minor groove results directly from the insertion of the hinge helix and indirectly from the recognition helix compression of the major groove. Despite the unusually wide minor groove, the helical rise, twist and slide of base pair steps 3 through 7 more closely resemble B-DNA (Table 2). The aberrant slide and twist angle values seen for Ade2-Thy2' could result from crystal packing effects as the PurR COOH-terminus, which is rich in aromatic and basic residues, contacts the 5'-nucleoside overhang. As a result, the DNA is not pseudo-continuous in the crystal. No bifurcated hydrogen bonds are observed between any base pair steps.

Many DNA-binding proteins bend their DNA recognition sites (11, 24). However, DNA kinking has been observed directly in only three transcription factor-DNA complexes, the catabolite gene activator protein, (CAP) (25), the TATA binding protein (TBP) (26, 27) and now PurR. CAP kinks its DNA binding site by ~40° at two dyad related TpG steps (25). Stabilization of the kink is brought about by several protein-major groove and protein-phosphate backbone interactions emanating, in large part, from residues of the recognition helix. *A. thaliana* TBP kinks the TATA element at the 5'-TpA and ApG steps of a TATAAAAG binding site (26) whereas yeast TBP kinks the 5'-TpA and ApA steps of a TATATAAA binding site (27). Unlike CAP, TBP stabilizes its ~45° kinks through minor groove base-phenylalanine stacking and phenylalanine-deoxyribose packing interactions, the locations of which are in or near β strands. Somewhat of a hybrid, PurR like CAP, uses an α helix to induce and stabilize its kink site, but like TBP, interacts in the minor groove by analogous van der Waals contacts to the base and deoxyribose ring. NMR studies on the sex determining factor, SRY, bound to a high-affinity DNA site, suggest that this HMG protein interacts with the minor groove in a manner similar to PurR whereby an isoleucine side chain appears to intercalate into a TpT base pair step (28).

The energetic compensation for kinking the DNA and unstacking the most favorable base pair step (29) comes in large part from the extensive hinge helix-minor groove and recognition helix-major groove interactions and the entropy gain related to freeing waters of hydration from the DNA. Another contribution likely arises from the local folding of the hinge helix upon *purF* operator binding. The thermodynamics of DNA binding-induced

protein folding have been described for several protein-DNA complexes and such folding contributes substantially to DNA binding affinity (30). Evidence that the hinge helix of PurR is unfolded in the absence of operator DNA is given by its susceptibility to rapid cleavage by a variety of proteases (5,22).

Operator specificity of LacI members

Inspection of a sequence alignment of twenty one LacI members reveals leucine is found at the position that corresponds to 54 in PurR in all but the Cytidine repressor, CytR, which has a valine (3). As conserved, is the alanine corresponding to Ala51 in PurR, again with the lone exception of CytR, which substitutes a glycine. From inspection of the PurR-*purF* operator complex it is evident that this position can be only an alanine or glycine as residue 51 directly abuts the DNA phosphate backbone (Fig 6D). The nearly complete conservation of these key residues of the hinge helix, combined with the conservation of a central CpG base-pair step in all LacI family operators, and biochemical studies demonstrating that several LacI family members bend their operators (31), strongly suggests similar modes of minor groove binding by all LacI family members. However, modification of the operator site can alter binding affinity. A case in point is the wild type *lac* operator, which contains an additional central C-G base pair that when deleted, increases LacI-*lac* operator affinity by 8 to 10 fold (32).

Operator discrimination by LacI family members relies on DNA deformability and base specific contacts. Sequence alignment of the recognition helices and position 55 of ten LacI proteins and their consensus operators (3) provides insight into the basis of DNA recognition by the LacI family (Fig 7). In PurR, the second residue of the recognition helix, Thr16, hydrogen bonds to an A-T base pair at position 6, whereas in LacI the corresponding residue, Gln18, contacts a G-C base pair also at position 6 (10, 12) (Fig 7). Of the eight remaining LacI members, only threonine, serine, and alanine are found and base pair 6 of their respective consensus operators is always an A-T, with the exception of the ribitol repressor, RbtR, which binds operators containing either C-G or A-T (Fig 7). This correlation suggests that alanine, serine and threonine strongly prefer an A-T pair at position 6 but that glutamine can bind only a G-C base pair. The identity of base pair 4 is determined directly or indirectly by the sixth residue of the recognition helix. In LacI, this residue is an arginine that interacts with Gua4 (10, 12). Other LacI family members that have an arginine at position six, have either a G-C or C-G base pair at position 4 (Fig 7). PurR recruits Arg26 from the loop following the recognition helix, to interact with Gua4, an interaction that is directed by His20 located at position six of the

recognition helix (Fig 6A and 7). Hydrophobic residues at position six appear to select for a T-A base pair at operator position 4 (Fig 7).

A third DNA discriminating contact is made by the hinge helix residue that corresponds to Lys55 in PurR. Lys55 contacts C2 and N3 of Ade8 (Fig 6A and 6D). Other LacI members that have a basic residue at this position also discriminate against a G-C base pair at position 8 most likely the result of steric clash between the side chain and guanine N2 exocyclic group. However, when this residue is small, as in GalR and LacI, there is no discrimination against G-C at position 8.

Mapping LacI mutants

Biochemical and genetic studies, which have resulted in the generation of over 4000 LacI mutants (33), have provided insight into the structure and function correlates of this repressor in the absence of a complete three-dimensional structure. The phenotypic characterization of LacI mutants identifies four categories: repressors defective in operator binding (I^-), repressors defective in effector binding (I^s), repressors that are tight operator binders (I^{tb}), and repressors defective in operator and effector binding.

Now that the structure of an intact LacI member is available we can begin to ascribe the structural basis for these mutant classes. When the LacI sequence is appropriately imposed on the PurR structure it is evident that the many of the mutations are clustered (Fig 8). Most I^- mutations are located in the DNA-binding domain with residues of the HTH and hinge helix being particularly sensitive to substitution (Fig 8, blue). However, mutations leading to incorrect secondary, tertiary or quaternary structures also lead to the I^- phenotype. Many such substitutions are located in the effector binding domain and frequently map to β sheets that constitute the structural core of the protein or to the interior facing hydrophobic regions of helices.

The I^s mutations are clustered mainly in or around the ligand binding cleft (Fig 8, magenta), for example Ser193, which corresponds to a PurR residue Thr192 (Fig 4). Other mutants of this type are found in the dimerization interface in positions where substitutions likely perturb the opening and closing of the ligand binding pocket. The I^{tb} mutants are confined to LacI residues Val24, Ser28, Val52 and Ser61 and correspond to PurR residues Ile22, Arg26, Val50 and Thr59 (Fig 8, green). Of these residues all but Thr59, which is located in the short loop connecting the hinge helix to the CBD, are found in the DNA-binding domain of PurR. The LacI mutants, which display the I^- and I^s phenotypes, are very tightly clustered and are located mostly in the corepressor binding pocket, for example, those corresponding to PurR residues Ser191, Arg196, and Asp275,

or in the dimerization interface, for example, those corresponding to PurR residues Cys281 and Tyr282, (Fig 8, yellow).

In conclusion, the three-dimensional structure of the PurR-hypoxanthine-*purF* operator complex has revealed the atomic details of minor groove binding by α helices and the mechanism by which the side chains of aliphatic amino acids can intercalate into a base pair step and kink its DNA binding site. Furthermore, the structure has broadened our understanding of the DNA binding specificities of the entire LacI family.

References and Notes

1. H. Zalkin and J.E. Dixon, in *Progress in Nucleic Acid Research and Molecular Biology*, (Academic Press, Inc., 1992) vol. 42, pp. 259-287.
2. R.J. Rolfes and H. Zalkin, *J. Bacteriology*. **172**, 5758 (1990);
L. Meng, M. Kilstrup, P. Nygaard, *Eur. J. Biochem.* **187**, 373 (1990).
3. J.D. Gralla, in *Transcriptional Regulation 2*, S.L. McKnight and K.R. Yamamoto, Eds. (Cold Spring Harbor Laboratory Press, 1992), 629;
R.J. Rolfes and H. Zalkin, *J. Biol. Chem.* **263**, 19653 (1988);
M.J. Weichert and S. Adhya, *J. Biol. Chem.* **267**, 15869 (1992).
4. R. Jaenicke, I. Muiznieks, C. Aslanidis, R. Schmitt, *FEBS Lett.* **260**, 233 (1990);
S. Alberti, S. Oehler, B. von Wilcken-Bergmann H. Kramer, B. Müller-Hill, *New Biol.* **3**, 57 (1991); A. Chakerian *et al.*, *J. Biol. Chem.* **266**, 1371 (1991);
S. Alberti, S. Oehler, B. von Wilcken-Bergmann, B. Müller-Hill, *EMBO J.* **12**, 3227 (1993); W-I. Chang, J.S. Olson, K.S. Matthews, *J. Biol. Chem.* **268**, 17613 (1993); J.-C. Cortay, D. Negre, M. Scarabel, T. M. Ramseier, N.B. Vartak, J. Reizer, M.H. Saier, Jr., A.J. Cozzone, *J. Biol. Chem.* **269**, 14885 (1994).
5. K.Y. Choi and H. Zalkin, *J. Bacteriol.* **174**, 6207 (1992).
6. C.A. Mauzy and M.A. Hermodson, *Protein Sci.* **1**, 843 (1992);
J.U. Bowie, R. Lüthy, D. Eisenberg, *Science* **253**, 164 (1991);
N.B. Vartak *et al.*, *Res. Microbiol.* **142**, 951 (1991);
M.A. Schumacher, J.R. Macdonald, J. Björkman, S.L. Mowbray, R.G. Brennan, *J. Biol. Chem.* **268**, 12282 (1993);
C.S. Barbier and S.A. Short, *J. Bacteriol.* **174**, 2881 (1992);
B. Müller-Hill, *Nature* **302**, 163 (1983).
7. J.C. Nichols, N.K. Vyas, F.A. Quicho, K.S. Matthews, *J. Biol. Chem.* **268**, 17602 (1993).
8. M. Hsieh, P. Hensley, M. Brenowitz, J.S. Fetrow, *J. Biol. Chem.* **269**, 13825 (1994).
9. E.R.P. Zuiderweg, R. Kaptein, K. Wüthrich, *Proc. Natl. Acad. Sci. U.S.A.* **80**, 5837 (1983);
R. Kaptein, E.R.P. Zuiderweg, R.M. Scheek, R. Boelens, W.F. van Gunsteren, *J. Mol. Biol.* **182**, 179 (1985).
10. V.P. Chuprina *et al.*, *J. Mol. Biol.* **234**, 446 (1993);
R.M.J.N. Lamerichs *et al.*, *Biochem.* **28**, 2985 (1989).

11. R.G. Brennan, *Curr. Opin. Struct. Biol.* **1**, 80 (1991);
S.C. Harrison and A.K. Aggarwal, *Ann. Rev. Biochem.* **59**, 933 (1990);
R.G. Brennan and B.W. Matthews, *J. Biol. Chem.* **264**, 1903 (1989);
T.A. Steitz, *Q. Rev. Biophys.* **23**, 205 (1990).
C.O. Pabo and R.T. Sauer, *Ann. Rev. Biochem.* **61**, 1053 (1992).
12. N. Lehming, J. Sartorius, S. Oehler, B. von Wilcken-Bergmann, B. Müller-Hill,
Proc. Natl. Acad. Sci. U.S.A. **85**, 7947 (1988).
13. M.A. Schumacher and R.G. Brennan, unpublished results.
14. R.A. Laskowski, M.W. MacArthur, D.S. Moss, J.M. Thornton, *J. Appl. Crystallogr.* **26**, 283 (1993).
15. A.J. Sharff, L.E. Rodseth, J.C. Spurlino, F.A. Quioco, *Biochem.* **31**, 10657 (1992);
G.A. Olah, S. Trakhanov, J. Trehella, F.A. Quioco, *J. Biol. Chem.* **268**, 16241 (1993);
B-H. Oh *et al.*, *J. Biol. Chem.* **268**, 11348 (1993).
16. F.A. Quioco, *Curr. Opin. Struct. Biol.* **1**, 922 (1991);
N.K. Vyas, *Curr. Opin. Struct. Biol.* **1**, 732 (1991).
17. S.L. Mowbray and L.B. Cole, *J. Mol. Biol.* **225**, 155 (1992).
18. K.Y. Choi, F. Lu and H. Zalkin, *J. Biol. Chem.*, accepted (1994).
19. C.O. Pabo *et al.*, *Science* **247**, 1210 (1990).
20. F. Lu and H. Zalkin, unpublished results.
21. B. He, A. Shiau, K.Y. Choi, H. Zalkin, J.M. Smith, *J. Bacteriol.* **172**, 4555 (1990);
B. He, J.M. Smith, H. Zalkin, *J. Bacteriol.* **174**, 130 (1992);
B. He and H. Zalkin, *J. Bacteriol.* **174**, 7121 (1992);
B. He, K.Y. Choi, H. Zalkin, *J. Bacteriol.* **175**, 3598 (1993);
B. He and H. Zalkin, *J. Bacteriol.* **176**, 1009 (1994).
22. K.Y. Choi and H. Zalkin, *J. Bacteriol.* **176**, 1767 (1994).
23. J.C. Wang, M.D. Barkley, S. Bourgeois, *Nature* **251**, 247 (1974).
24. A.A. Travers, *Curr. Opin. Struct. Biol.* **1**, 114 (1991);
A.A. Travers, *Curr. Opin. Struct. Biol.* **2**, 71 (1992);
P.J. Hagerman, *Biochim. Biophys. Acta* **1131**, 125 (1992).
25. S.C. Schultz, G.C. Shields, T.A. Steitz, *Science* **253**, 1001 (1991).
26. J.L. Kim, D.B. Nikolov, S.K. Burley, *Nature* **365**, 520 (1993).
J.L. Kim and S.K. Burley, *Nature struct. Biol.* **1**, 638 (1994).
27. Y. Kim, J.H. Geiger, S. Hahn, P.B. Sigler, *Nature* **365**, 512 (1993).
28. C.-Y. King and M.A. Weiss, *Proc. Natl. Acad. Sci., USA*, **90**, 11990 (1993).
J., **12**, 1311 (1993).

29. R.L. Ornstein, R. Rein, D.L. Breen, R.D. MacElroy, *Biopolymers*. **17**, 2341 (1978).
30. R.S. Spolar and M.T. Record Jr., *Science* **263**, 777 (1994).
31. C. Zwieb, J. Kim and S. Adhya, *Genes & Development*, **3**, 606, (1989);
J. Kim, C. Zwieb, C. Wu and S. Adhya, *Gene*, **85**, 15 (1989);
F. Lu and H. Zalkin, unpublished results.
32. J.R. Sadler, H. Sasmor, J.L. Betz, *Proc. Natl. Acad. Sci. U.S.A.* **80**, 6785 (1983).
A. Simons, D. Tils, B. von Wilcken-Bergmann, B. Müller-Hill, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 1624 (1984).
33. R.O. Spotts, A.E. Chakerian, K.S. Matthews, *J. Biol. Chem.* **266**, 22998 (1991);
W-I. Chang, P. Barrera, K.S. Matthews, *Biochemistry* **33**, 3607 (1994);
L.G. Kleina and J.H. Miller, *J. Mol. Biol.* **212**, 295 (1990);
A.J.E. Gordon *et al.*, *J. Mol. Biol.* **200**, 239 (1988);
A.E. Chakerian and K.S. Matthews, *J. Biol. Chem.* **266**, 22206 (1991);
P. Markiewicz, L.G. Kleina, C. Cruz, S. Ehret, J.H. Miller, *J. Mol. Biol.*, **240**,
421 (1994).
34. M.A. Schumacher, K.Y. Choi, H. Zalkin, R.G. Brennan, *J. Mol. Biol.*, **242**, 302
(1994).
35. N.h. Xuong, C. Nielsen, R. Hamlin, D. Anderson, *J. Appl. Crystallogr.* **18**, 342
(1985).
36. T.C. Terwilliger and D. Eisenberg, *Acta Crystallogr.* **A39**, 813 (1983).
37. B-C. Wang, in *Methods in Enzymology*, H.W. Wykoff, C.H.W. Hirs, S.N.
Timasheff, Eds. (Academic Press, New York, 1985), vol. 115, part B, pp. 90.
38. T.A. Jones, in *Methods in Enzymology*, H.W. Wykoff, C.H.W. Hirs, S.N.
Timasheff, Eds. (Academic Press, New York, 1985) Vol. 115, Part B, pp. 157.
39. D.E. Tronrud, L.J. Ten Eyck, B.W. Matthews, *Acta Cryst.* **A43**, 489 (1987).
40. R. Lavery and H. Sklenar, *J. Biomol. Struct. Dynam.* **6**, 63 (1988).
41. G. Ravishanker *et al.*, *J. Biomol. Struct. Dynam.* **6**, 669 (1989).
42. W. Saenger, *Principles of Nucleic Acid Structure* (Springer-Verlag, New York,
1984).
43. We thank Dr. D. L. Beveridge and M. Young for their help in the analysis of the
DNA and P.A. Stange for excellent technical assistance. Supported by March of
Dimes grant no. 505 and by Public Health Service grant GM 49244 to R.G.B., by a
National Defense Science and Engineering Graduate Fellowship in Biosciences to
M.A.S. and by Public Health Service grant GM 24658 to H.Z.. The coordinates are
in the process of being deposited in the Brookhaven Protein Data Bank but are
available from R.G.B..

TABLE 1 LEGEND:

Detailed conditions for crystallization of the PurR-hypoxanthine-*purF* operator complex are described elsewhere (34). Briefly, the crystals are grown by vapor diffusion from polyethylene glycol solutions and take the orthorhombic space group C222₁ with cell dimensions, $a = 175.85 \text{ \AA}$, $b = 94.79 \text{ \AA}$ and $c = 81.84 \text{ \AA}$. Intensity data were collected at room temperature with a San Diego Multiwire Systems (SDMS) Area Detector (35) and a Rigaku RU200-H rotating anode generator as the X-ray source set at 40 kV, 150 mA. The data were processed with software provided by SDMS. The structure was determined by Multiple Isomorphous Replacement (MIR) and revealed one PurR monomer-hypoxanthine-*purF* operator half site per asymmetric unit. Heavy atom parameters were refined and MIR phases were calculated with the program HEAVY (36) initially with data from 10.0 \AA to 3.0 \AA . The final figure of merit to 3.0 \AA was 0.59. The initial electron density map used for tracing was generated after solvent flattening (37) and revealed clear density for most of the protein backbone, the hypoxanthine and some of the DNA. FRODO (38) was used to build in a partial polyalanine model for the protein which was subsequently refined in real space with TNT (39). After 10 cycles of real space refinement, the partial model was used as a source of phases which were combined with the MIR phases to give a new figure of merit of 0.78. The resulting map showed clear density for most of the side chains and the DNA. At that point PurR residues, 3 - 330, and the *purF* half-site were fit. After 20 cycles of positional refinement using TNT, the R-factor dropped from 0.47 to 0.29. Several rounds of rebuilding were followed by more positional refinement, after which time electron density for residues 331-340 was found and the data were extended to 2.7 \AA . The R-factor converged at 0.235 at which time tightly restrained B-factor refinement was begun. At the present stage of refinement, the agreement factor of the model to crystallographic data is 0.195, for all data from 10.0 \AA to 2.7 \AA .

TABLE 1 FOOTNOTES:

*Indicates derivatives in which pseudo twofold related sites were both iodinated. †mHgCl₂ is an abbreviation for methyl-mercuric chloride. ‡ $2R_{\text{sym}} = \sum |I_o - \langle I \rangle| / I_o$, where I_o = observed intensity, $\langle I \rangle$ = average intensity obtained from multiple observations of symmetry-related reflections. § $R_{\text{iso}} = \sum |F_{\text{PH}}| - |F_{\text{P}}| / \sum |F_{\text{P}}|$, where $|F_{\text{P}}|$ = protein structure factor amplitude, $|F_{\text{PH}}|$ = heavy atom derivative structure factor amplitude. || Phasing power = r.m.s. ($|F_{\text{H}}| / E$), $|F_{\text{H}}|$ = heavy atom structure factor amplitude, E = residual lack of closure. ¶ $R_c = \sum |F_{\text{der}} \pm F_{\text{nat}}| - |F_{\text{H}(\text{calc})}| / \sum |F_{\text{der}} - F_{\text{nat}}|$ for centric reflections, where $F_{\text{H}(\text{calc})}$ is the calculated heavy atom structure factor. #Figure of merit = $\int P(\theta) \exp(i\theta) d\theta / \int P(\theta) d\theta$

where P is the probability distribution of θ , the phase angle. **R factor = $\Sigma|F_{\text{obs}} - F_{\text{calc}}|/\Sigma F_{\text{obs}}$. r.m.s. bond lengths and r.m.s. bond angles are the respective root-mean-square deviations from ideal values.

Table 1. Summary of crystallographic analysis

| | Native | Iodo-dC9 | Iodo-dU8* | Iodo-dU7 | Iodo-dU6* | mHgCl ₂ † | Thimerosal |
|---------------------------------------|--------|----------|------------|----------|-----------|----------------------|------------|
| Resolution (Å) | 2.7 | 2.8 | 3.0 | 2.8 | 3.0 | 3.0 | 3.0 |
| Unique reflections | 18,242 | 16,115 | 12,870 | 15,980 | 13,094 | 11,656 | 13,547 |
| Completeness (%) | 97 | 86 | 68 | 85 | 70 | 62 | 72 |
| I/σ(I) for data | 9.4 | | | | | | |
| from 10.0Å to 2.7Å | | | | | | | |
| I/σ(I) for data | 2.5 | | | | | | |
| from 2.8Å to 2.7Å | | | | | | | |
| R _{symm} (%)† | 5.6 | 6.8 | 10.2 | 7.1 | 8.8 | 9.0 | 8.1 |
| R _{iso} (%)§ | | 12.2 | 14.9 | 11.0 | 15.1 | 19.2 | 21.3 |
| Number of Sites | | 1 | 2 | 1 | 2 | 1 | 1 |
| Phasing power | | 1.27 | 0.89 | 1.25 | 1.18 | 1.02 | 1.01 |
| R _c ¶ | | 0.80 | 0.70 | 0.68 | 0.69 | 0.74 | 0.71 |
| Mean overall figure of merit (to) 3Å# | | | | | | | |
| | | | | | | | |
| Refinement statistics | | | | | | | |
| Resolution (Å) | | | 10.0 - 2.7 | | | | |
| R-factor** | | | 0.195 | | | | |
| Total number of atoms | | | 3007 | | | | |
| Water molecules | | | 20 | | | | |
| Rms deviations | | | | | | | |
| Bond angles (°) | | | 2.35 | | | | |
| Bond lengths (Å) | | | 0.012 | | | | |

TABLE 2 LEGEND :

a. Inter base pair and intra base pair parameters of the *purF* operator site calculated with CURVES (40). Inter base pair parameters refer to those between base pairs within the same strand and intra base pair parameters refer to those within a base pair. Typical roll angles, rise values and twist angles are 0°, 3.38 Å and 34.3° for B-DNA and 0°, 2.56 Å, and 32.7° for A-DNA (41,42).

b. Minor groove width of the *purF* operator site. Minor groove width is defined as the shortest O4'- O4' distance minus 2.8 Å (the van der Waals radii of two oxygen atoms). The average minor groove width for B-DNA is 5.7 Å and for A-DNA, 11.7 Å (41).

Table 2: DNA helical parameters of the *purF* operator half-site

a. Selected inter- and intra-base pair parameters

| Base Pair | Inter-Base Pair | | | | Intra-Base Pair | |
|-----------|-------------------|----------|----------|-----------|---------------------|------------|
| | Helical Twist (°) | Roll (°) | Rise (Å) | Slide (Å) | Propeller Twist (°) | Buckle (°) |
| 2 A·T | 5.69 | 14.36 | 3.60 | -1.02 | 7.27 | -19.20 |
| 3 C·G | 38.68 | -3.14 | 3.47 | -0.60 | -2.18 | -0.28 |
| 4 G·C | 49.69 | 2.87 | 3.48 | 0.35 | 1.01 | -0.49 |
| 5 A·T | 21.80 | -1.11 | 3.68 | -0.13 | -16.12 | 6.96 |
| 6 A·T | 41.84 | 0.71 | 3.52 | -0.01 | -1.74 | 3.59 |
| 7 A·T | 39.93 | 4.57 | 3.60 | -0.49 | -18.87 | 1.95 |
| 8 A·T | 16.93 | -1.65 | 2.84 | -1.17 | -22.93 | 7.17 |
| 9 C·G | 27.22 | 44.86 | 6.45 | 0.65 | -3.99 | 21.38 |
| 9' G·C | | | | | -3.99 | 21.38 |

←--DYAD

b. Minor groove width (Å) of *purF* operator site

| | | | | | | | | | |
|-------|-------|-------|-------|-------|-------|-------|-------|------|-------|
| T(1) | A(2) | C(3) | G(4) | C(5) | A(6) | A(7) | A(8) | C(9) | G(9') |
| | 9.26 | 10.31 | 9.46 | 8.50 | 7.80 | 8.98 | 9.75 | 9.20 | 9.20 |
| T(2') | G(3') | C(4') | G(5') | T(6') | T(7') | T(8') | G(9') | C(9) | |

Fig 1. (A) Sequence of the duplex deoxyoligonucleotide used in cocrystallization. The numbering scheme is used throughout the text and the locations of iodine-labelled bases used in structure determination are identified by an asterisk (*). Shaded is the half-site used in the design of the perfect palindrome site. **(B)** Sequence of the perfect palindromic *purF* site. The pseudo-dyad and dyad axes of each sequence is indicated by the symbol(-).

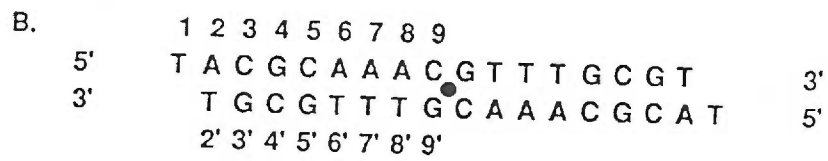
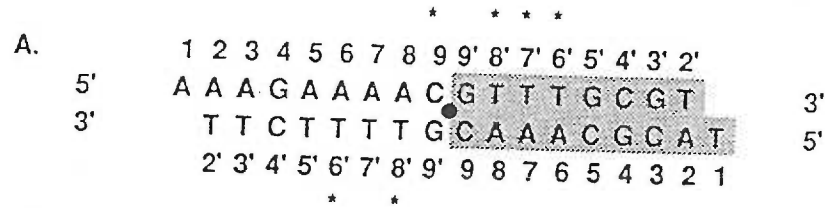


Fig 2. Electron density showing how the crystallographically related "leucine levers" (Leu54 and Leu54') intercalate into the minor groove and are wedged between the central Cyt9pGua9' base pairs. The 45° roll of the central Cyt9pGua9' base-pair step is evident.

(A) The refined $2(F_{\text{obs}} - F_{\text{calc}})$ electron density map contoured at 1.5° .

(B) Omit map ($F_{\text{obs}} - F_{\text{calc}}$) with nucleotides Cyt9, Gua9' and residue Leu54 omitted from the model refinement. The contour level is 3.7° .

A

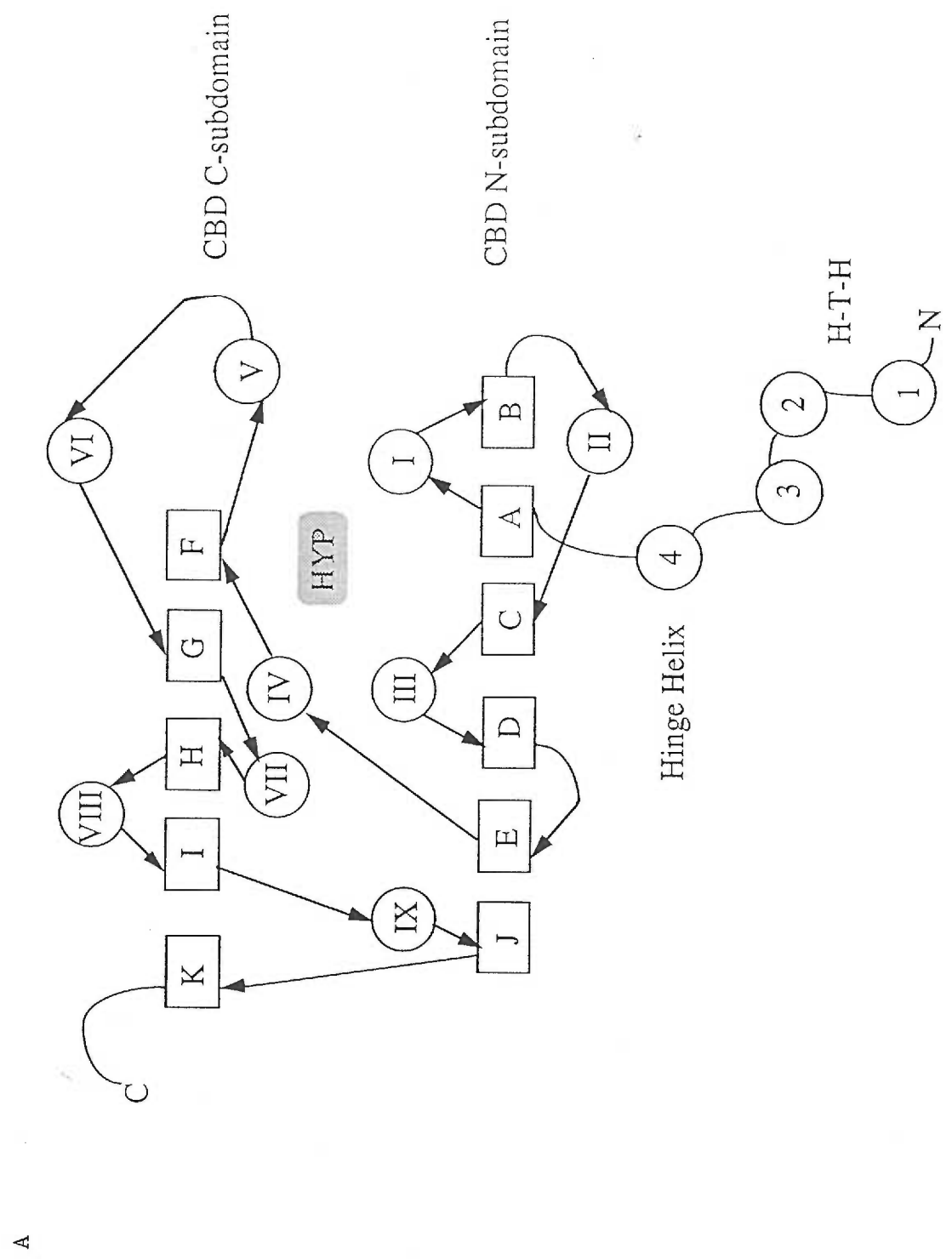


B



Fig 3. (A) Topology diagram of the PurR monomer. α helical segments are shown as circles and β strands as squares. The NH_2 -terminus and COOH -terminus of the protein are identified with N and C respectively. The DNA-binding domain consists of (residues enclosed in parentheses) helix 1 (4-10) and helix 2 (15-23) (the HTH, labelled), helix 3 (30-43) and helix 4 (the hinge helix, labelled) (48-56). The corepressor binding domain or CBD consists of a CBD NH_2 -subdomain (labelled), which contains strand A (61-66), helix I (72-88), strand B (91-96), helix II (101-113), strand C (118-121), helix III (128-134), strand D (142-147), strand E (156-159), helix IX (294-310) and strand J (318-320), and a CBD COOH -subdomain (labelled), helix IV (162-174), strand F (181-184), helix V (191-204), helix VI (223-234), strand G (241-245), helix VII (248-260), strand H (270-275), Helix VIII (280-282), strand I (288-291) and strand K (324-326). The hypoxanthine corepressor is represented by a shaded oval.

(B) Stereo view of the PurR-hypoxanthine-DNA complex. The DNA is shown as yellow stick bonds and the hypoxanthine as blue stick bonds. The PurR dimer is represented by a ribbon with one subunit colored green and the other red. The secondary structural elements are indicated as in white for one monomer subunit. (figure generated with Biosym Insight II).



B

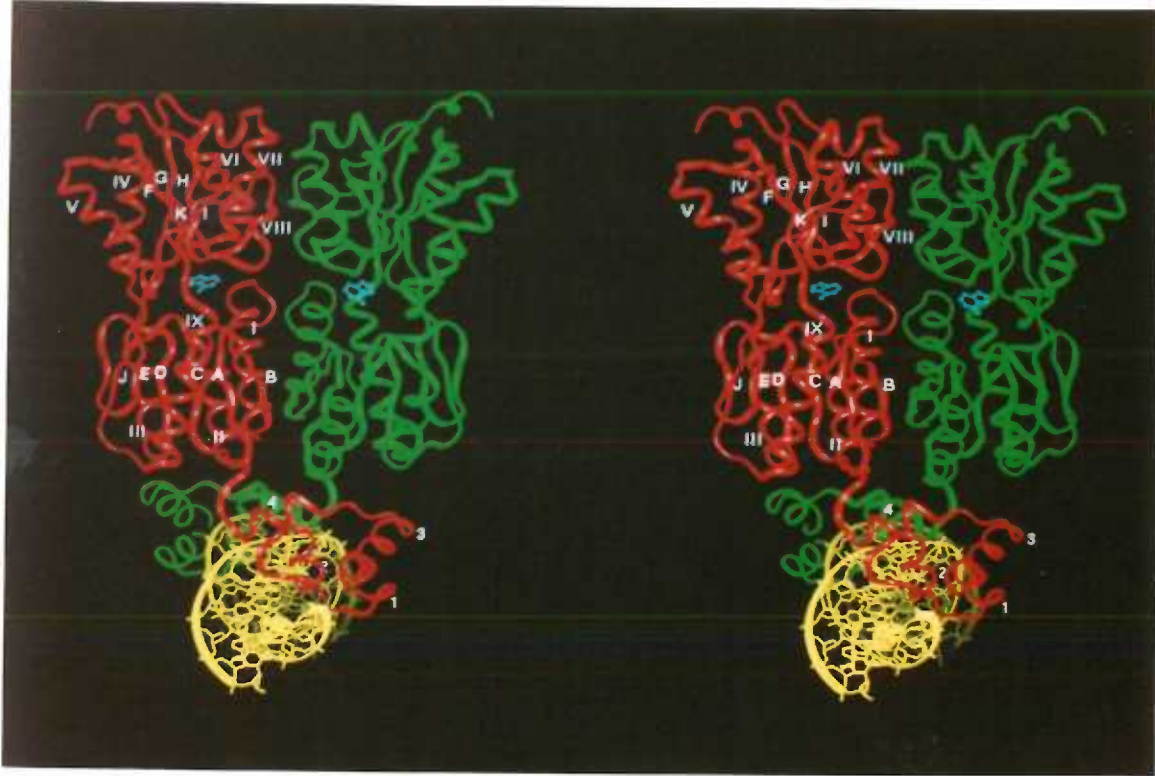


Fig 4. Stereo diagram of the corepressor binding pocket. The hypoxanthine molecule is shown in red. PurR residues involved in hypoxanthine binding and relevant hydrogen bonds are indicated in yellow. Also shown in yellow and represented as a sphere is a water molecule, Wat1, involved in hypoxanthine binding. The architecture of the purine binding pocket provides insight into the ability of PurR to bind guanine with high affinity. Water molecule, Wat2 (in purple), is located within 4.34 Å of the hypoxanthine C2 and hydrogen bonds to the carboxylate side chain of Glu222 (3.49 Å). A simple rotation of this side chain would provide a hydrogen bond to the guanine N2. (figure generated with Biosym Insight II).

Fig 5. Superposition of the HTH motifs (helices 1 and 2) of PurR (magenta) and LacI (blue). Also shown is the side chain of PurR residue Asn12, which is found at the invariant glycine position of the turn. Although not included in the calculation, the C α atoms of the loop that follows helix 2, helix 3 and the residues that follow show a high degree of structural overlap. (figure generated with Biosym Insight II).

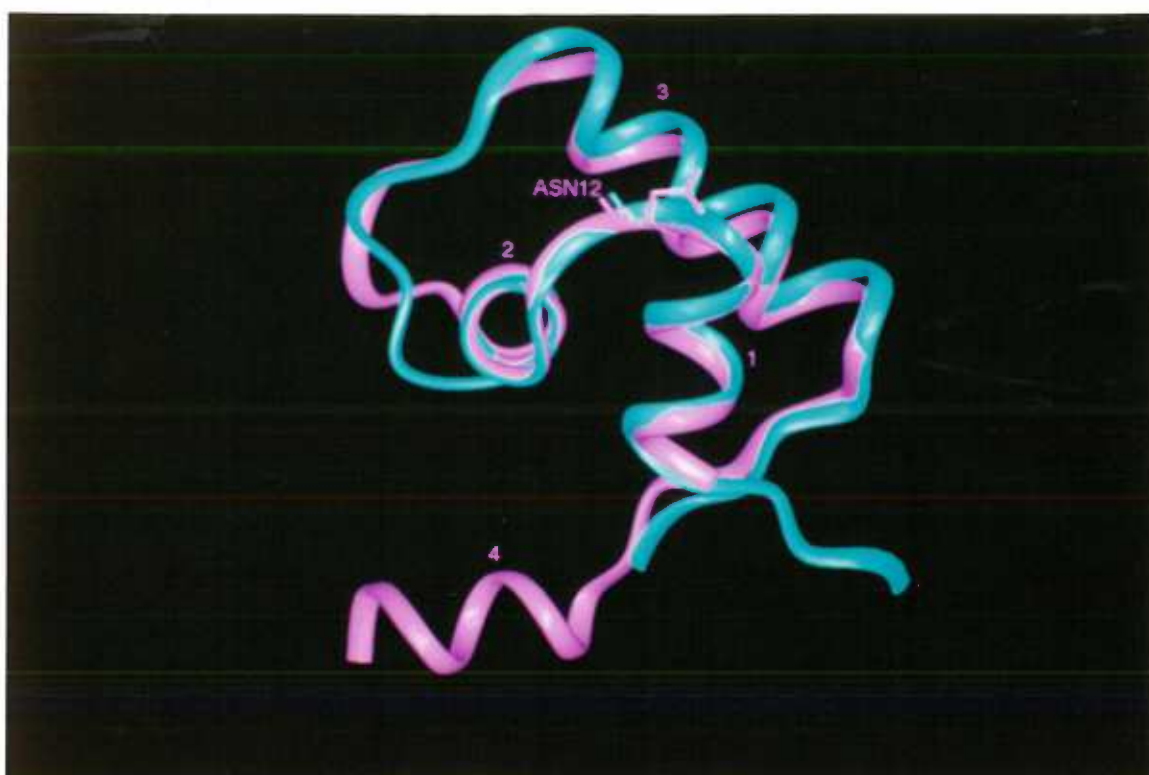


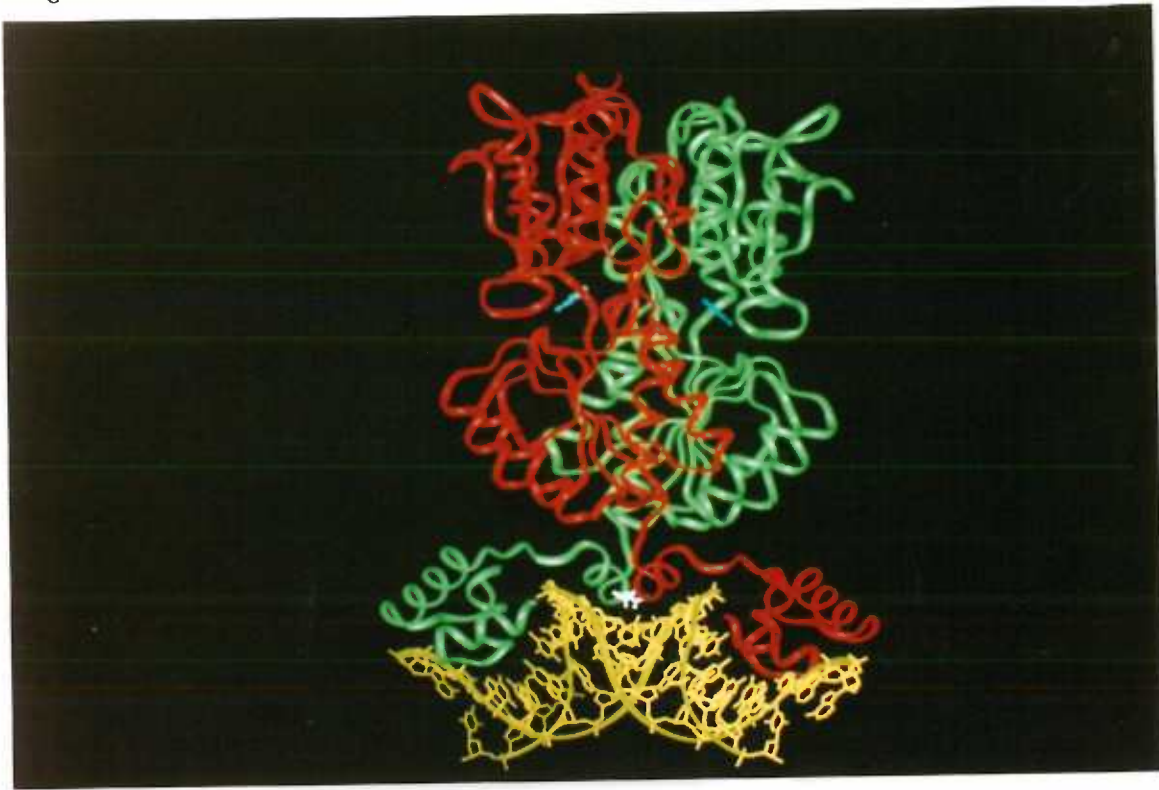
Fig 6. PurR-DNA interactions. (A) Stereo representation showing DNA base contacts made by the DNA-binding domain of one PurR monomer. The DNA is shown in magenta stick bonds and PurR residues 3-56, which includes the helices 1 through 3 and the hinge helix, are represented by a blue ribbon. Residues contacting the DNA are depicted as blue sticks. Selected hydrogen bond distances are shown in blue. A blue sphere representing a water molecule which mediates the contact between Thr15 and Thr15-Thy7' is also shown.

(B) Schematic view of the PurR-DNA interaction. The DNA is represented as a cylindrical projection. Contacts made to the major groove (M) are differentiated from contacts made to the minor groove (m). Bases involved in van der Waals interactions are shaded light gray, bases involved in hydrogen bonds are shaded medium gray and bases involved in both are shaded dark gray. Deoxyribose and phosphate groups contacted from either the major or minor groove are filled in with stripes while phosphate and sugar groups contacted from both the major and minor grooves are hatched. The possible Thr16-Cyt5 interaction is flagged by an asterisk.

(C) Overall view of the PurR-hypoxanthine-DNA complex. The DNA is represented as yellow stick bonds, the hypoxanthine molecules as blue balls and sticks, and PurR as a ribbon with one monomer subunit green and the other red. Shown as white sticks are the side chains of Leu54 and Leu54'. The expansion of the minor groove and severe kink caused by their insertion is evident.

(D) View of the two-fold related hinge helices and the minor groove of the *purF* operator. The crystallographic two-fold axis is perpendicular to the plane of the figure and bisects the central CpG' base pair step. Shown as blue sticks are key hinge helix residues, Val50 and Val50' (V50 and V50'), Leu54 and Leu 54' (L54 and L54'), and Ala51 and Ala51' (A51 and A51'). The two HTH motifs are located in the major grooves in the lower left and upper right of the figure. (figure generated with Biosym Insight II).

C



D

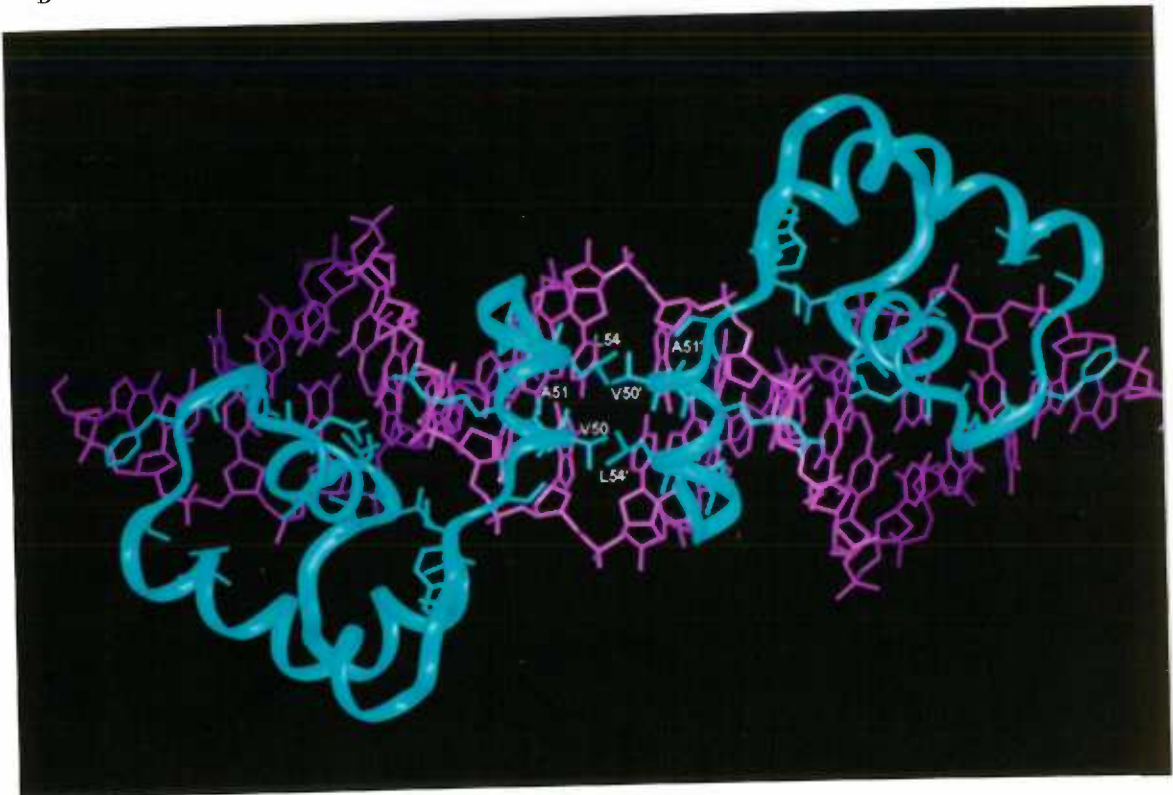
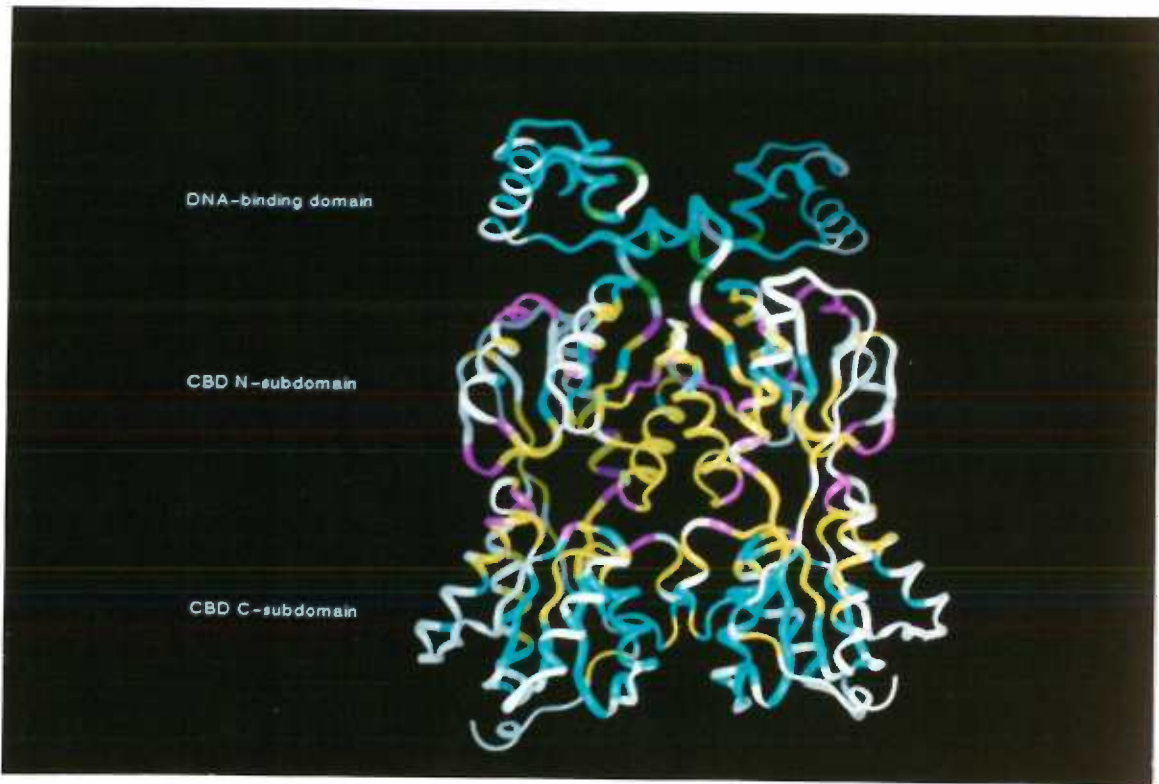


Fig 7. LacI family member-operator site specificity. Residues of ten LacI family members and the base pairs they contact or are predicted to contact are shaded similarly. Specifically, residue 2 from helix 2, the recognition helix, contacts base pair 6; residue 6 from helix 2 contacts base pair 4, and LacI family member residues corresponding to PurR residue Lys55 (indicated by asterisk) contact base pair 8. Upper, consensus operator half sites from LacI family members, the operators of which have been well characterized. Q = any base; O = Ade or Cyt; R = Gua or Thy; S = Ade or Thy; Y = Cyt or Thy. The four letter abbreviations for the LacI family members not provided in the text are: the galactose isorepressor (GalS), amylase repressor (CcpA), maltose repressor (MalI), rhamnose repressor and sucrose repressor (ScrR). Lower, sequences of the putative or known HTH motifs of these LacI family members.

Fig 8. LacI mutants mapped onto the three-dimensional structure of PurR (based on our previous sequence alignment (6)). Labelled are the DNA-binding domain and the CBD N- and COOH-subdomains of PurR. The positions of I^r mutants are blue, I^s mutants are magenta, I^{tb} are green and mutations that lead to both the I^r and I^s phenotype are yellow.



Manuscript #4**Crystallization and Preliminary X-ray Studies on the Co-repressor
Binding Domain of the *Escherichia coli* Purine Repressor**

Maria A. Schumacher¹, Kang Yell Choi², Howard Zalkin²
and Richard G. Brennan¹

¹Department of Biochemistry and Molecular Biology
Oregon Health Sciences University
Portland, OR 97201-3098

²Department of Biochemistry
Purdue University
West Lafayette, IN 47907

ABSTRACT

The purine repressor is a putative helix-turn-helix DNA-binding protein that regulates several genetic loci important in purine and pyrimidine metabolism in *Escherichia coli*. The protein is composed of two domains, an NH₂-terminal DNA-binding domain and a COOH-terminal core that binds the co-repressor, guanine and hypoxanthine. The co-repressor binding domain (residues 53-341) has been crystallized from polyethylene glycol 600-MgCl₂ solutions. They are of the monoclinic form, space group P2₁, with $a = 38.2 \text{ \AA}$, $b = 125.7 \text{ \AA}$, $c = 61.8 \text{ \AA}$ and $\beta = 100.2^\circ$. They diffract to a resolution of at least 2.2 \AA and contain two monomers per asymmetric unit. The importance of the structural determination of this domain is underscored by the high degree of sequence homology displayed within the effector binding sites among a sub-class of helix-turn-helix proteins, of which LacI and GalR are members. The structure of the PurR co-repressor binding domain will provide a high resolution view of one such domain and could serve as a possible model for future effector site structural determinations. Perhaps more important will be this structure's contribution to the further understanding of how protein-DNA interactions are modulated.

De novo synthesis of purine nucleotides requires, overall, 14 enzymatic steps to convert 5-phosphoribosyl-1-pyrophosphate to AMP or GMP. In *Escherichia coli*, the purine repressor regulates seven known loci that encode enzymes involved in this synthesis as well as four other loci encoding enzymes for pyrimidine nucleotide synthesis and salvage (Rolfes & Zalkin, 1990a,b). In addition, *purR* is under autogenous control (Rolfes & Zalkin, 1990a,). The *purR* gene, mapped to co-ordinate 1755 kb on the *E. coli* restriction map (Kohara *et al.* 1987), encodes a protein of 341 amino acid residues ($M_r = 38,179$) that consists of an NH₂-terminal helix-turn-helix DNA-binding domain (Brennan & Matthews, 1989) and a COOH-terminal purine co-repressor binding domain. Repression requires the presence of the purine co-repressors, hypoxanthine and guanine, which have apparent dissociation constants of 9.1 mM and 1.7 mM respectively (Rolfes & Zalkin, 1990b).

After overexpression and purification of PurR, a proteolytic fragment containing the co-repressor binding domain can be prepared by tryptic cleavage at arginine 52. This product runs as a single band on SDS/PAGE ($M_r = 32,487$). We report here crystallization conditions for this proteolytic fragment that give rise to a crystal form that diffracts to at least 2.2 Å (1 Å = 0.1 nm). The method used is hanging drop, vapor diffusion. Protein, at 12 mg/ml, is mixed 1: 1 with 20% PEG 600, 0.1 M TrisHCl (pH = 7.6), 0.2 M MgCl₂ and allowed to stand for one week at room temperature. The resulting crystals are elongated rods that can grow to dimensions of 1.0 mm X 0.5 mm X 0.4 mm. Precession photographs reveal that the crystals are monoclinic and take the space group P2₁ (Fig. 1). Unit cell dimensions are $a = 38.2$ Å, $b = 125.7$ Å, $c = 61.8$ Å and $\beta = 100.2^\circ$. The crystals are stable to 50 hours' exposure to graphite-monochromated CuK α X-rays using a Rigaku RU-200 X-ray generator set at 50 kV and 100 mA with a collimator size of 0.75 mm. Assuming one monomer per asymmetric unit (ASU), the crystal packing parameter, V_m is 4.49 Å³/dalton. This falls outside the normal values (2.0 to 3.0 Å³/dalton) for protein crystals (Matthews, 1968). However, assuming two monomers per asymmetric unit, the V_m is 2.25 Å³/dalton. Therefore, it is likely that there are two PurR co-repressor monomers, possibly a dimer, occupying the ASU. This is also consistent with the finding that PurR is dimeric both in the presence and absence of co-repressors (Choi & Zalkin, unpublished results) and with the self-Patterson rotation function which identified a non-crystallographic two-fold at $\phi=33.9^\circ$, $\varphi=78.9^\circ$ and $\kappa=180^\circ$. (However, the self-Patterson rotation function failed to identify any obvious noncrystallographic symmetry axis. This failure may reflect slight differences in the relative orientations of the putative N and COOH-terminal domains of the two PurR monomers, which are likely connected by a

flexible hinge region as observed in the evolutionarily related periplasmic binding proteins (Vartak *et al.*, 1991; Spurlino *et al.*, 1991).)

Attempts are being made to soak the sparingly soluble co-repressors into the large fragment crystals as well as to co-crystallize them *de novo*. Small crystals of the intact form have also been grown under different conditions but are not yet of data quality. In conjunction with this, crystallization of the holorepressor-DNA complex is being attempted. Collection of high-resolution native three-dimensional data for the large fragment crystal form and screening for isomorphous derivatives are underway.

The X-ray structure of the large fragment crystals will provide, at atomic resolution, the details of the co-repressor binding domain. Recent protein homology analysis suggests that the co-repressor binding domain of PurR may be structurally similar to the sugar (inducer) binding domain of several DNA-binding proteins containing the helix-turn-helix motif, including the Lac, Cyt, Gal, Mal, Raf, and Rbs repressors (Rolfes & Zalkin, 1990a). Surprisingly, this region also appears to be homologous to several periplasmic binding proteins, indicating a close evolutionary relationship (Vartak *et al.*, 1991). The high-resolution three-dimensional structures of several of these proteins have been elucidated by X-ray crystallography (Vyas *et al.*, 1991; Spurlino *et al.*, 1991; Vyas, 1991; Quioco, 1986). If, as suggested by the sequence homology, PurR and the periplasmic binding proteins are structurally similar, molecular replacement using structurally one or more of these may serve to hasten the structure determination of the co-repressor binding domain of the purine repressor. This structure may provide an important first step in the further understanding of the mechanisms by which effector molecules influence the affinity of these helix-turn-helix proteins their cognate DNA.

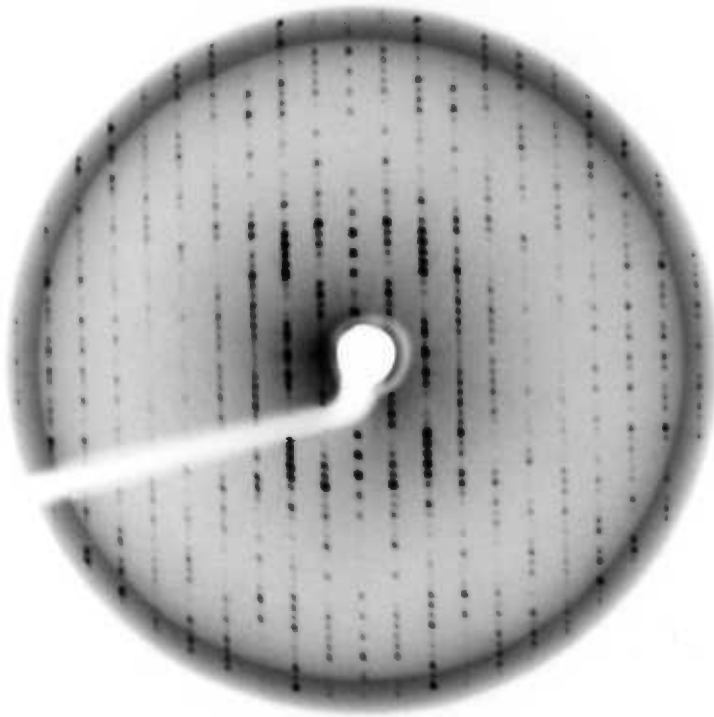
Acknowledgments

This work was supported by Public Health Service grant GM24658 (H.Z.) and by Basil O'Connor Starter Scholar Research Award no. 50726 from the March of Dimes Birth Defects Foundation (R.G.B.) and a grant from the Oregon Community Foundation (R.G.B.).

REFERENCES

- Brennan, R. G. & Matthews, B. W. (1989). The helix-turn-helix DNA binding motif. *J. Mol. Chem.* **264**, 1903-1906.
- Kohara, Y., Akiyama, K. & Isono, K. (1987). The physical map of the whole *E. coli* genomic library. *Cell*, **50**, 495-508.
- Matthews, B. W. (1968). Solvent content of protein crystals. *J. mol. biol.* **33**, 491-497.
- Quioco, F. A. (1986). Carbohydrate-binding proteins: tertiary structures and protein-sugar interactions. *Annu. Rev. Biochem.* **55**, 287-315.
- Rolfes, R. J. & Zalkin, H. (1990a). Autoregulation of *Escherichia coli* purR requires two control sites downstream of the promoter. *J. Bacteriol.* **172**, 5758-5766.
- Rolfes, R. J. & Zalkin, H. (1990b). Purification of the *Escherichia coli* purine regulon repressor and identification of corepressors. *J. Bacteriol.* **172**, 5637-5642.
- Spurlino, J. C. Lu, G.-Y. & Quioco, F. A. (1991). The 2.3 Å resolution structure of the maltose- or maltodextrin-binding protein, a primary receptor of bacterial active transport and chemotaxis. *J. Biol. Chem.* **226**, 5202-5219.
- Vartak, N. B., Reizer, J., Reizer, A., Gripp, J. T., Groisman, A. E., Wu, L., Tomich, J. M. & Saier, M. H. (1991). Sequence and evolution of the FruR protein of *Salmonella typhimurium*: a pleiotropic transcriptional regulatory protein possessing both activator and repressor functions which is homologous to the periplasmic ribose binding protein. *Res. Microbiol.* **142**, 951-964.
- Vyas, N. K. (1991). Atomic features of protein-carbohydrate interactions. *Curr Opinions Struct. biol.* **1**, 732-740.
- Vyas, N. K., Vyas, M. N. & Quioco, F. A. (1991). Comparison of the periplasmic 5226-5237.

Figure 1. $12^\circ hk0$ precession photograph of the co-repressor binding domain of the purine repressor from *E. coli* using Ni-filtered $\text{CuK}\alpha$ X-rays set at 30 kV, 30 mA and a 1.0 mm collimator. Exposure time is 18h.



Manuscript #5:**Ligand-Free Purine Repressor has an Open Conformation:
Implications for DNA Binding**

Maria A. Schumacher¹, Kang Yell Choi², Fu Lu², Howard Zalkin²
& Richard G. Brennan^{1*}

¹Department of Biochemistry and Molecular Biology
Oregon Health Sciences University
Portland, Oregon 97201-3098, U.S.A
Telephone: (503) 494-4427
FAX: (503) 494-8393

²Department of Biochemistry
Purdue University
West Lafayette, Indiana 47907, U.S.A.

*Person to whom correspondence should be addressed.

ABSTRACT

The three dimensional structure of the ligand-free Corepressor Binding Domain (CBD) of the Purine Repressor, PurR, residues 53 to 341, has been determined to 2.2 Å resolution by x-ray crystallography. In its unliganded state the globular subdomains of each monomer subunit of the CBD dimer are rotated open by 17° and 23° relative to corepressor-bound CBD. These rotations lead to the generation of a twelve-stranded, intermonomer β-sheet between the NH₂-subdomains in ligand-free CBD and a major restructuring of the corepressor binding pocket. Comparisons of the structures of the ligand-free and ligand-bound CBD suggests a mechanism by which corepressor binding activates PurR to bind its cognate DNA.

The purine repressor, PurR, is a 341 amino acid residue dimer that functions as the master regulator of de novo purine biosynthesis and to a lesser extent, de novo pyrimidine biosynthesis in *Escherichia coli* by repressing transcription of at least 14 genes (1). Specific DNA binding by PurR requires its binding of a purine corepressor, either guanine (K_d of 1.5 μM) or hypoxanthine (K_d of 9.3 μM) (2). Such modulation of the affinity of DNA-binding proteins for their cognate DNA sites by small molecule effectors is common in both prokaryotes and eukaryotes and was first demonstrated for the lactose repressor (LacI) (3). PurR shares significant sequence homology with LacI and together with at least 21 other bacterial transcription repressors forms the LacI family (4). All members of the LacI family are regulated by small molecule effectors that act either as inducers or corepressors. The high sequence homology displayed by LacI family proteins suggests their structural similarity (4). However, a full understanding of their effector induced-DNA binding mechanisms has been hampered by the lack of structural information.

Recently we reported the crystal structure determination of a PurR-hypoxanthine-*purF* operator ternary complex (5). The structure revealed that PurR consists of an NH₂-terminal DNA-binding domain (residues 3 to 56) that contains Helix-Turn-Helix (HTH) and hinge helix DNA-binding motifs, and a larger COOH-terminal Corepressor Binding Domain (CBD) (residues 60 to 341) that is responsible for corepressor binding and dimerization. The CBD of each monomer subunit is strikingly similar in structure to the monomeric bacterial periplasmic binding proteins (PBPs) (6), a large family of proteins that functions in active transport and in some cases chemotaxis, the dimeric transcription anti-terminator regulator protein, AmiC (7), and lactoferrin, which functions in iron transport in mammals (8). The PBPs exist in an equilibrium between "open" and "closed" forms, the latter of which is stabilized by ligand binding (9). The large, predominantly quaternary changes between the open and closed forms of these proteins are key in mediating the signal transduction functions of these proteins, that is, only the closed, ligand-bound form can transduce signal.

PurR also functions in signal transduction, however, for PurR as well as the other LacI family members, the intracellular signal is transduced first intramolecularly from the effector binding domain to the covalently attached DNA-binding domain and second intermolecularly to affect specific DNA binding. In the case of PurR, purine binding signals the overabundance of hypoxanthine or guanine and acts as an environmental switch to stop de novo synthesis of purines. Biochemical studies have suggested that the CBD of PurR undergoes a conformational change upon binding corepressor (10). However, whether the ligand-free, dimeric CBD of PurR, or the effector binding domain of any other LacI family member, undergoes the type and magnitude of conformational transition

observed in the monomeric PBPs (9), remains an open question. Clearly, the structural mechanisms of signal transduction, evoked by the presence or absence of effector molecule, and transcription regulation are tightly linked.

To assess more fully the conformational changes that occur in the absence of corepressor and how these changes affect the ability of PurR to bind DNA, we have determined the structure of the corepressor-free (unliganded) Corepressor Binding Domain of PurR (11). The structure reveals substantial structural changes and hinge-bending rotations of 17° and 23° between the NH₂- and COOH-subdomains of each monomer subunit relative to the corepressor bound form (5) that results in an "open" CBD conformation. As a direct consequence of these rotations the hinge helices, which make key minor groove interactions and are responsible for kinking the DNA (5), would be pulled apart and destabilized. This destabilization would also result in the relocation of the HTH motifs such that they could no longer be in position to bind in successive major grooves (5). Combined, these changes would lead to the complete disruption of specific DNA binding. A similar mechanism of effector-regulated specific DNA binding can be envisioned for the entire LacI family.

The corepressor binding domain of PurR was prepared by trypsin-cleavage and results in a protein that contains the entire corepressor binding and dimerization domain (residues 60 to 341) but is extended slightly to include residues 53 through 59 of the DNA-binding domain (12). The structure of the CBD was solved by molecular replacement (13) (Table 1). There is one CBD dimer in the asymmetric unit and the present model consists of residues 60-187 and 194-341 of each CBD monomer with residues 53 to 59 and 188 to 193 being disordered, 243 water molecules and 3 magnesium ions. The current R-factor is 15.6% using data from 10.0 - 2.2 Å. The model displays excellent stereochemistry (14) (Table 1). A typical section of the current $2F_o - F_c$ electron density map is shown in Figure 1.

The CBD consists of two structurally similar, noncontiguous subdomains per monomer, the CBD NH₂-subdomain and CBD COOH-subdomain, that are connected by three short crossover strands (Fig. 2). The purine corepressor binding pocket is found between the two subdomains. The tertiary structure of the unliganded CBD COOH-subdomain is virtually identical to that of its hypoxanthine-bound form as revealed by the root mean square deviations (rmsds) of 0.34 Å and 0.40 Å between the C α atoms of residues 162-187, 194-290 and 324-326 of each unliganded CBD COOH-subdomain and the liganded CBD COOH-subdomain (5) (Fig. 3A). Moreover, the CBD COOH-subdomain dimerization interface is maintained as the COOH-subdomain monomers change little relative to each other (2.4° rotation of each COOH-subdomain monomer). The rmsd

between the C α atoms of the unliganded CBD COOH-dimer and its liganded counterpart is 0.60 Å (Fig. 3B).

Unlike the CBD COOH-subdomains, the CBD NH₂-subdomains undergo large structural changes upon corepressor binding with rmsds of 1.72 Å and 1.61 Å between all 146 C α s of the liganded and unliganded NH₂-subdomain (residues 60-159 and 296-321). A similar superimposition using only the C α s of residues 60-146 and 296-321, results in an rmsd of 1.3 Å and 1.2 Å (see Fig. 3 legend) (15). Furthermore, the superimposition of the NH₂-subdomain dimer of unliganded CBD, using only the C α s of residues 60-146 and 296-321 of each subunit onto the corresponding atoms of liganded CBD results in an rmsd of 3.5 Å. This underscores the large quaternary change between each NH₂-subdomain relative to the other in going from the ligand-free to the ligand-bound form.

Since the CBD COOH-subdomains of corepressor-free and bound CBD are virtually identical, they can serve as a reference frame to examine any conformational changes that occur upon ligand binding. Superimposition of the 126 corresponding C α atoms of each unliganded COOH-subdomain onto their liganded CBD counterparts, reveals that, despite the dimeric nature of PurR, the NH₂-subdomains have undergone rotations of 17° and 23° that result in an open conformation relative to the closed conformation of the hypoxanthine-bound CBD (Fig. 3). Thus, the CBD COOH-subdomain dimer acts as a structural anchor for the large rotations of the NH₂-subdomains. These rotations are a direct consequence of the hinge bending that arises primarily from changes in the backbone torsion angles of residues in the CBD subdomain crossovers. Residues Asp160 and Asn161 of the first crossover, show the largest changes in which $\Delta\psi$ is 61.2° and $\Delta\phi$ is 14.5° for Asp160 and $\Delta\phi$ is 143.2° and $\Delta\psi$ is 26.8° for Asn161. The rotations result in a restructuring of the NH₂-subdomain dimerization interface in ligand-free CBD that includes the formation of two new hydrogen bonds between β -strands B and B' (where the prime indicates the other monomer). These hydrogen bonds lead to the generation of a twelve-stranded, intermonomer β sheet in which the six strands of each monomer are parallel but their interfacing B and B' strands are antiparallel (Fig. 1). Specifically, new hydrogen-bonds are formed between the carbonyl oxygen of Val92 and the amide of Gly94' and the amide of Val92 and the carbonyl oxygen of Gly94' as well as the related two-fold contacts. The intersubunit sheet is further stabilized by two hydrogen bonds between the side chain OH of Tyr107 and the carbonyl oxygen of Tyr90' and their dyad-related mates. In corepressor-bound CBD there are only two B-B' hydrogen bonds, which occur between the carbonyl oxygen of Val92 and the amide of Gly94' and the dyad-related interaction, and Tyr107 stacks against Phe86' and hydrogen bonds to Glu82', both located on helix I.

The largest local structural changes found between ligand-free and bound CBD are confined to the corepressor binding pocket and involve corepressor-binding residues Tyr73 and Phe74 from the CBD NH₂-subdomain and Arg190 and Thr192 from the CBD COOH-subdomain (5). Residue Arg190, which makes two specificity determining hydrogen bonds to the corepressor, resides on a loop (residues 185 to 190), the specificity loop, that, in part, becomes disordered in the absence of ligand (Fig. 2). Residue Thr192, which makes a hydrogen bond and van der Waals contacts to the corepressor, is the part of helix V (residues 191 to 204) that, in ligand-free CBD, unwinds by about a turn, leaving residues 191 through 193 disordered. Ligand dissociation also results in the loss of hydrogen bonds between Arg190 and Glu125 of the CBD NH₂-subdomain and between Asn191 and crossover residue Asp160. In liganded CBD these additional contacts act to further brace this region and their loss, upon ligand dissociation, leaves this region untethered. The disordering of residues 188 to 193 is the only structural change within the COOH-subdomains between the ligand-free and bound forms.

The NH₂-subdomain, on the other hand, undergoes several large changes. One is the rearrangement of the β -sheet contact region between β -strands D (residues 142 to 147) and E (residues 156 to 159) whereby the two hydrogen bonds observed in hypoxanthine-bound PurR, from the amide NH of Val144 to the carbonyl of Asp156 and the amide NH of Asp156 to the carbonyl of Lys142, are replaced by four new hydrogen bonds found between the amide NH of Val144 to the side chain O δ 2 of Asp156, the amide NH of Asp146 to the carbonyl of Asp156, the amide NH of Asp156 to the carbonyl of Val144, and the amide NH of Val158 to the carbonyl of Asp146. Despite the ratcheting of the β -sheet hydrogen bonds between β -strands D and E, the β -sheet hydrogen bonds between β -strands D and C and those between β -strands E and J are the same as those in the corepressor-bound form of the CBD. Linked to this structural rearrangement are the structural changes observed directly within the corepressor binding pocket that involve residues Tyr73 and Phe74, located at the NH₂-terminus of helix I, and residues 147 to 159, which include the COOH-terminus of β -strand D, the loop that follows, and β -strand E (Fig. 4). In ligand-free CBD, Trp147 no longer stacks against Tyr126 but has rotated into the ligand binding pocket (Fig. 4) resulting in translations of 3.3 Å and 10.7 Å of its C α and N ϵ atoms, respectively. This rotation also pulls residues 148 through 159, as a unit, toward the ligand binding pocket and accounts for the ratcheting motion of β -strands D and E. The movement of residues 147 through 159 as a unit separate from residues 60-146 is necessitated by hinge bending. Indeed, superimposition of the corresponding C α s of the entire NH₂-subdomain (after the superimposition of the liganded and unliganded CBD COOH-subdomains), reveals that were residues 60 through 159 to move as a rigid

body in going from the unliganded to the liganded state, NH₂-subdomain loop residues 148 through 155 would collide with the COOH-subdomain specificity loop.

In ligand-free CBD, the indole moiety of Trp147 stacks against Tyr73 and Phe74 and its Ne forms a hydrogen bond to the side chain OH of Tyr73 (Fig. 4). In this way, Trp147 acts as a structural, but not functional, replacement for the corepressor. In addition to the hydrogen bond between Trp147 and Tyr73, the Tyr73 hydroxyl group also hydrogen bonds to the side chain of cross-over residue Asp160. Because residue Asp160 is located on the "outer edge" of the corepressor binding pocket, the hydrogen bond between this residue and Tyr73 stabilizes the open conformation. In the ligand-bound form, the side chain OH of Tyr73 switches hydrogen bonding partners to the side chain Oe of Gln292 of the second cross-over, as does the side chain of Asp160 which is hydrogen bonds to the side chain of Asn191 from the specificity loop (Fig. 4). These different sets of interactions directly link corepressor binding to the hinge bending motion of the CBD and suggest a possible mechanism for corepressor binding. In this mechanism, residues Tyr73, Phe74 and Trp147 create a solvent accessible hydrophobic patch that would serve as the initial binding site for corepressor (Fig. 4). Upon corepressor binding, Trp147 would be displaced from the purine binding pocket as would the region encompassing residues 148 to 159. This displacement would lead directly to the disruption of the key hydrogen bond between Tyr73 and Asp160. As a consequence, Asp160 would be free to make a hydrogen bond to Asn191, which would aid in the concomitant ordering of the specificity loop. This mechanism suggests that corepressor would bind first to the NH₂-subdomain, which is consistent with crystallographic studies on the structurally related periplasmic binding protein LIVBP (leucine-isoleucine-valine binding protein) that show when leucine is soaked into crystals of a locked open form, the substrate interacts only with residues of the NH₂-subdomain (16).

Recently, structures for the open and closed forms of two PBPs, the Maltose Binding Protein (MBP) and the Lysine-Arginine-Ornithine Binding Protein (LAOBP), have been described (9). These structures allow us to compare and contrast the open and closed conformations of the CBD of dimeric PurR to the corresponding conformations of these monomeric PBPs. In general, the CBD of PurR and MBP and LAOBP display similar ligand-free open and ligand-bound closed conformations. However, the 17° and 23° rotations between the subdomains of the CBD are much smaller than those taken by MBP (35°) and LAOBP (53°). Also, unlike the PBPs (17), the subdomains of the CBD display little twisting relative to each other. Although we cannot eliminate the possibility of larger rotations, modelling demonstrates that rotations on the order of 35° for each CBD monomer subunit would disrupt completely the NH₂-subdomain dimerization interface. A

second key difference between the monomeric PBPs and PurR is that the periplasmic binding proteins show little structural rearrangements in going from the open to closed conformation. The structural changes observed in PurR may compensate for the lack of twisting, which, in the PBPs, serves to open further and reorient the two globular domains. A twisting motion like those observed for the PBPs would disrupt the dimerization interface in PurR.

The conformational changes in the PBPs allow the membrane-bound components of active transport or chemotaxis to differentiate between the liganded forms, which are bound by these proteins, and the unliganded forms, which are not bound (18). Similarly, the intracellular environmental signal of excess purine is transduced via PurR which binds *pur* regulon operators under physiological conditions only in its liganded (closed) form. In addition to allowing PurR to bind specific DNA sites, the structural changes in the CBD NH₂-subdomains present the possibility for additional effector-controlled protein-protein interactions. Although PurR is not known to interact with other proteins, the cytidine repressor, CytR, interacts with the catabolite gene activator protein, CAP, only in the absence of inducer (19-22), leading to repression of transcription of the CytR regulon (23), and the amylase repressor, CcpA, binds phosphorylated HPr (24).

To date, the "activation" of specific DNA binding by repressor through corepressor binding has been characterized structurally for only the *E. coli* Trp (25) and MetJ repressors (26). For the Trp repressor, the L-tryptophan corepressor orients the HTH reading heads to fit into successive major grooves. In addition, the L-tryptophan corepressor makes DNA phosphate contacts (25). Corepressor (S-adenosylmethione) binding to the MetJ repressor, on the other hand, does not cause a conformational change and corepressor activation appears to arise from electrostatic effects (26). Unlike these proteins, which are similar in that they bind corepressor within their DNA binding domains, PurR binds its corepressor in a separate domain, the CBD, at a distance of at least 40 Å from the DNA binding domain (over 50 Å from the DNA). Although the structure of the unliganded CBD does not include the DNA-binding domain, overlaying the C α s of residues 60 to 146 and 296 to 321 of the CBD NH₂-subdomain of corepressor-bound PurR, onto the corresponding C α s of each of the NH₂-subdomains of unliganded CBD suggests the mechanism by which corepressor binding to the CBD modulates DNA binding by PurR. Simply, in the ligand-free conformation, the two-fold related DNA-binding domains of PurR, which include the HTH (residues 4 to 23) and the hinge helix (residues 48 to 56), are relocated and the hinge helix regions are pulled apart (Fig. 5). As a measure of the magnitude of this motion, the distance between the C α of residue Lys60 to the C α of residue Lys60' (the first residue observed in each CBD monomer subunit), increases from

17.3 Å in the corepressor bound form to 20.8 Å in unliganded CBD. A likely consequence of the relocation of the hinge helices would be their destabilization due to the removal of the DNA-induced van der Waals interactions between them. Proteolysis studies have revealed that the hinge helix region is highly susceptible to cleavage when PurR is not making specific DNA contacts, a strong indication that DNA binding is a requisite for the stable formation of the hinge helices (27). Furthermore, recent NMR studies on residues 1 to 56 and 1 to 62 of PurR, in the absence of DNA, have demonstrated that residues 48 through 56 take no discernable structure (28). These data suggest a reaction coordinate whereby corepressor binding induces the necessary structural and conformational changes to bring the hinge helix regions of each monomer subunit into apposition. Local folding of the hinge helices can then take place in the presence of a specific DNA site. The probable role of the HTH motif, which together with helix 3 forms a stable globular domain (5), is to initiate specific DNA binding by making appropriate operator site base pair contacts.

The sequence homology between PurR and other LacI family members and their DNA operator sites strongly suggests that they take similar structures and use related DNA-binding mechanisms (4). However, unlike PurR, most LacI members are activated to bind their DNA sites only in the absence of effectors, a role indicative of the fact that they repress catabolic pathways and not biosynthetic pathways as does PurR. This necessitates a different spatial arrangement of the NH₂-subdomains in these proteins such that the hinge helices and HTH motif are positioned optimally to bind DNA in their unliganded forms. Alternatively, these proteins might display the same open and close conformations observed for PurR but link differently their DNA binding and effector binding/oligomerization domains such that in their open forms specific DNA binding is effected. The confirmation or modification of these hypotheses awaits the structures of other LacI family members bound to DNA (29).

In summary, the crystal structure of the corepressor-free form of the CBD of PurR reveals an open conformation relative to its corepressor bound form. This conformational change is effected by hinge bending rotations of 17° and 23° between the NH₂- and COOH-subdomains with the COOH-subdomains acting as an anchor in this rotation. Residues Tyr73, Trp147 and Asp160 are critical to this process. In the presence of corepressor these rotations serve to position the DNA-binding domains of each PurR monomer subunit so that hinge helix formation can occur in the presence of specific DNA. We expect an analogous conformational change for other members of the LacI family.

References and Notes

1. H. Zalkin and J.E. Dixon, in *Progress in Nucleic Acid Research and Molecular Biology*, (Academic Press, Inc., 1992) vol. 42, pp. 259-287.
2. K.Y. Choi and H. Zalkin, *J. Bacteriol.* **174**, 6207 (1992).
3. J.D. Gralla, in *Transcriptional Regulation 2*, S.L. McKnight and K.R. Yamamoto, Eds. (Cold Spring Harbor Laboratory Press, 1992), 629.
4. M.J. Weickert and S. Adhya, *J. Biol. Chem.* **267**, 15869 (1992).
5. M.A. Schumacher, K.Y. Choi, H. Zalkin and R.G. Brennan, *Science* **266**, 763 (1994).
6. C.A. Mauzy and M.A. Hermodson, *Protein Sci.* **I**, 843 (1992); J.U. Bowie, R. Lüthy, D. Eisenberg, *Science* **253**, 164 (1991); N.B. Vartak *et al.*, *Res. Microbiol.* **142**, 951 (1991).
7. L. Pearl, B. O'Hara, R. Drew and S. Wilson, *EMBO J.* **13**, 5810 (1994).
8. B.F. Anderson, H.M. Baker, G.E. Norris, S.V. Rumball and E.N. Baker, *Nature* **344**, 784.
9. A.J. Sharff, L.E. Rodseth, J.C. Spurlino, F.A. Quijoch, *Biochem.* **31**, 10657 (1992); G.A. Olah, S. Trakhanov, J. Trehwella, F.A. Quijoch, *J. Biol. Chem.* **268**, 16241 (1993); B-H. Oh *et al.*, *J. Biol. Chem.* **268**, 11348 (1993).
10. K.Y. Choi and H. Zalkin, *J. Bacteriol.* **174**, 6207 (1992).
11. The terms corepressor-free, unliganded and ligand-free CBD are used interchangeably as are the terms corepressor-bound, liganded and ligand-bound CBD.
12. R.J. Rolfes and H. Zalkin, *J. Bacteriol.* **172**, 5637 (1990).
13. P.M.D. Fitzgerald, *J. Appl. Crystallogr.* **21**, 273 (1988).
14. D.E. Tronrud, L.F. Ten Eyck and B.W. Matthews, *Acta crystallogr.* **A43**, 489 (1987).
15. The slightly elevated rmsd for the NH₂-subdomains residues 60-146 (rmsd of 0.84 Å and 0.91Å) results from structural differences in the NH₂-terminus of helix I and the NH₂-subdomain dimerization region. The slightly elevated rmsd for superimposition of residues 60-146 and 296-321 (text), results from the above noted differences as well as differences in positions of these noncontiguous regions between unliganded and liganded CBD. The largest differences (described in text) are located in regions 147-159.
16. J.S. Sacks, M.A. Saper and F.A. Quijoch, *J. Mol. Biol.* **206**, 171 (1989).
17. C.L. Careaga, J. Sutherland, J. Sabeti, and J.J. Falke, *Biochemistry* **34**, 3048 (1995).

18. F.A. Quiococho, *Curr. Opin. Struct. Biol.* **1**, 922 (1991).
19. L.Søgaard-Andersen, H. Pedersen, B. Holst, and P. Valentin-Hansen, *Mol. Microbiol.* **5**, 969 (1991).
20. L. Søgaard-Andersen, J. Martinussen, N.E. Møllegaard, S.R. Douthwaite, and P. Valentin-Hansen, *J. Bacteriol.* **172**, 5706 (1990).
21. L. Søgaard-Andersen, N.E. Møllegaard, S.R. Douthwaite, and P. Valentin-Hansen, *Mol. Microbiol.* **4**, 1595 (1990).
22. L. Søgaard-Andersen and P. Valentin-Hansen, *Mol. Gen. Genet.* **231**, 76 (1991).
23. K. Hammer-Jespersen, in *Metabolism of Nucleotides, Nucleosides and Nucleobases in Microorganisms.*, (Academic Press, 1983), 203.
24. C.J. Hueck, and W. Hillen, personal communication.
25. Z. Otwinowski, R.W. Schevitz, R-G Zhang, C.L. Lawson, A. Joachimiak, R.Q. Marmostein, B.F. Luisi and P.B. Sigler, *Nature* **334**, 321 (1988).
26. J.B.Rafferty, W.S. Somers, I. Saint-Girons and S.E.V. Phillips, *Nature* **341**, 705 (1989); W.S. Somers and S.E. Phillips, *Nature* **359**, 387 (1992); K. Phillips and S.E.V. Phillips, *Structure* **15**, 309 (1994).
27. K.Y. Choi, and H. Zalkin, *J. Bacteriol.* **176**, 1767 (1994).
28. Personal communication, Nagadoi, A., Morikawa, S., Nakamura, H., Enari, M., Kobayashi, K., Yamamoto, H., Sampei, G., Mizobuchi, K., and Nishimura, Y.
29. T.A. Steitz, R.E. Stenkamp, N. Geisler, K. Weber and J. Finch, *Biomol. Struct.Comform. Funct.& Evol.* **1**, 425 (1980); H.C. Pace, P. Lu and M. Lewis, *Proc.Natl. Acad. Sci. U.S.A.* **87**, 1870 (1990).
30. M.A. Schumacher, K.Y. Choi, H. Zalkin, and R.G. Brennan, *J. Mol. Biol.* **225**, 1131 (1992).
31. T.A. Jones, *Methods Enzymol.* **115**, part B, 157.
32. P.J. Kraulis, *J. Appl. Cryst.* **24**, 946 (1991).
33. Supported by March of Dimes grant no. 505 and by Public Health Service grant GM 49244 to R.G.B., by a National Defense Science and Engineering Graduate Fellowship in Biosciences to M.A.S. and by Public Health Service grant GM 24658 to H.Z.. The coordinates are in the process of being deposited in the Brookhaven Protein Data Bank but are available from R.G.B.. E-mail address, brennan@ohsu.edu.

TABLE 1: Summary of Selected Crystallographic Data

| | |
|--------------------------------|---|
| Space group | P2 ₁ |
| Cell Dimensions | a=38.04 Å, b=125.26 Å, c=61.29 Å, β=100.17° |
| Data Collection | |
| Resolution (Å) | 2.2 |
| Number of Reflections | 25,876 |
| R _{SYM} (%) | 3.6 |
| Number of atoms | 4,436 |
| Number of solvent molecules | 243 |
| Refinement | |
| R-factor (%) | 15.6 |
| Deviations from Ideality (rms) | |
| Bond distances (Å) | 0.018 |
| Bond angles (degrees) | 2.799 |

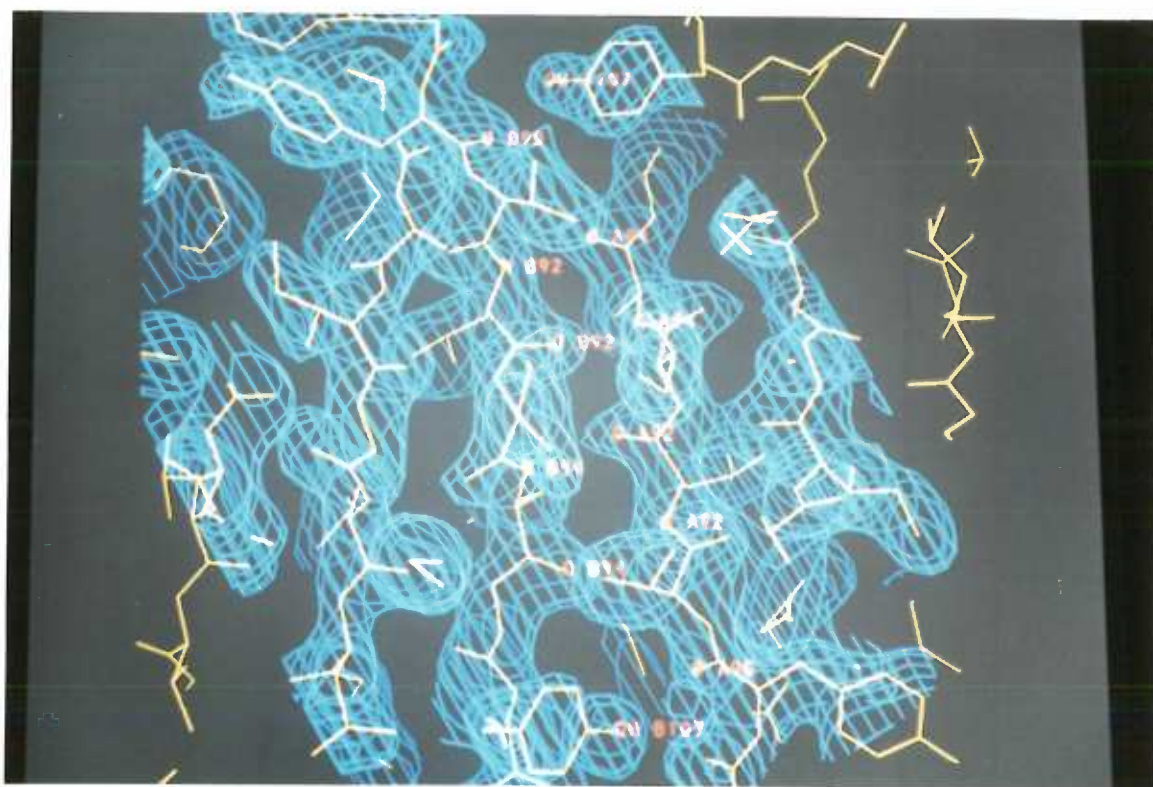
TABLE 1 LEGEND:

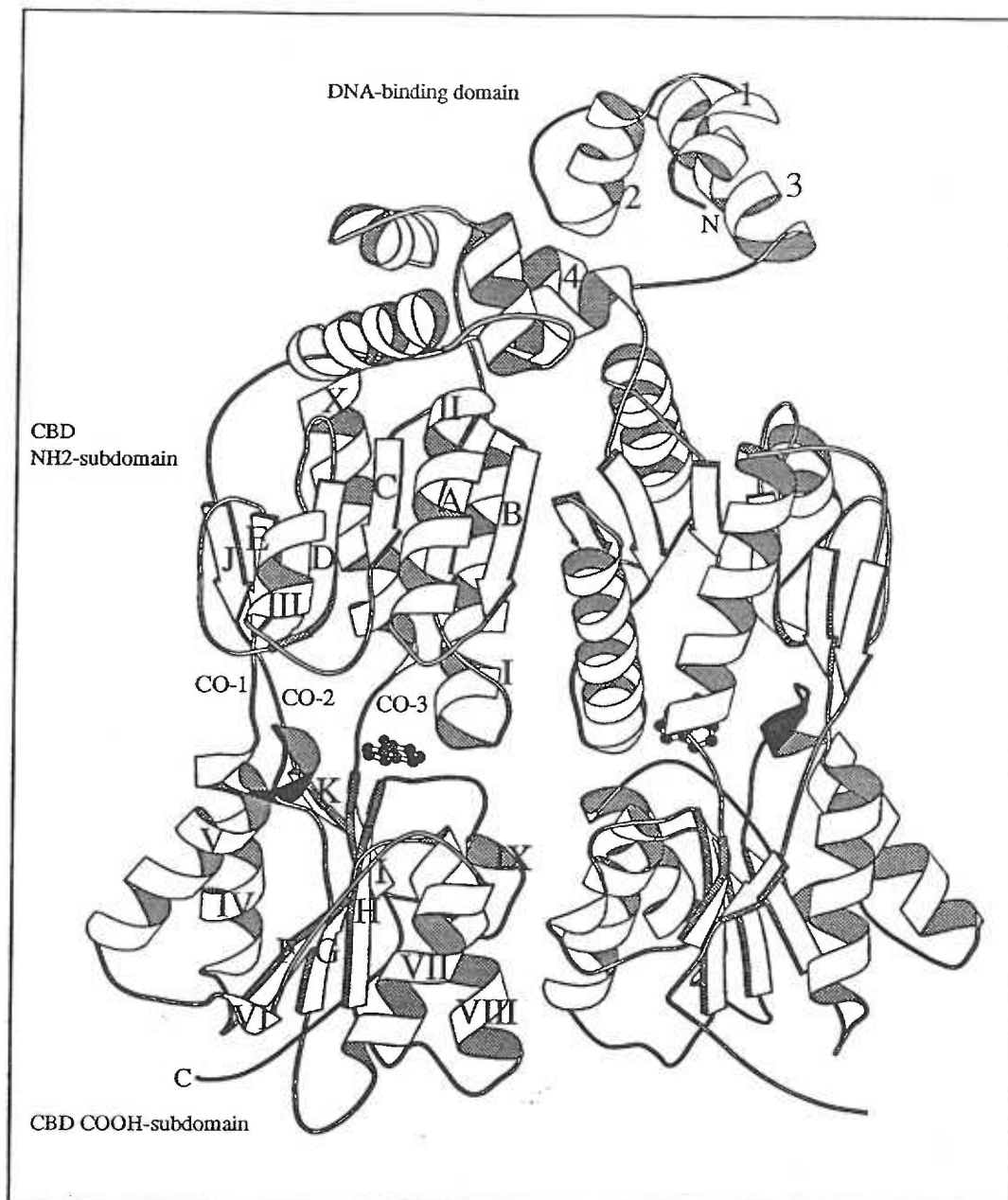
The corepressor binding domain (CBD) of PurR (residues 53-341) was prepared by trypsin cleavage of the intact purine repressor and was dialyzed exhaustively to remove any bound corepressor. Crystals of the CBD were grown in PEG 600, MgCl₂ solutions and were characterized by precession photography as previously described (30). X-ray intensity data were collected at room temperature with an Area Detector Systems Corporation (ADSC) multiwire area detector using a Rigaku RU200-H rotating anode generator for the X-ray source (40kV, 150 mA). The data were processed with the software provided by ADSC. The structure was solved by Molecular Replacement with MERLOT (13). Initial values for the rotation function were obtained using the Crowther fast rotation function and the resulting solutions were fine tuned with the Lattman rotation function. Subsequently, the Crowther and Blow translation function was used to position the search model. Search models were based on residues 60-340 of the PurR-hypoxanthine-*purF* complex structure. As corepressor loss was expected to result in conformational changes, models consisting of residues from the NH₂-subdomain dimer, the COOH-subdomain dimer, a single COOH-subdomain and a single NH₂-subdomain were used. All NH₂-subdomain models failed, but both the COOH-subdomain monomer as well as the COOH-subdomain dimer models produced clear solutions for both the rotation (9.0 and 8.6 standard deviations above the mean for the monomers and 11.5 standard deviations above the mean for the dimer) and translation (6.0 and 5.8 standard deviations above the mean and 8.5 standard deviations above the mean) functions. The euler angles solutions using the COOH-subdomain monomer, $\alpha=317.5^\circ$, $\beta=140.0^\circ$, $\gamma=105.0^\circ$ and $\alpha=137.5^\circ$, $\beta=40.0^\circ$, $\gamma=250.0^\circ$, were related by the strong self-rotation peak previously identified at $\phi=33.9^\circ$, $\varphi=78.9^\circ$ and $\kappa=180^\circ$ confirming that there was a CBD dimer in the asymmetric unit. A $2F_o - F_c$ electron density map calculated with phases from this partial model was of sufficient quality to visualize and fit a substantial part of the structure of one NH₂-subdomain. After rigid body refinement, a new electron density map was calculated that allowed for the fitting of part of the other NH₂-subdomain. The R-factor for this model was 39.9% after rigid body refinement. Positional least-squares refinement with TNT (14), using data from 10.0 Å to 3 Å, dropped the R-factor to 25.5%. This was followed by successive cycles of model rebuilding, using FRODO (31), and refinement. After fitting both the NH₂-subdomains, excluding residues 53 to 59 and 188 to 193, the resolution was extended to 2.2 Å and positional and thermal parameter refinement was carried out until convergence. Difference density maps calculated with coefficients, $F_{\text{HgCl}} - F_{\text{Nat}}$ and $F_{\text{Thimerosal}} - F_{\text{Nat}}$, and α_{Nat} , where F_{HgCl} were

the structure factors of a mercuric chloride derivative and $F_{\text{Thimerosal}}$ were the structure factors of a thimerosal derivative, revealed large two-fold related peaks coincident with locations of Cys85 for the thimerosal data and Cys85 and Cys123 for the mercuric chloride derivative data, confirming the different locations of the two NH_2 -subdomains. The final model consists of residues 60-187 and 194-341 of each CBD monomer with residues 53 to 59 and 188-193 being disordered, 243 water molecules and 3 magnesium ions.

Figure 1. The refined $2F_o - F_c$ electron density map contoured at 1.0σ showing the intersubunit β -sheet created between the NH_2 -subdomains in unliganded CBD. The β -sheet is formed by hydrogen bonds between the carbonyl oxygens of Val92 and Val92' (where the prime indicates the other subunit) and the amide groups of Gly94' and Gly94, respectively, and from the amide groups of Val92 and Val92' to the carbonyl oxygens of Gly94' and Gly94, respectively. Hydrogen bonds between the side chain hydroxyls of Tyr107 and Tyr107' to the carbonyl oxygens of Tyr90' and Tyr90, brace the sheet at both ends.

Figure 2. Ribbon diagrams of (A) The corepressor-bound form of PurR from PurR-hypoxanthine-*purF* operator complex. Labelled are the DNA-binding domain, the NH₂- and COOH subdomains and the NH₂-terminal residue (N) and the COOH-terminal residue (C). The three crossover regions are labelled CO-1, CO-2 and CO-3. The specificity region (residues 188-193) is shown in bold. (B) Corepressor-free form of the CBD of PurR. The NH₂-subdomain, the COOH-subdomain and the NH₂-terminal residue (N) and the COOH-terminal residue are labeled. The crossover regions are labelled as in (A). The specificity region, which is disordered in unliganded CBD, (residues 188 to 193) is shown in bold. Unliganded CBD NH₂-subdomain consists of (residues enclosed in parentheses) strand A (61 to 66), α -helix I (72 to 88), β -strand B (91 to 96), α -helix II (101 to 113), β -strand C (118 to 122), α -helix III (128 to 136), β -strand D (142 to 146), β -strand E (156 to 160), α -helix IX (296 to 310) and β -strand J (318 to 321). The unliganded COOH-subdomain consists of α -helix IV (162 to 175), β -strand F (182 to 184), α -helix V (194 to 206), 3_{10} -helix VI (213 to 215), α -helix VII (223 to 234), β -strand G (242 to 245), α -helix VIII (248 to 260), β -strand H (270 to 275), 3_{10} -helix IX (280 to 282), β -strand I (288 to 291) and β -strand K (324 to 326). These figures were generated using MOLSCRIPT (32).





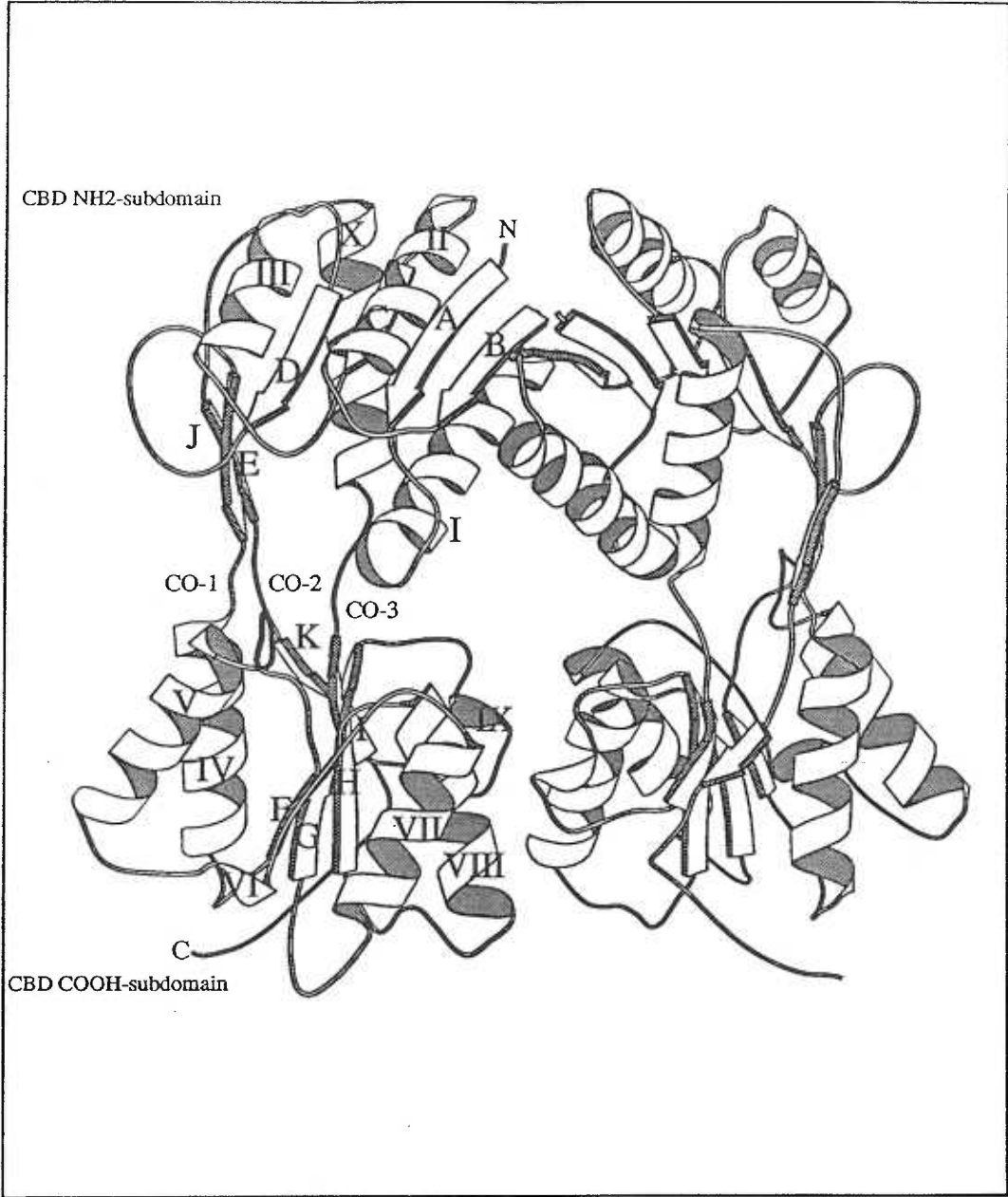
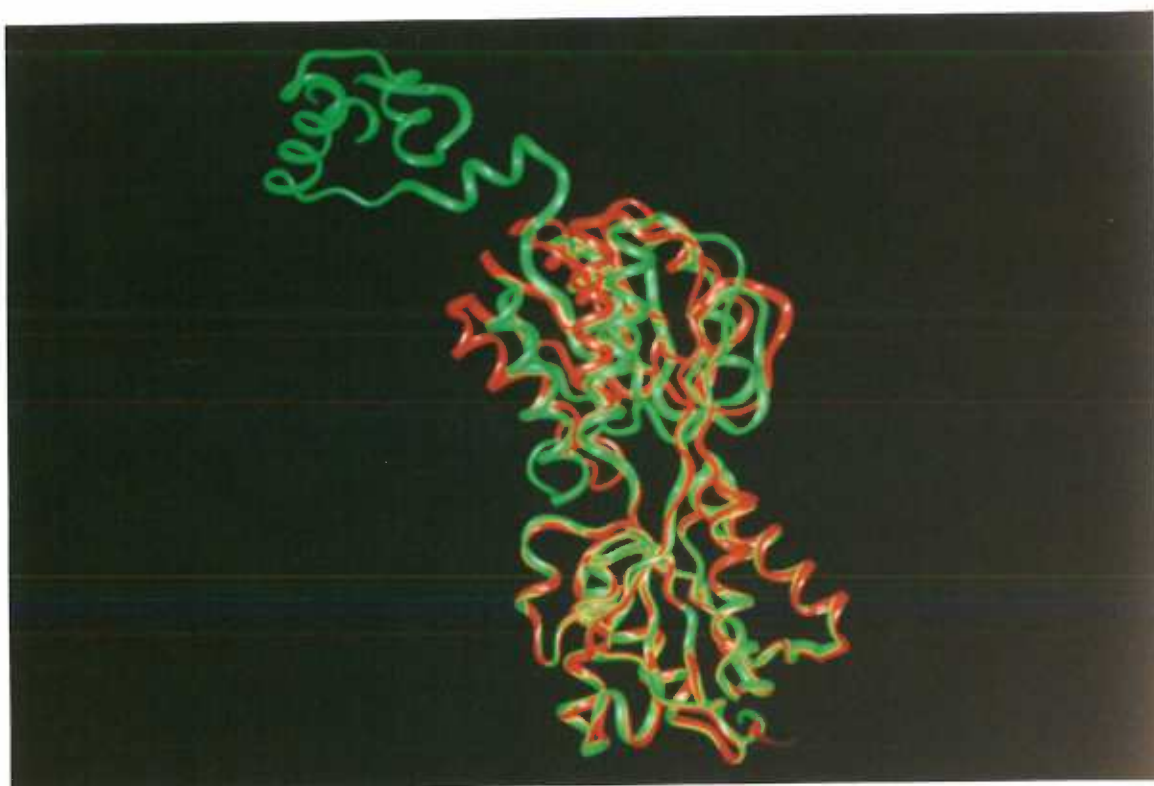


Figure 3: C α overlays of the unliganded CBD COOH-subdomains onto the liganded CBD COOH-subdomains (5). Unliganded CBD is represented as a solid red ribbon and liganded CBD as a solid green ribbon. (A) C α overlays of the crystallographically independent COOH-subdomains of unliganded CBD onto the COOH-subdomain from the crystallographic monomer of liganded CBD. The resulting rmsd for 126 atoms (residues 162 to 187, 194 to 290 and 324 to 326) is 0.34 Å for one CBD COOH monomer subunit and 0.40 Å for the other. Superimposing the two unliganded CBD COOH-subdomains results in an rmsd of 0.31 Å. As is evident from the locations of the CBD NH₂-subdomains, unliganded CBD is in an open conformation relative to the liganded form. Rotations of 17° and 23° are required to bring residues 60 to 146 and 296 to 321 of the two NH₂-subdomains into coincidence with the corresponding liganded NH₂-subdomains. Residues 60 to 146 and 296 to 321 of the two crystallographically independent NH₂-subdomains superimpose with rmsd of 0.62 Å while C α superimposition these residues between each unliganded CBD NH₂-subdomain and the liganded form results in rmsds of 1.2 Å and 1.3 Å. Superimposition of the NH₂-subdomain dimer of unliganded CBD (residues 60 to 146 and 296 to 321 of each subunit) onto that of liganded CBD results in an rmsd of 3.49 Å, underscoring the large conformational change of the NH₂-subdomain of one monomer relative to the other, in going from the ligand-free to the ligand-bound form. (B) Superimposition of the COOH-subdomain dimers of the two forms results in a root mean squared deviation (rmsd) of 0.60 Å (for 252 corresponding atoms) indicating that the COOH-subdomain dimers do not differ structurally in the presence or absence of corepressor and providing evidence that the COOH-subdomain acts as a structural anchor for PurR. Figure generated with BIOSYM Insight II.



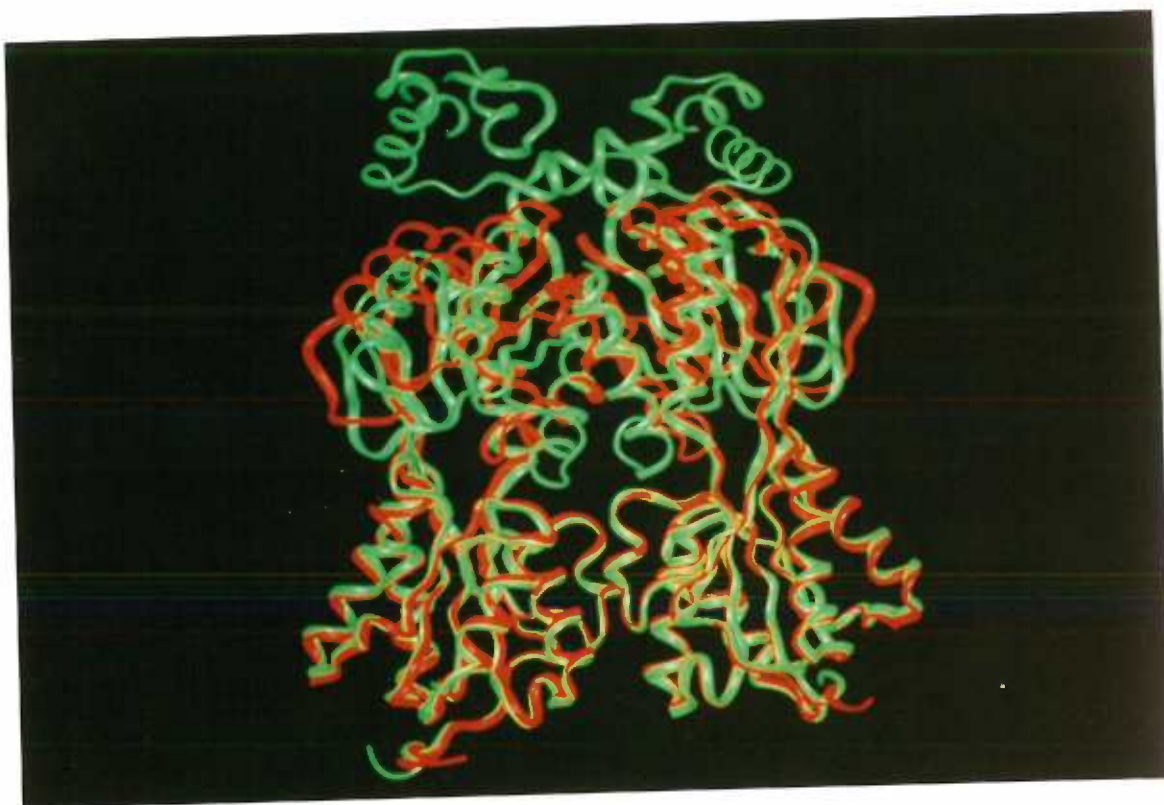
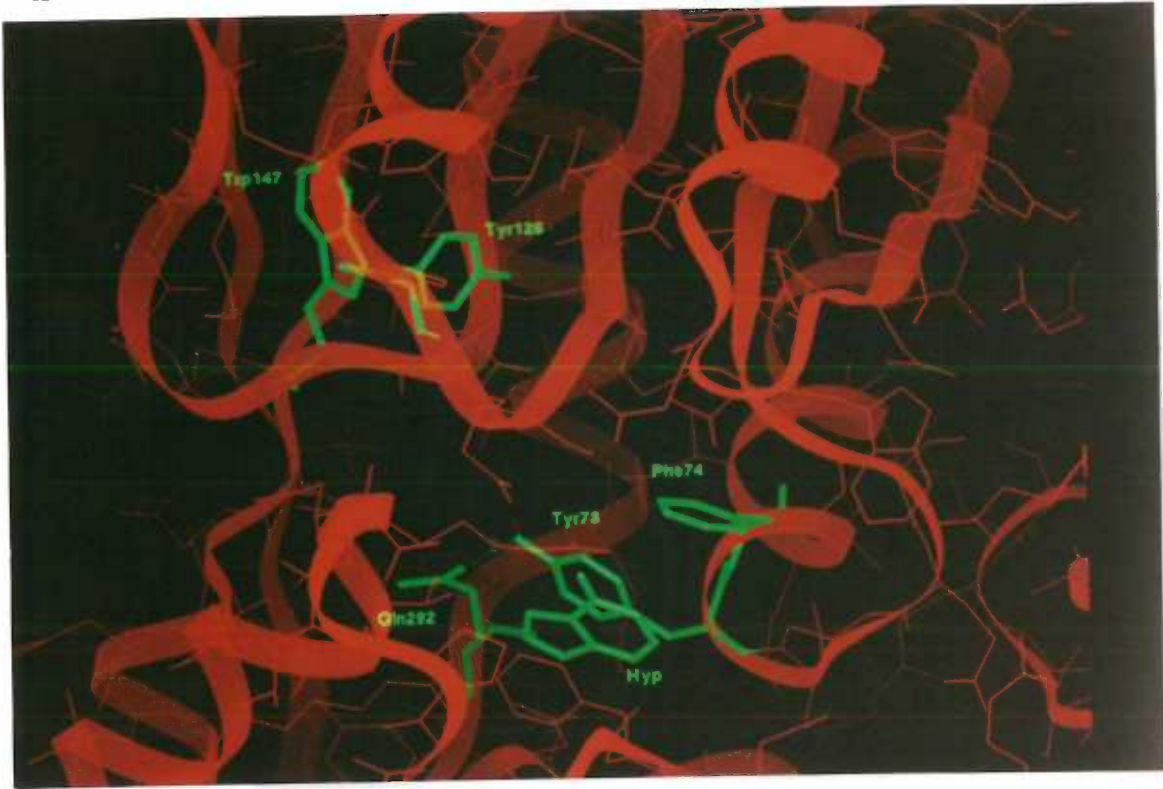


Figure 4: The switch residues of the CBD NH₂-subdomains in corepressor-free and corepressor-bound CBD.

(A) In liganded CBD, Tyr73 and Phe74, the second and third residues of helix I, interact with the hypoxanthine corepressor. The side chain of Tyr73 also hydrogen bonds to Gln292, located on the second crossover between subunits. Trp147 is far removed from the corepressor binding pocket and stacks against Tyr126.

(B) In unliganded CBD, the N-terminus of helix I becomes distorted thereby allowing Tyr73 and Phe74 to interact with key switch residue, Trp147. Residue Trp147, which is located on the loop between β -strands D and E, has rotated into the corepressor binding pocket. In addition to hydrogen bonding to the N ϵ of Trp147, the side chain OH of Tyr73 also hydrogen bonds to Asp160, located on the first crossover. Figure generated with BIOSYM Insight II.

A



B

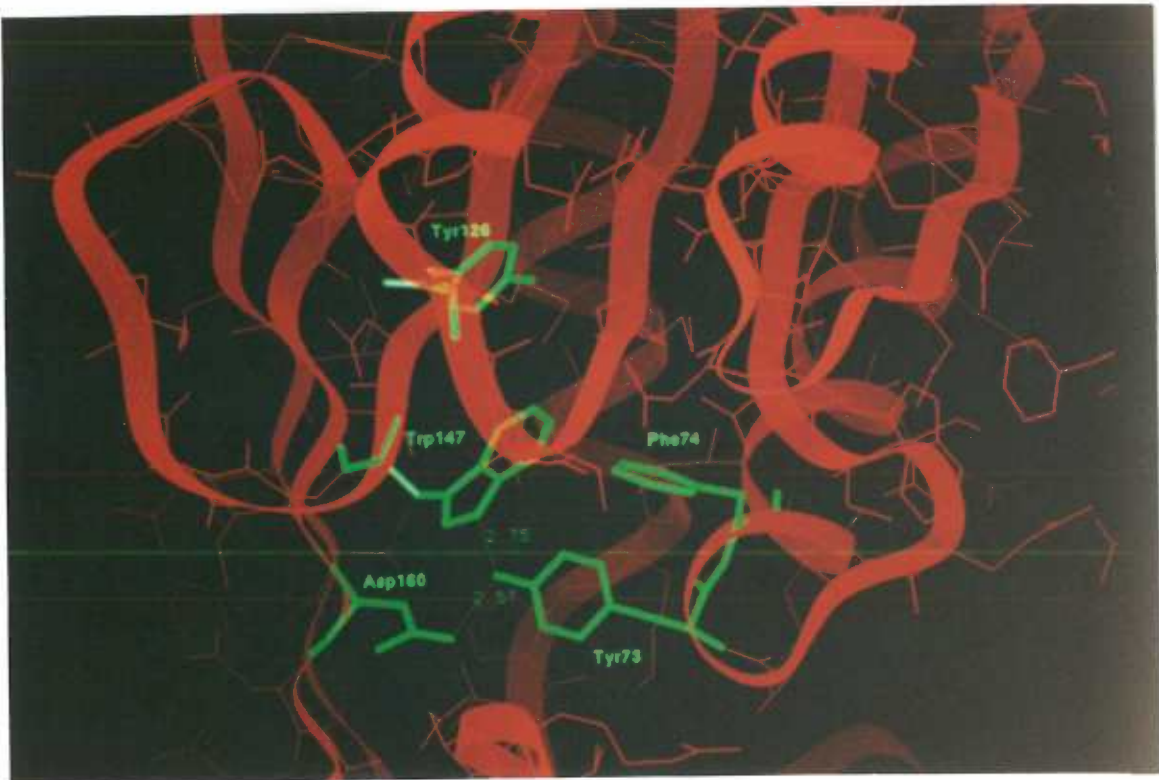
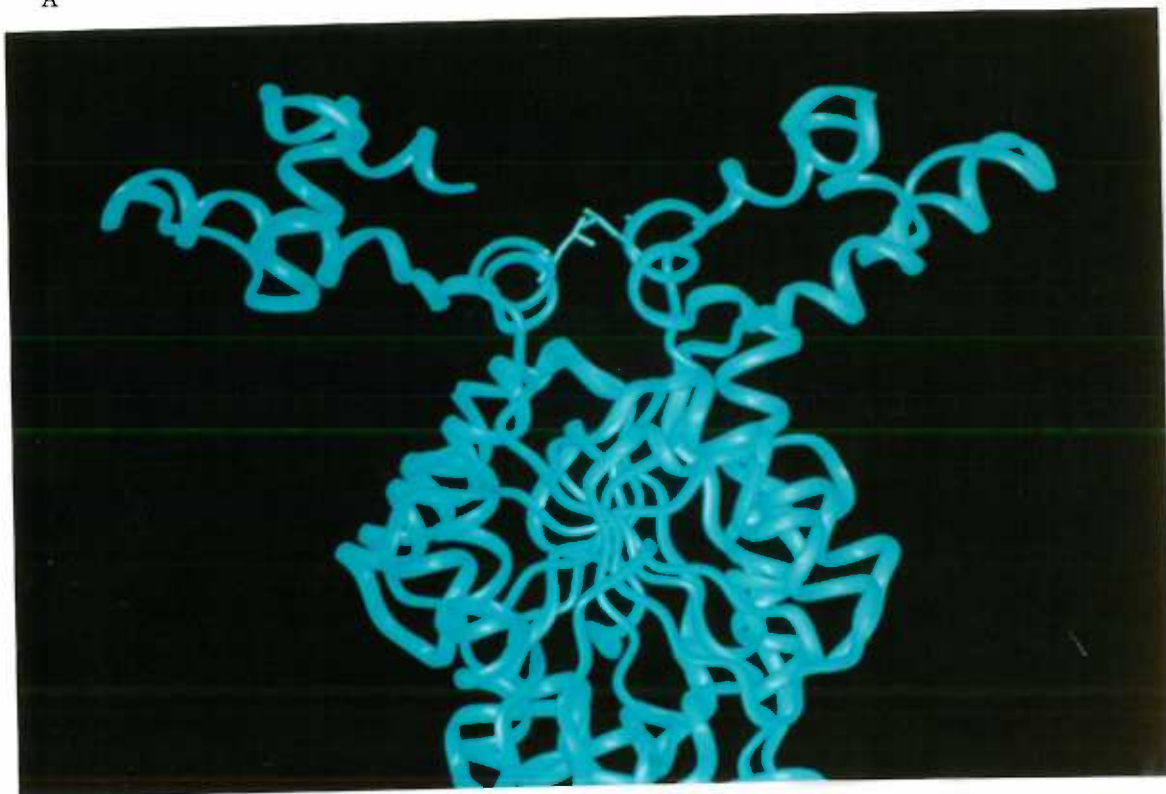
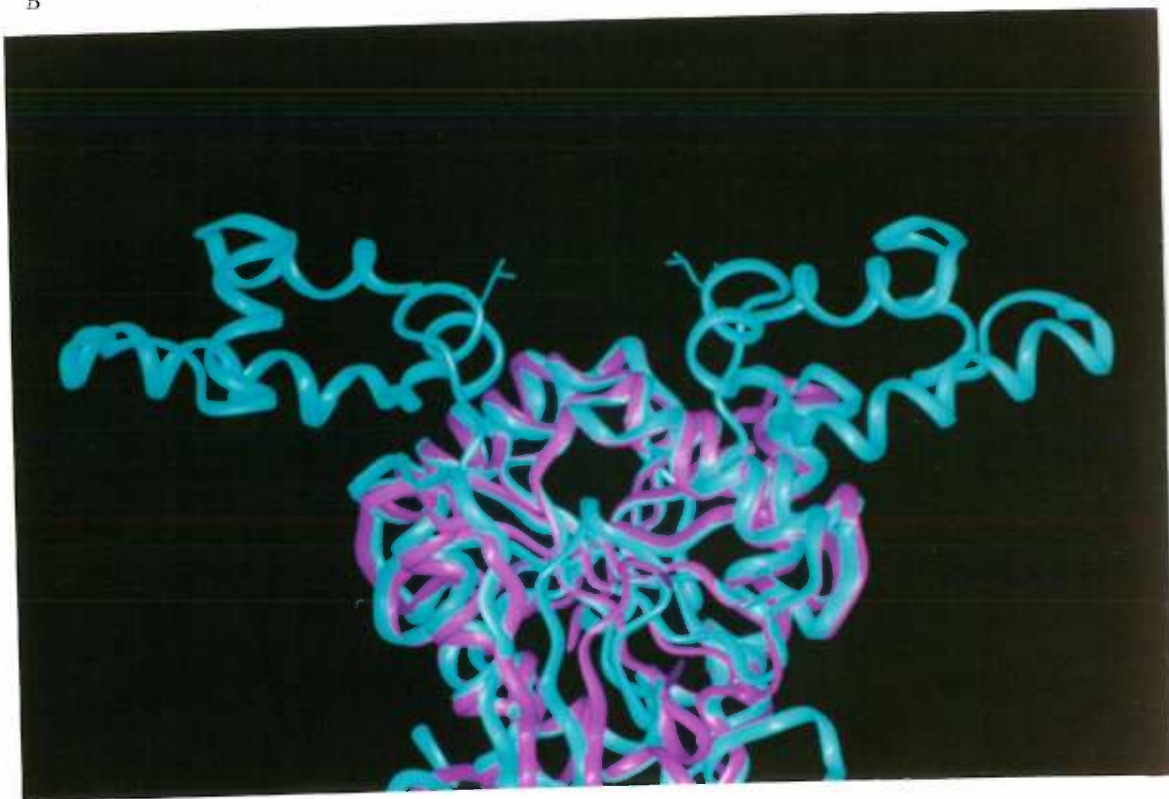


Figure 5: Implications for DNA binding. (A) View of the hinge helix interactions in the PurR-hypoxanthine-DNA complex. The leucine "lever" side chains (Leu54 and Leu54') are also shown. (B) Overlay of the NH₂-subdomains (C α residues 60-146 and 296-315) from the PurR-hypoxanthine-*purF* structure, including the DNA binding domains (shown in blue), onto the corresponding C α s of the NH₂-subdomains of unliganded CBD (magenta). The overlay demonstrates dramatically that the hinge helices would be broken apart, disrupting the weak interhelical interactions that act to stabilize the helices and ultimately disrupting the DNA binding domains. Figure generated with BIOSYM Insight II.

A



B



DISCUSSION AND CONCLUSION

The work presented in this thesis examines, at the atomic level, the basis of DNA binding specificity by the regulatory protein, the Purine repressor or PurR, and the mechanism by which a metabolic signal in the form of a purine corepressor is transmitted to effect gene regulation by activating PurR to bind DNA. Prior to the work of this thesis it had been established that PurR was a 341 amino acid, dimeric DNA binding protein that functioned as the master regulator of *de novo* purine biosynthesis and, to a lesser extent, *de novo* pyrimidine biosynthesis in *E. coli* (1-17). The importance of PurR in *E. coli* gene regulation is underscored by recent studies which have demonstrated that PurR also regulates the transcription of several genes involved in nucleotide metabolism (18). PurR carries out its control by a simple negative feedback loop, whereby it senses the level of nucleotides in the *E. coli* cell via the concentrations of its corepressors, hypoxanthine and guanine, which it binds with Kds of 9.3 μM and 1.5 μM , respectively (9). Corepressor binding activates PurR to bind to its cognate DNA sites within the *pur* regulon, leading to repression of the transcription of these genes.

PurR is a member of the highly homologous LacI family of bacterial transcription regulators (19). All LacI family members, including PurR, can be proteolytically digested into two domains: An NH₂-terminal subdomain, containing the DNA binding function, and a larger COOH-terminal subdomain, which is responsible for the corepressor binding/dimerization functions. Prior to this thesis, the only three-dimensional structural information for the LacI proteins had been obtained from NMR studies on the LacI DNA binding domain (Lac headpiece), both free and bound to a *lac* operator half site. Those studies confirmed the presence of a helix-turn-helix motif in this domain (20). However, these studies left in question the structure of the effector binding/dimerization domain. Furthermore, these studies could not address the key question of how the signal of effector binding is transduced to the DNA binding domain to effect gene regulation.

The high degree of sequence homology observed within the LacI family, as well as the operator sites they recognize, suggests that all members are structurally homologous and interact with their DNA sites in a very similar manner. One goal of this thesis was to provide a model, or prototype, for the, over 21, LacI family members through the structure determination of PurR. The specific questions which were addressed were 1) what is the basis for the sequence specific DNA binding of PurR and 2) what is the means by which corepressor binding to the corepressor binding domain activates the DNA binding domain of PurR to bind DNA. In great part, these questions were answered by the determination of two structures by X-ray crystallography. The first structure determined was the ternary

complex of PurR bound to both its corepressor, hypoxanthine, and a 16 base pair *purF* operator site. The second was the corepressor free form of the corepressor binding domain of PurR (CBD). The structure of the ternary complex was solved by multiple isomorphous replacement to 2.7 Å resolution and refined to an R-factor of 19.5%. As anticipated, the structure of PurR contains two domains: An NH₂-terminal DNA binding domain and a COOH-terminal corepressor binding/dimerization domain. The DNA binding domain could be divided into two functionally important regions. The first region contains a helix-turn-helix motif that is part of a larger globular, helix-turn-helix-loop-helix, domain (residues 3 to 43) and similar to the NMR structure of the LacI headpiece (residues 1 -51) (20). Key specificity determining contacts are made by residues from helix two, the recognition helix. Hydrogen bonds are provided from residue Thr16 of this helix to both the adenine and thymine of base pair 6. Similar to the λ repressor-DNA structure, a residue outside the recognition helix also makes a base specific contact. In this contact Arg26, which is located on the loop following helix two, hydrogen bonds to Gua4. This interaction is critical in terms of specificity and explains why guanine4 is the only guanine in the *purF* operator to show DNA methylation interference sensitivity and why it is conserved in all known *pur* operators.

Following the helix-turn-helix-loop-helix is the second DNA binding element. This element represents a novel DNA binding structural element and consists of two-fold related "hinge" helices that bind deeply in the minor groove of the DNA. The hinge helices interact with each other through a series of weak van der Waals interactions between the two fold-related residues Val50 and Val50' and Leu54 and Leu54'. The accommodation of two helices into a B-DNA minor groove clearly requires its expansion. This is effected by the insertion of the two Leu54s between the central CpG base pair step. This insertion unstacks the central base pairs and kinks the DNA by 45°. As a result, the central base pair step displays an unusually large roll angle of 45° and helical rise of 6.4 Å. Additional DNA contacts are provided by "hinge" helix residues including phosphate contacts from Arg52 and Ser48. Lys55, which hydrogen bonds to Adenine 8, contributes to operator site specificity in that it discriminates against a guanine at position 8 due to the steric clash that would result between this large side chain and the exocyclic N2 of guanine. This is corroborated by the finding that position 8, although it can be a cytosine or thymine, is never a guanine in all known *pur* operators.

Biochemical studies have shown that many DNA binding proteins bend their DNA sites. However, DNA kinking, which is distinguished from DNA bending by the presence of unstacked base pairs, has been directly observed in only three transcription factor-DNA complex structures: The CAP-cAMP-DNA structure (21), the TATA binding protein-DNA

structures (22-23) and now, the PurR-hypoxanthine-DNA structure. CAP kinks its DNA at two dyad-related TpG steps, which results in an overall bending angle of $\sim 90^\circ$ (21). This kinking is elicited by contacts made by CAP to the major groove of the DNA. TBP kinks the TATA element at the ends of an eight base pair TATA element. Unlike CAP, TBP stabilizes its two $\sim 45^\circ$ kinks through minor groove interactions, primarily the intercalation of a pair of phenylalanines between the each of the two base pair steps (22-23). Although the mechanism of kinking by PurR is unique, it is similar to CAP in that α helices are used to induce kinking, but like TBP and unlike CAP, the interactions are with the DNA minor groove. Although minor groove binding by α helices had not previously been observed, NMR studies on the eukaryotic protein, SRY (sex determining region on chromosome Y), bound to a high affinity DNA site, suggested that this HMG protein interacts with the minor groove in a manner similar to PurR whereby an isoleucine side chain, located on a purported α helix, appears to intercalate into the minor groove (24). Thus, PurR may serve as a model for DNA binding by SRY.

The role of DNA kinking and bending in gene expression is not clear. However, one obvious result of DNA bending is that several DNA binding factors may be brought closer together than would be possible on linear DNA. For example, it has been postulated that DNA bending by CAP might facilitate contacts between RNA polymerase and the DNA upstream from CAP. This could explain how CAP can activate transcription from a variety of positions. DNA bending by TBP may likewise juxtapose the transcriptional initiation factors needed to form the pre-initiation complex. Unlike CAP and TBP, PurR functions as a repressor. What role DNA bending may play in the repression is unclear. However, it implies that the simple "roadblock" mechanism, previously postulated to explain repression, may require revision. Perhaps, in addition to steric effects, the significant perturbation of the DNA structure caused by PurR binding would prohibit RNA polymerase, and other initiation factors from properly "reading" or opening a promoter site.

The high degree of sequence homology displayed between PurR and other LacI members, especially in the DNA binding domain (60% identical), strongly suggests that all LacI members bind their DNA sites in a manner similar to PurR. Particularly noteworthy is the absolute conservation of the critical "hinge" helix residues, corresponding to Ala51 and Leu54 in PurR. Ala51 is an important residue in terms of stereochemical constraints as it directly abuts the DNA backbone, indicating that any other residue, except perhaps glycine, would disrupt DNA binding. Leu54 is the critical "leucine lever" which, along with its dyadic partner, is responsible for the severe kink in the central CpG base pair step of the *purF* operator.

The tight binding of PurR to its operator is contingent upon corepressor binding (9). Several studies have predicted that the corepressor binding domain (CBD) of PurR is structurally similar to the bacterial periplasmic binding proteins (PBPs) which function in the transport of metabolites across the bacterial membrane and, in some cases, chemotaxis. The structure of the PurR-hypoxanthine-*purF* operator ternary complex revealed that the CBD is, in fact, remarkably similar the PBPs, especially the ribose binding protein, RBP (25). Like the PBPs, PurR binds its corepressor in a cleft formed between the two structurally similar subdomains of the CBD through a series of polar, nonpolar and aromatic interactions. The structure clearly reveals the basis for corepressor binding specificity. It is the hydrogen bonds from Arg190 to the exocyclic O6 acceptor of hypoxanthine (also found in guanine) that allow PurR to read the exocyclic atom at position 6 and, consequently, discriminate against the amino group found at this position in adenine. Unlike the PBPs, however, PurR is a dimer. The PurR dimerization interface is extensive, excluding 2,242 Å² of protein surface, and is formed, for the most part, equally by both CBD subdomains.

Structural and biochemical studies on the periplasmic binding proteins have provided evidence that the PBPs exist in an equilibrium between an open, unliganded form and a closed, liganded form and that ligand binding acts to stabilize the closed form (26-27). The conformational change, mostly quaternary, that accompanies the transition from the open to the closed form allows the PBP receptor proteins to distinguish between the two forms. Although the structure of the PurR ternary complex cannot explain the means of corepressor binding activation, it suggested that, perhaps, PurR may exist in an open unliganded conformation similar to the PBPs and that corepressor binding may stabilize the closed form, transmitting the signal to the DNA binding domain.

To address this question, the structure of the unliganded form of the corepressor binding domain of PurR (residues 53 - 341) was solved by X-ray crystallography. The structure was determined by molecular replacement using the CBD COOH-subdomain as a probe model. The structure has been refined to an R-factor of 15.6% using data to 2.2 Å resolution. This structure reveals an open conformation as the two globular subdomains of each monomer subunit are rotated by 17° and 23° relative to that seen in the corepressor-bound form of the PurR-hypoxanthine-*purF* ternary complex. The CBD COOH-subdomains change little upon corepressor binding. However, the NH₂-subdomains rotate open. This leads to a large reorientation of these domains relative to each other. As a result, the six-stranded β-sheets of each NH₂-subdomain become linked and form a continuous twelve-stranded β-sheet between subdomains.

Aside from the significantly different disposition of the NH₂-subdomains relative to each other, the largest structural changes that can be noted when comparing corepressor-free to corepressor-bound CBD all reside within the corepressor binding pocket. In the CBD COOH-subdomains, the NH₂-terminus of helix V, which contains Thr192, and the loop containing Arg190, become completely disordered upon corepressor release. There are also large structural rearrangements that occur in the CBD NH₂-subdomains in which the β -sheet interface between strand D and strand C is disrupted and replaced by a new set of β -sheet hydrogen bonds. At the heart of this alteration is the rearrangement of the region from residues 147-159. In the corepressor-free form of the CBD, Trp147 swings into the corepressor binding pocket to interact with Phe74 and Tyr73 from helix I, whereas in the corepressor-bound form it swings out of the corepressor binding pocket and stacks against Tyr126 from helix III. In this way Trp147 serves as a structural, although not functional, replacement for the corepressor, stabilizing Phe74 and Tyr73 in the absence of corepressor. Thus, residues 147-159 move, as a separate unit, from residues 60-146 upon corepressor binding. This separation appears to be made possible by the formation of two different yet stable hydrogen bonding β -sheet interfaces between strands D and C. A hydrogen bond between Tyr73 and Asp160 in unliganded CBD is also critical to the stability of the unliganded form.

Although the structure of the unliganded CBD does not include the DNA-binding domain, overlaying the NH₂-subdomains (C α residues 60-146) of the corepressor-bound form of PurR onto the corresponding C α residues of the NH₂-subdomains of unliganded CBD suggests a mechanism for how the signal of corepressor binding acts to modulate DNA binding. As revealed in the PurR-hypoxanthine-*purF* operator crystal structure, the CBD of PurR is attached to the DNA-binding domain by a short linker (residues 57-59). The functional DNA-binding domain is made up of a helix-turn-helix motif (residues 4-23) that interacts with the major groove of the DNA and a hinge helix (residues 48-56) which interacts with the minor groove and is directly attached to the CBD via the short peptide (residues 57-59). Weak interactions between the two-fold hinge helices are critical, not only for minor groove binding, but also for stabilizing PurR's dimeric DNA binding domain. When PurR is in the "open" form, the hinge helices would be pulled apart. (the distance from the C α of residue 60 to that of residue 60' increases from 17.3 Å in the corepressor bound form to 20.8 Å in unliganded CBD). Disruption of the weak interactions between these helices would greatly destabilize them, which supports the hypothesis that these helices may be nascent and require the corepressor bound form of the CBD, as well as DNA interactions, to stabilize their formation. Proteolysis studies (28) and a recent structure determination of the PurR headpiece (residue 1-61) (29) provide

strong evidence for this hypothesis. The proteolysis studies demonstrated that residues 48-56 are susceptible to proteolytic cleavage when PurR is not making contacts to a specific DNA site (operator) and the NMR structure of the PurR headpiece reveals that, although the first three helices are present, as seen in the PurR ternary structure, the fourth or hinge helix is absent and that residues 48-56 take a random conformation. These data, combined with the lack of electron density for residues 53-59 in the CBD electron density map, support the notion that this region folds into stable helices only when activated and bound to DNA. This finding is consonant with recent data demonstrating the importance of local "folding" in DNA binding for transcription factors. Although most LacI members are specifically bound to DNA only in the absence of inducer (the exceptions being PurR and CcpA, the Amylase repressor), a similar mechanism likely occurs upon effector binding for all proteins. The different outcomes could be explained by the different relative arrangements of the NH₂-subdomains of a given effector binding domain such that the hinge regions would be closer in the unliganded form of some members, e.g., LacI, and the opposite would be true in others, e.g., PurR. A complete understanding of this difference, however, necessitates that structures be determined for LacI members of the former class.

Currently, the only examples of gene regulatory proteins effected by small molecule binding that have been well characterized structurally are the *E. coli trp* (30) and MetJ (31) repressors. However, corepressor activation of PurR can be contrasted to that of the *trp* and Met repressors in that the corepressors of PurR are bound at a much larger distance from the DNA binding domain (~40 Å). Structural studies on the *trp* repressor have shown that its corepressor, L-tryptophan, is bound within the DNA binding domain and functions to orient the flexible reading heads so that the helix-turn-helix motifs can interact with successive major grooves of the operator. Furthermore, the L-tryptophan directly contacts the DNA through a phosphate contact via its indole ring nitrogen. Conversely, studies have shown that no conformational change occurs in the Met repressor upon binding its corepressor, S-adenosylmethionine (SAM) (32). Moreover, the SAM molecules do not interact with the DNA, although they are bound very close to the DNA binding domain of MetJ. It is believed that SAM activation of MetJ is the result of electrostatic effects (33). This is supported by studies which demonstrate that the SAM analogue, S-adenosylhomocysteine, although bound by MetJ, does not increase its affinity for its operator (31).

In conclusion, the structure of the PurR-hypoxanthine-*purF* operator complex and the corepressor free form of the corepressor binding domain of PurR have revealed: 1) the mechanism by which PurR and, likely, other LacI members specifically bind to their

cognate DNA sites; 2) a novel DNA binding motif in which two symmetry related α helices specifically interact with DNA minor groove and which may serve as a model for DNA binding by SRY; 3) a mechanism of small molecule signalling in which the small molecule (corepressor) is bound at a site distant from that which elicits the response.

REFERENCES

1. Zalkin, H. and Nygaard, P. (1994). Bioynthesis of Purine Nucleotides. In press
2. Zalkin, H. and Dixon, J.E. (1992). *De novo* Purine Nucleotide Biosynthesis
Prog.Nucl. Acid Research **42**, 259-286.
3. Gots, J.S., Benson, C.E., Jochimsen, B., and Koduri, K.R. (1977). Microbial models and regulatory elements in the control of purine metabolism. *Ciba Found. Symp.* **48**, 23-41.
4. Kilstrup, M., Meng, L.M., Neuhard, J., and Nygaard, P. (1989). Genetic Evidence for a Repressor of Synthesis of Cytosine Deaminase and Purine Biosynthesis Enzymes in *Escherichia coli*. *J. Bacteriol.* **171**, 2124-2127.
5. Rolfes, R. and Zalkin, H. (1988). Regulation of *Escherichia coli purF*. *J. Biol. Chem.* **263**, 19649-19652.
6. Rolfes, R. and Zalkin, H. (1988). *Escherichia coli* Gene *purR* Encoding a Repressor for Purine Nucleotide Synthesis. *J. Biol. Chem.* **263**, 19653-19661.
7. Rolfes, R. and Zalkin, H. (1990). Purification of the *Escherichia coli* Purine Regulon Repressor and Identification of Corepressors. *J. Bacteriol.* **172**, 5637-5642.
8. Rolfes, R. and Zalkin, H. (1990). Autoregulation of *Escherichia coli purR* Requires Two Control Sites Downstream of the Promoter. *J. Bacteriol* **172**, 5758-5766.
9. Choi, K.Y. and Zalkin, H. (1992). Structural characterization and corepressor binding of the *Escherichia coli* purine repressor. *J. Bacteriol.* **174**, 6207-6214.
10. Choi, K.Y., Lu, F., and Zalkin, H. (1994). Mutagenesis of amino acid residues required for binding of corepressors to the purine repressor. *J. Biol. Chem.* **269**, 24066-24072.
11. He, B. and Zalkin, H. (1992). Repression of *Escherichia coli purB* is by a Transcriptional Roadblock Mechanism. *J.Bacteriol* **174**, 7121-7127.
12. He, B. and Zalkin, H. (1994). Regulation of *Escherichia coli purA* by Purine Repressor, One Component of a Dual Control Mechanism. *J.Bacteriol* **176**, 1009-1013.
13. He, B., Smith, J.M. and Zalkin, H. (1992). *Escherichia coli purB* Gene: Cloning, Nucleotide Sequence, and Regulation by *purR*. *J. Bacteriol.* **174**, 130-136.
14. Meng, L.M., Kilstrup, M., and Nygaard, P. (1990). Autoregulation of PurR repressor synthesis and involvement of *PurR* in the regulation of *purB*, *purC*, *purL*, *purMN* and *guaBA* expression in *Escherichia coli*. *Eur. J. Biochem.* **187**, 373-379.

15. Steiert, J.G, Rolfes, R., Zalkin, H., and Stauffer, G.V. (1990). Regulation of the *Escherichia coli glyA* Gene by the *purR* Gene Product. *J. Bacteriol.* **172**, 3799-3803.
16. Choi, K.Y. and Zalkin, H. (1990). Regulation of *Escherichia coli pyrC* by Purine Regulon Repressor Protein. *J. Bacteriol.* **172**, 3201-3207.
17. Wilson, H. and Turnbough, C.L. (1990). Role of the purine repressor in the regulation of pyrimidine gene expression in *Escherichia coli* K-12. *J. Bacteriol.* **172**, 3208-3213.
18. He, B., Choi, K.Y., and Zalkin, H. (1993). Regulation of *Escherichia coli glnB*, *prsA*, and *speA* by the Purine Repressor. *J. Bacteriol.* **175**, 3598-3606.
19. Weickert, M.J. and Adhya, S. (1992). A Family of Bacterial Regulators Homologous to Gal and Lac Repressors. *J. Biol. Chem.* **267**, 15869-15874.
20. Chuprina, V.P., Rullman, J.A.C., Lamerichs, R.M.J.N., van Boom, J.H., Boelens, R., and Kaptein, R. (1993). Structure of the Complex of *lac* Repressor Headpiece and an 11 Base-pair Half-operator Determined by Nuclear Magnetic Resonance Spectroscopy and Restrained Molecular Dynamics. *J. Mol. Biol.* **234**, 446-462.
21. Schultz, S.C., Shields, G.C., and Steitz, T.A. (1991). Crystal Structure of a CAP-DNA Complex: the DNA is Bent by 90°. *Science* **253**, 1001-1007.
22. Kim, Y., Geiger, J.H., Hahn, S., and Sigler, P.B. (1993). Crystal Structure of a Yeast TBP/TATA-box Complex. *Nature* **365**, 512-520.
23. Kim, J.L., Nikolov, D.B., and Burley, S.K. (1993). Co-Crystal Structure of TBP Recognizing the Minor Groove of a TATA Element. *Nature* **365**, 520-527.
24. King, C-Y and Weiss, M.A. (1993). The SRY High-Mobility -Group Box Recognizes DNA by partial Intercalation in the Minor Groove: A Topological Mechanism of Sequence Specificity. *Proc. Natl. Acad. Sci. USA* **90**, 11990-11994.
25. Mowbray, S.L. and Cole, L.B. (1992). 1.7 Å Structure of the Periplasmic Ribose Receptor from *Escherichia coli*. *J. Mol. Biol.* **225**, 155-175.
26. Sharff, A.J., Rodseth, L.E., Spurlino, J.C., and Quioco, F.A. (1992). Crystallographic Evidence of a Large Ligand-Induced Hinge-Twist Motion between the Two Domains of the Maltodextrin Binding Protein Involved in Active Transport and Chemotaxis. *Biochem.* **31**, 10657-10663.
27. Olah, G.A., Trakhanov, S., Trewella, J., and Quioco, F.A. (1993). Leucine/Isoleucine/Valine-binding Protein Contracts upon Binding of Ligand. *J. Biol. Chem.* **268**, 16241-16247.
28. Choi, K.Y. and Zalkin, H. (1994). Role of Purine Repressor Hinge Sequence in Repressor Function. *J. Bacteriol.* **176**, 1767-1772.

29. Personal communication, Nagadoi, A., Morikawa, S., Nakamura, H., Enari, M., Kobayashi, K., Yamamoto, H., Sampei, G., Mizobuchi, K., and Nishimura, Y.
30. Otwinowski, Z., Schevitz, R.W., Zhang, R.G., Lawson, C.L., Joachimiak, A., Marmorstein, R.Q., Luisi, B.F., and Sigler, P.B. (1988). Crystal Structure of *trp* Repressor/Operator Complex at Atomic Resolution. *Nature* **335**, 321-329.
31. Somers, W.S. and Phillips, S.E.V. (1992). Crystal structure of the *met* repressor-operator complex at 2.8 Å resolution reveals DNA recognition by β -strands. *Nature* **359**, 387-393.
32. Rafferty, J.B., Somers, W.S., Saint-Girons, I., and Phillips, S.E.V. (1989). Three-dimensional crystal structures of *Escherichia coli met* repressor with and without corepressor. *Nature* **341**, 705-710.
33. Phillips, K. and Phillips, S.E.V. (1994). Electrostatic activation of *Escherichia coli* methionine repressor. *Structure* **15**, 309-316.

APPENDIX 1:**Structures of Potential Intermediates of the T to R Transition of Hemoglobin**

Maria A. Schumacher*, Melinda M. Dixon[^], Ronald Kluger[#],
Richard T. Jones* & Richard G. Brennan*⁺

*Department of Biochemistry and Molecular Biology
Oregon Health Sciences University,
3181 S.W. Sam Jackson Park Road
Portland, Oregon 97201-3098, U.S.A.

[#]Lash Miller Laboratories, Department of Chemistry, University of Toronto
Toronto, Ontario M5S 1A1, Canada

[^]Biophysics Research Division, University of Michigan
Ann Arbor, Michigan 48109, U.S.A.

⁺To whom all correspondence should be addressed.

Telephone: (503) 494-4427

FAX: (503) 494-8393

Electronic Mail: BRENNANR@OHSU.EDU

ABSTRACT:

The structural end points of hemoglobin's transition from its low oxygen affinity (T) to high oxygen affinity (R) state, have long been established by X-ray crystallography¹⁻⁷. However, short-lived intermediates have proven more refractory to X-ray studies. One approach to characterize these intermediates structurally is through chemical cross-linking. Here, we describe the X-ray structures of three hemoglobins, $\alpha_2\beta^1S^{82}\beta$, $\alpha_2\beta^1Tm^{82}\beta$ and $\alpha_2\beta^{1,82}Tm^{82}\beta$, that were cross-linked between the amino groups of residues β Val1 and β Lys82 by 3,3'-stilbenedicarboxylic acid (S) or trimesic acid (Tm) while in the deoxy state, and saturated with carbon monoxide prior to crystallization. $\alpha_2\beta^1S^{82}\beta$, which displays near normal oxygen affinity, is completely R state in structure. However, $\alpha_2\beta^1Tm^{82}\beta$ and $\alpha_2\beta^{1,82}Tm^{82}\beta$, which have low oxygen affinities, have been stopped from fully completing their transitions into the R state conformation and display many features expected of a transitional intermediate. These hemoglobins thus appear to provide a snapshot of the nascent R state.

The cross-linked hemoglobins were prepared by reacting human deoxyhemoglobin with the diacyl bis(methyl phosphate) derivatives of 3,3'-stilbenedicarboxylic acid (S)⁸ or trimesic acid (Tm)⁹ and were crystallized in the carbonmonoxy (CO) form by the method of Perutz¹⁰. $\alpha_2\beta^1S^{82}\beta$, which displays a slightly higher O₂ affinity than HbA and high cooperativity (Table 1), crystallizes isomorphously with respect to COHbA^{1,7}. The structure has been refined¹¹ to a crystallographic R factor of 14.6% at 2.4 Å resolution (Table 1). Both $\alpha_2\beta^1Tm^{82}\beta$ and $\alpha_2\beta^{1,82}Tm^{82}\beta$, display markedly lower O₂ affinities than HbA, high cooperativity, and crystallize in space group C2 (Table 1). The structure of $\alpha_2\beta^1Tm^{82}\beta$ was solved by molecular replacement¹² and found to contain a tetramer per asymmetric unit. Refinement converged to an R-factor of 15.0% at 1.8 Å resolution (Table 1). This cross-linked hemoglobin served as the starting model, minus cross-linker, for refinement of the triply linked $\alpha_2\beta^{1,82}Tm^{82}\beta$, which converged to a final R-factor of 13.5% at 1.8 Å resolution (Table 1). The final 2F_O - F_C electron density map corresponding to the cross-linker region of $\alpha_2\beta^{1,82}Tm^{82}\beta$ is shown in Figure 1. To our knowledge, the structure of $\alpha_2\beta^{1,82}Tm^{82}\beta$ provides the first example of a rationally designed, triply linked protein⁹.

Upon binding carbon monoxide, the $\alpha_1\beta_1$ dimer of HbA undergoes a 13.6° rigid body rotation with respect to the $\alpha_2\beta_2$ dimer¹. This rotation is also observed for $\alpha_2\beta^1S^{82}\beta$ (Fig. 2). However, the $\alpha\beta$ dimers of $\alpha_2\beta^1Tm^{82}\beta$ and $\alpha_2\beta^{1,82}Tm^{82}\beta$ (TmHbs) are restricted by the trimesoyl group and rotate only 7.4° and 8.5°, respectively, suggesting that they are trapped in a transitional conformation (Fig. 2). The tertiary and quaternary differences between the cross-linked hemoglobins and deoxy and COHbA were analyzed by a series of C α coordinate-difference plots (CDP) (Fig. 3). These plots confirm the conformational identities of $\alpha_2\beta^1S^{82}\beta$ and COHbA and demonstrate that simply cross-linking the α -NH₃ of β Val1 of one $\alpha\beta$ dimer to the ϵ -NH₃ of β Lys82 of the other $\alpha\beta$ dimer does not induce large structural perturbations. Most important, these plots reveal that, unlike the structures of other proposed transitional intermediates such as R2-HbA¹³, Ypsilanti Hb¹⁴, and CN-met Hb¹⁵, those of the TmHbs lie directly on the T to R conformational pathway.

The most significant structural differences between the TmHbs and COHbA and deoxyHbA are found in elements involved directly in the T to R transition, namely, the α subunits' A, C, E and H helices and EF turns and the β subunits' A, E, F and H helices and FG turns (Fig. 3). Inspection of these structural elements reveals relaxation to the R conformation is incomplete or has not commenced and their locations result in unfavorable steric pressure or strain, which is the likely underlying cause for the reduced oxygen affinities of both subunits of the TmHbs. Specifically, their α subunit E helices are

essentially in the T state position (Fig. 4A), whereas the E helices of the β subunits are located between the T and R positions (Fig. 4B). Ultraviolet Raman studies have shown that movement of the E helices of hemoglobin is critical to relieve the strain imposed by ligand binding¹⁶. Also intermediate are the locations of the β subunits' F helices, which have moved only 0.7 Å in $\alpha_2\beta^1\text{Tm}^{82}\beta$ and 0.6 Å in $\alpha_2\beta^{1,82}\text{Tm}^{82}\beta$ from the T state position, as contrasted to a 1.4 Å shift in COHbA (Fig. 4B). Their fully ligated hemes have also not reached the R state location. However, there is no heme doming. To maintain proper coordination to the heme Fe^{2+} , the proximal histidine, βHis92 (F8), which retains its T-state χ^1 torsion angle, is significantly displaced towards the heme implying a parabolic trajectory of F8 in the T to R transition (Fig 4B). These results suggest the movements of the β subunits' E and F helices and hemes are coordinated.

In contrast to the β F helices, the F helices of the α subunits have reached the R state location in the TmHbs (Fig. 4A). However, the α hemes, which are also fully ligated and planar, are in the T state location, implying they move independently of their F helices. The conformational transition from the T state E helices to the R state F helices is accommodated by the EF turns which are structurally intermediate (Fig. 4A). The α subunit proximal histidine, αHis87 (F8), is also positioned between its T and R state locations to coordinate properly the heme Fe^{2+} . Thus, the T to R state transitional pathways of the α and β hemes are different. This finding is in accord with the structures of the T-state liganded hemoglobins, T(α -oxy, β -deoxy)Hb and T(α -met, β -met)Hb^{17,18} and likely reflects the different packing and chemical environments of the α (tight) and β (looser) hemes.

Additional notable intermediates features are found in the switch region and the β subunit carboxy-termini of the TmHbs. Inspection of the switch reveals that, although located in its R state position between $\alpha_1\text{Thr38}$ and $\alpha_1\text{Thr41}$, $\beta_2\text{His97}$ is displaced significantly towards its T state location (Fig. 4C). At the carboxy-termini all critical T state salt bridges, including that between βHis146 and βAsp94 , are disrupted⁷, yet the side chains of their neighboring residues, βTyr145 and the reactive βCys93 ^{19,20}, maintain their T state locations suggesting that their movements occur late in the T to R transition (Fig. 4D).

In conclusion, the carbonmonoxy structures of $\alpha_2\beta^1\text{Tm}^{82}\beta$ and $\alpha_2\beta^{1,82}\text{Tm}^{82}\beta$ have provided likely snapshots of hemoglobin's nascent R state and a rationale for their lowered oxygen affinities. Using these cross-linked hemoglobins as reagents, this hypothesis may be tested further by other, more dynamic approaches such as ultraviolet Raman^{16,21,22} and ultra-fast near infrared^{23,24} spectroscopies. Our structural results, combined with the finding of an inverse correlation between the O_2 affinity and cross-linker bridging distance

in a series of $\beta^1X^{82}\beta$ cross-linked hemoglobins⁸, suggests that the structures of additional hemoglobin intermediates, either more or less R-like, can be isolated by using cross-linkers of different lengths, an idea currently being pursued. Such an approach might be applicable to structural studies on other multi-state proteins.

Acknowledgments: We thank T.S. Fujita for preparing the cross-linked hemoglobins and X.-J. Zhang for help with our molecular replacement. This work was supported in part by the National Institutes of Health, the Department of Defense and the American Heart Association of Oregon.

References and Notes

1. Baldwin, J. & Chothia, C. *J. Mol. Biol.* **129**, 175-220 (1979).
2. Fermi, G., Perutz, M., Shaanan, B. & Fourme, R. *J. Mol. Biol.* **175**, 159-174 (1984).
3. Shaanan, B. *J. Mol. Biol.* **171**, 31-59 (1983).
4. Perutz, M.F. *Nature* **228**, 726-739 (1970).
5. Dickerson, R.E. & Geis, I. *Hemoglobin: structure, function, evolution, and pathology* pp. 3-63 (Benjamin/Cummings, Menlo Park, CA).
6. Perutz, M.F. *Ann. Rev. Physiol.* **52**, 1-25 (1990).
7. Derewenda, Z., Dodson, G., Emsley, P., Harris, D., Nagai, K., Perutz, M. & Reynaud, J.-P. *J. Mol. Biol.* **211**, 515-519 (1990).
8. Jones, R.T., Head, C.G., Fujita, T.S., Shih, D.T.-b., Wodzinska, J. & Kluger, R. *Biochemistry* **32**, 215-223 (1993).
9. Kluger, R., Wodzinska, J., Jones, R.T., Head, C.G., Fujita, T.S. & Shih, D.T.-b. *Biochemistry* **31**, 7551-7559 (1992).
10. Perutz, M. *J. Crystal Growth* **2**, 54-56 (1968).
11. Tronrud, D.E., Ten Eyck, L.F. & Matthews, B.W. *Acta crystallogr.* **A43**, 489-501 (1987).
12. Fitzgerald, P. M. D. *J. Appl. Crystallogr.* **21**, 273-278 (1988).
13. Silva, M.M., Rogers, P.H. & Arnone, A. *J. Biol. Chem.* **267**, 17248-17256 (1992).
14. Smith, F.R., Lattman, E.E. & Carter Jr., C.W. *Proteins: Structure, Function and Genetics*, **10**, 81-91 (1991).
15. Smith, F. R. & Simmons, K.C. *Proteins: Structure, Function and Genetics*, **18**, 295-300 (1994).
16. Rodgers, K. R. & Spiro, T. G., *Science* **265**, 1697-1699 (1994).
17. Liddington, R., Derewenda, Z., Dodson, G. & Harris, D. *Nature* **331**, 725-728 (1988).
18. Liddington, R., Derewenda, Z., Dodson, E., Hubbard, R. & Dodson, G. *J. Mol. Biol.* **228**, 551-579 (1992).
19. Perutz, M.F., Kilmartin, J.V., Nishikura, K., Fogg, J.H., Butler, P.J.G. & Rollema, H.S., *J. Mol. Biol.* **138**, 649-670 (1980).
20. Martin de Llano, J.J., Jones, W., Schneider, K., Chait, B.T., Manning, J.M., Rodgers, G., Benjamin, L.J. & Weksler, B. *J. Biol. Chem.* **268**, 27004-27011 (1993).
21. Franzen, S., Lambry, J.C., Bohn, B., Poyart, C. & Martin, J.L., *Nature Structural Biology* **1**, 230-233 (1994).

22. Mukerji, I. & Spiro, T.G. *Biochemistry* **33**, 13132-13139 (1994).
23. Anfinrud, P.A., Han, C., Hochstrasser, R.M., *Proceedings of the National Academy of Sciences of the USA* **86**, 8387-8391 (1989).
24. Lim, M., Jackson, T.A. & Anfinrud, P.A., *Proceedings of the National Academy of Sciences of the USA* **90**, 5801-5804 (1993).
25. Jones, T. A. *J. Appl. Crystallogr.* **11**, 268-272 (1978).

TABLE 1: Summary of Selected Biochemical and Crystallographic Data

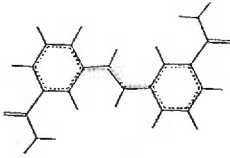
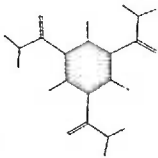
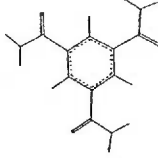
| Cross-linked Hemoglobin | $\alpha_2\beta^1S^{82}\beta$ | $\alpha_2\beta^1Tm^{82}\beta$ | $\alpha_2\beta^1.82Tm^{82}\beta$ |
|--------------------------------|---|--|---|
| Cross-linker |  |  |  |
| P_{50} (mmHg) | 3.4 ⁸ | 17.1 ⁸ | 18.1 ⁹ |
| Hill Coefficient | 2.6 ⁸ | 2.7 ⁸ | 2.6 ⁹ |
| Space Group | P4 ₁ 2 ₁ 2 | C2 | C2 |
| Cell Dimensions (Å) | $a=b=54.13$ $c=196.23$ | $a=104.43$, $b=72.16$ $c=88.03$ $\beta=108.25^\circ$ | $a=104.43$, $b=72.16$ $c=88.03$ $\beta=108.25^\circ$ |
| $\alpha\beta$ dimer per ASU | 1 | 2 | 2 |
| Data Collection | | | |
| Resolution (Å) | 2.4 | 1.8 | 1.8 |
| Number of Reflections | 9,807 | 48,608 | 42,152 |
| $R_{SYM}(\%)$ | 3.6 | 6.3 | 3.0 |
| Number of atoms | 2,300 | 4,576 | 4,577 |
| Number of solvent molecules | 110 | 393 | 429 |
| Refinement | | | |
| R-factor (%) | 14.6 | 15.0 | 13.5 |
| Deviations from Ideality (rms) | | | |
| Bond distances (Å) | 0.019 | 0.010 | 0.017 |
| Bond Angles (degrees) | 2.90 | 2.34 | 3.01 |

TABLE 1 LEGEND:

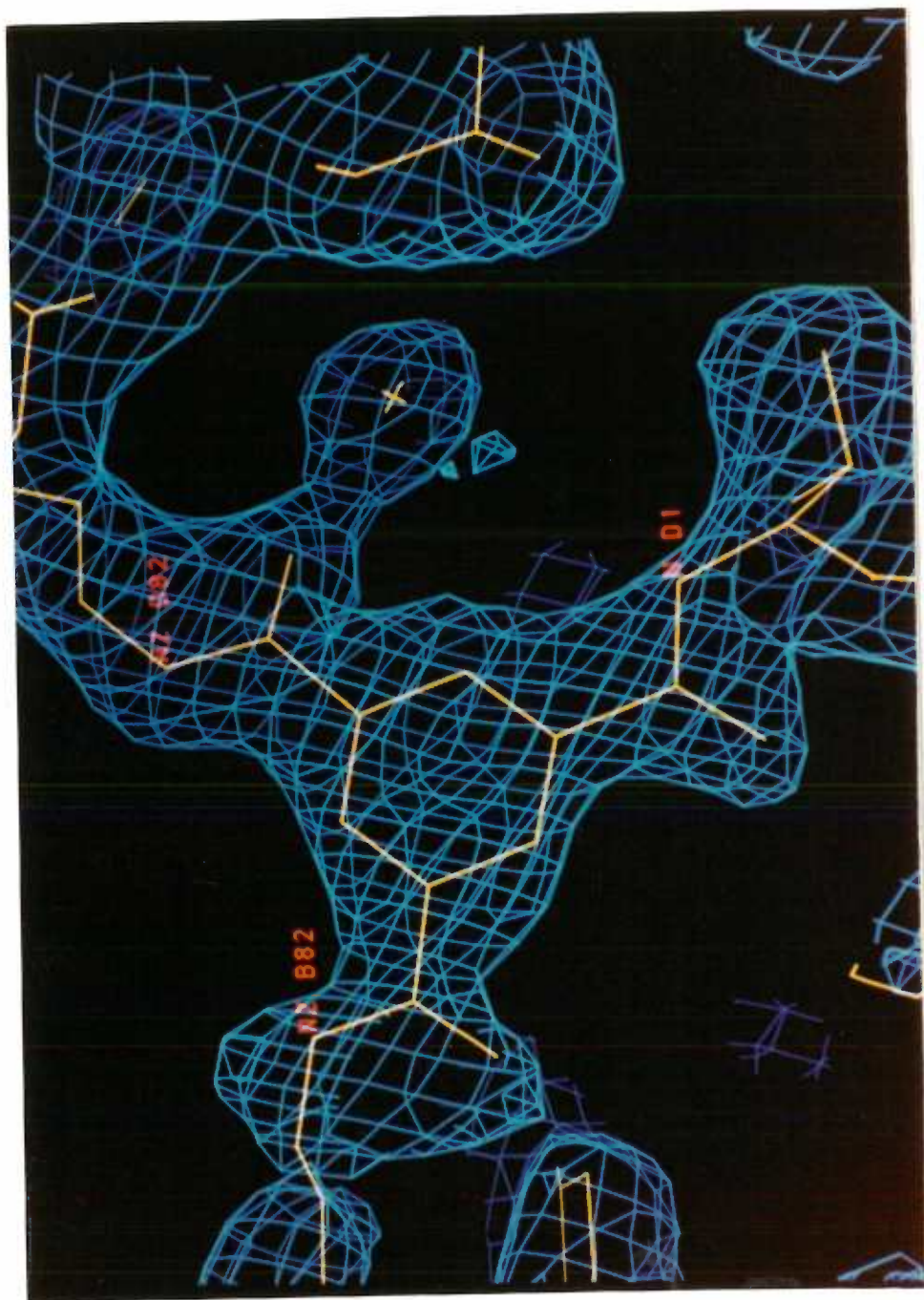
$\alpha_2\beta^1S^{82}\beta$, $\alpha_2\beta^1Tm^{82}\beta$ and $\alpha_2\beta^{1,82}Tm^{82}\beta$ were cross-linked in their deoxy forms as previously described^{8,9} by reacting human deoxyhemoglobin with the diacyl bis(methyl phosphate) derivatives of 3,3'-stilbenecarboxylic acid (S) or trimesic acid (Tm). $\alpha_2\beta^1S^{82}\beta$ and $\alpha_2\beta^1Tm^{82}\beta$ are cross-linked through the α -NH₃ of β_1 Val1 to the ϵ -NH₃ of β_2 Lys82, where the subscripts 1 and 2 indicate different β subunits only. $\alpha_2\beta^{1,82}Tm^{82}\beta$ is cross-linked through the α -NH₃ of β_1 Val1 and the ϵ -NH₃ of β_1 Lys82 to the ϵ -amino group of β_2 Lys82. The cross-linkers are depicted in their acid forms in Table 1. The negative charge of the methyl phosphate leaving groups serves to target the cross-linkers to the cationic 2,3-*bis*-phosphoglycerate (BPG) binding pocket where β Lys82 and β Val1 are located. As can be ascertained from the comparison of the P₅₀ values and Hill coefficients of the cross-linked hemoglobins to those of native HbA (P₅₀=5.0 mmHg, Hill coefficient=3.0), $\alpha_2\beta^1S^{82}\beta$ binds oxygen with an increased affinity, whereas $\alpha_2\beta^1Tm^{82}\beta$ and $\alpha_2\beta^{1,82}Tm^{82}\beta$ display reduced oxygen affinities. However, all remain cooperative^{8,9}. Crystals of the carbonmonoxy form of $\alpha_2\beta^1S^{82}\beta$, $\alpha_2\beta^1Tm^{82}\beta$ and $\alpha_2\beta^{1,82}Tm^{82}\beta$ were grown at room temperature by the batch method of Perutz¹⁰ in 2.45 M, 2.35 M and 2.35 M sodium-potassium phosphate solutions, respectively, following carbon monoxide saturation. The $\alpha_2\beta^1S^{82}\beta$ crystals are isomorphous with native carbonmonoxy hemoglobin while the Tm cross-linked hemoglobins, $\alpha_2\beta^1Tm^{82}\beta$ and $\alpha_2\beta^{1,82}Tm^{82}\beta$, are not. Data for the $\alpha_2\beta^1S^{82}\beta$ and $\alpha_2\beta^{1,82}Tm^{82}\beta$ hemoglobins were collected at room temperature with an Area Detector Systems Corporation (ADSC) area detector using a Rigaku RU200-H rotating anode generator for the X-ray source (40 kV, 150 mA) and the data were processed with the software provided by ADSC. Data for the $\alpha_2\beta^1Tm^{82}\beta$ were collected with an R-Axis II imaging plate at Molecular Structure Corporation (The Woodlands, Texas) and processed with software provided by MSC. The native carbonmonoxy hemoglobin structure minus waters was used as the starting model for the $\alpha_2\beta^1S^{82}\beta$ structure and the cross-linker was located in a difference Fourier map and fitted using FRODO²⁵. The model was refined using TNT¹¹. The $\alpha_2\beta^1Tm^{82}\beta$ hemoglobin was solved by Molecular Replacement using the MERLOT¹² software package. The native carbonmonoxy $\alpha\beta$ dimer⁷, minus waters, was used as the starting model. Two large peaks in both the rotation and translation functions located the molecule and confirmed the presence of a tetramer in the asymmetric unit (ASU). Refinement¹¹ and difference Fourier maps located the trimesoyl cross-linker, which was fitted using FRODO²⁵. Subsequent cycles of refinement followed until convergence. This structure, minus the

cross-linker and waters, was used as the starting model for the $\alpha_2\beta^{1,82}\text{Tm}^{82}\beta$ model. The cross-linker was located in a difference Fourier map and the structure was refined to convergence. Coordinates are being deposited in the Brookhaven Protein Data Bank and are available from R.G.B. Full details of the structures will be reported elsewhere (M.A.S. *et al.*, in preparation). $R_{\text{SYM}} = \Sigma |I_o - \langle I \rangle| / I_o$, where I_o is the observed intensity, and $\langle I \rangle$ is the average intensity from multiple observations of symmetry related reflections.

$R\text{-factor} = \Sigma |F_{\text{obs}} - F_{\text{calc}}| / \Sigma F_{\text{obs}}$.

Figure 1. $2F_o - F_c$ electron map of the triply linked $\alpha_2\beta_1^{1.82}\text{Tm}^{82}\beta$ hemoglobin contoured at 1.0σ . The atoms are represented as yellow sticks and the electron density map is indicated in blue²⁵. The cross represents the location of a water molecule that is hydrogen bonded to the carbonyl group of the cross-linker. Shown is the trimesoyl cross-linker bonded to the $\epsilon\text{-NH}_3$ of $\beta_1\text{Lys82}$ (N ζ D82), the $\alpha\text{-NH}_3$ of $\beta_1\text{Val1}$ (N D1) and the $\epsilon\text{-NH}_3$ of $\beta_2\text{Lys82}$ (N ζ B82), providing the first example of a rationally designed triply linked protein.

Figure 2. Stereo view of the overlay of the $\alpha_1\beta_1$ interface of deoxyhemoglobin (blue C α trace), $\alpha_2\beta^1\text{Tm}^{82}\beta$ (yellow C α trace), $\alpha_2\beta^{1,82}\text{Tm}^{82}\beta$ (green C α trace) and $\alpha_2\beta^1\text{S}^{82}\beta$ (white C α trace) onto COHbA (red C α trace) using the method of Baldwin and Chothia¹. This figure dramatically illustrates the quaternary changes that occur in the T to R transition. In this overlay method the regions in which no structural changes occur in going from T to R are used as a reference frame and include residues $\alpha 30 - \alpha 36$ (B helix); $\alpha 102 - \alpha 113$ (G helix); $\alpha 117 - \alpha 127$ (H helix); $\beta 30 - \beta 36$ (B helix); $\beta 51 - \beta 55$ (D helix); $\beta 107 - \beta 132$ (G and H helices). After overlaying the above described region of the $\alpha_1\beta_1$ dimers, a rotation of 13.6° is required to bring the $\alpha_2\beta_2$ interfaces of the deoxy and COHbA into coincidence. The TmHbs appear intermediate in this transition as they have rotated only 7.4° and 8.5° about an axis of rotation nearly coincident to that used by native COHbA. Their translation components, approximately half (0.44) of that of native Hb, are also intermediate. No additional rotation or translation is required for the $\alpha_2\beta^1\text{S}^{82}\beta$ hemoglobin, consistent with its complete R state conformation. Coordinates used in all analyses were taken from references 7 (COHbA) and 2 (deoxyhemoglobin).



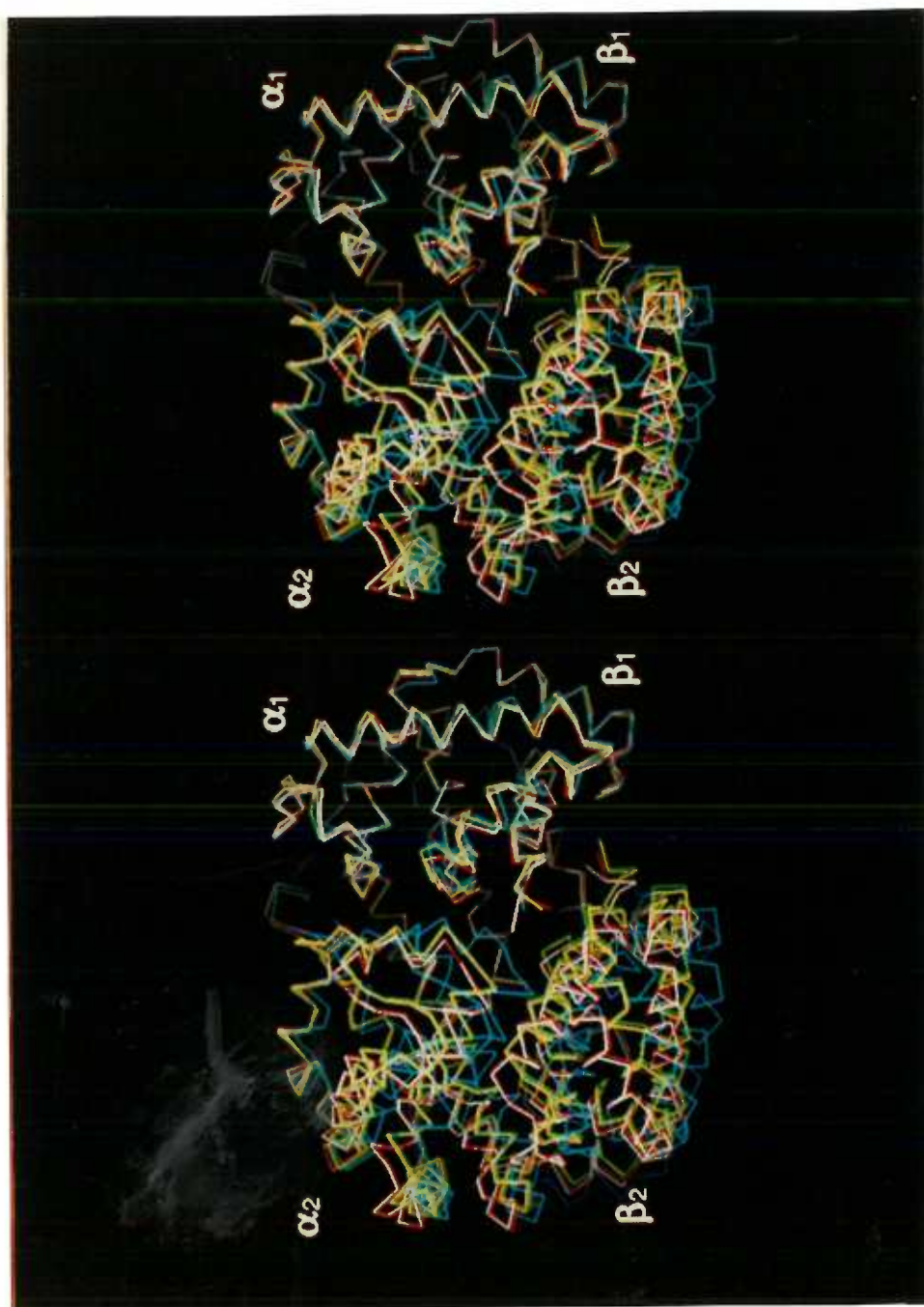


Figure 3. Coordinate difference plot (CDP) showing the differences between corresponding C α positions of deoxyhemoglobin (background, shaded in blue), $\alpha_2\beta^1\text{Tm}^{82}\beta$ (near background, shaded in yellow), $\alpha_2\beta^{1.82}\text{Tm}^{82}\beta$ (foreground, shaded in green) and $\alpha_2\beta^1\text{S}^{82}\beta$ (unshaded, i.e., white) after their $\alpha_1\beta_1$ interfaces are overlaid on COHbA as in Figure 2. The ordinate indicates the displacement in Å for the corresponding C α atoms and the abscissa indicates the residue number where the A and B chains correspond to the α_1 and β_1 chains, respectively, and the C and D chains, to the α_2 and β_2 chains, respectively. The region designated "tertiary changes" corresponds to the $\alpha_1\beta_1$ region that was overlaid and refer to structural alterations between tertiary elements, for example, helical displacements. The structural elements that undergo the most significant displacements are labelled on the plot. The region labelled "quaternary changes" corresponds to the $\alpha_2\beta_2$ region, which was not overlaid, and corresponds to quaternary differences between deoxyhemoglobin, $\alpha_2\beta^1\text{Tm}^{82}\beta$, $\alpha_2\beta^{1.82}\text{Tm}^{82}\beta$ and $\alpha_2\beta^1\text{S}^{82}\beta$ and COHbA. The quaternary differences between the TmHbs and COHbA are made obvious by these plots and suggest that the TmHbs take an intermediate quaternary conformation. A similar overlay of the $\alpha_2\beta_2$ interfaces of deoxyhemoglobin, $\alpha_2\beta^1\text{Tm}^{82}\beta$, $\alpha_2\beta^{1.82}\text{Tm}^{82}\beta$ and $\alpha_2\beta^1\text{S}^{82}\beta$ onto that of COHbA produces a nearly identical CDP.

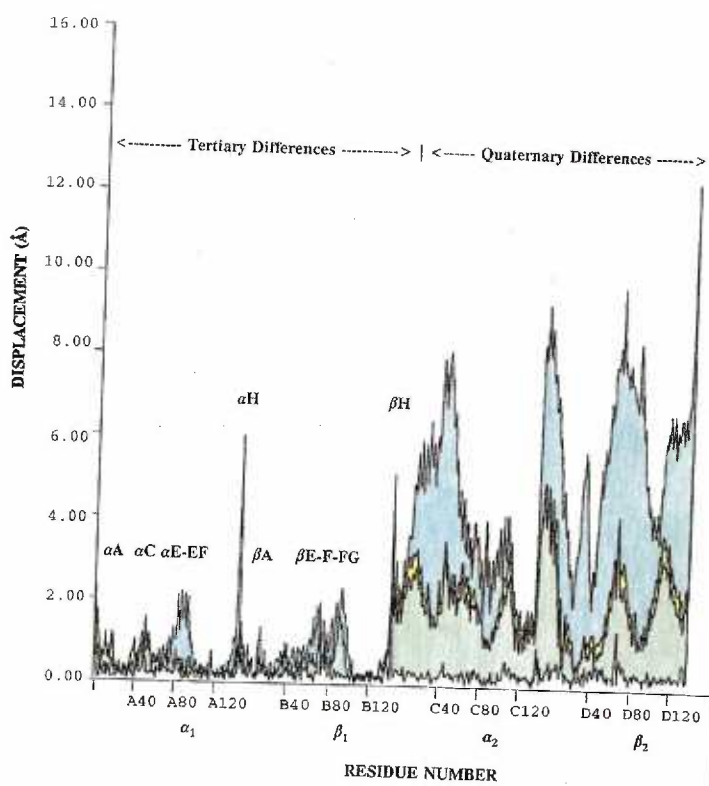


Figure 4. Snapshots of the T to R transition in hemoglobin. All views are the result of the $\alpha_1\beta_1$ overlay described in the Figure 2 legend. As in Figure 1, deoxyhemoglobin is shown in blue, $\alpha_2\beta^1\text{Tm}^{82}\beta$ in yellow, $\alpha_2\beta^{1,82}\text{Tm}^{82}\beta$ in green, $\alpha_2\beta^1\text{S}^{82}\beta$ in white and COHbA in red.

A. The α_1 heme environments showing the proximal histidines, $\alpha\text{His}87$ (F8) and the heme groups as solid rendered bonds and the A, E and F helices as C α traces. The T state positions of the E helices and hemes of the TmHbs are evident as are the R state positions of their F helices. The A helices of the TmHbs are also closer to the T state position, a finding which is consistent with recent UV raman studies¹⁶. The α hemes are fully ligated and planar. The T state positions of the E helices and the α hemes and the R state positions of the F helices results in steric pressure which requires F(8) to move to a position between its normal T and R state locations so that the Fe^{2+} atom of the α heme maintains proper octahedral geometry.

B. The β_1 heme environments showing the proximal histidines, $\beta\text{His}92$ (F8) and the heme groups as solid rendered bonds and the A, E and F helices as C α traces. The intermediate locations of the A, E and F helices of the TmHbs are evident as are the significant displacements of their proximal histidines towards the heme. Despite the strain resulting from the intermediate nature of the E and F helices of the TmHbs, the β hemes are fully ligated and planar. The shifts of the β F helices were quantitated (see text) by averaging the C α displacements (\AA) of each F helix residue of deoxyhemoglobin, $\alpha_2\beta^1\text{Tm}^{82}\beta$, $\alpha_2\beta^{1,82}\text{Tm}^{82}\beta$ and $\alpha_2\beta^1\text{S}^{82}\beta$ relative to the corresponding residues in COHbA.

C. The switch region of hemoglobin, which is composed of residues 97 through 102 from the β_2 FG corner and G helix and residues 38 through 44 from the C helix and CD corner of the α_1 subunit, demonstrating the significant displacements of the key switch residue $\beta_2\text{His}97$ of $\alpha_2\beta^1\text{Tm}^{82}\beta$ (yellow) and $\alpha_2\beta^{1,82}\text{Tm}^{82}\beta$ (green) toward the T state position. The distances from the $\beta_2\text{His}97$ C α atoms of $\alpha_2\beta^1\text{Tm}^{82}\beta$ and $\alpha_2\beta^{1,82}\text{Tm}^{82}\beta$ to the corresponding $\beta_2\text{His}97$ C α atom in COHbA (red) are 1.26 \AA and 1.47 \AA , respectively, compared to a distance of 7.12 \AA for deoxyhemoglobin (blue). Despite these displacements, the $\beta_2\text{His}97$ of the TmHbs are clearly in the R conformation as evidenced by their locations between $\alpha_1\text{Thr}38$ and $\alpha_1\text{Thr}41$. In deoxyhemoglobin, His97 sits between $\alpha_1\text{Thr}41$ and $\alpha_1\text{Pro}44$.

D. The pocket formed by the β F and H helices showing the T state position of the side chains of $\beta\text{Cys}93$ and $\beta\text{Tyr}145$ of $\alpha_2\beta^1\text{Tm}^{82}\beta$ (yellow) and $\alpha_2\beta^{1,82}\text{Tm}^{82}\beta$ (green). The side chains of $\beta\text{Cys}93$ and $\beta\text{Tyr}145$ of the $\alpha_2\beta^1\text{S}^{82}\beta$ hemoglobin (white), however, are clearly in the R state location. Deoxyhemoglobin and COHbA are shown in blue and red, respectively.

