

The Engrailed Homeodomain:  
Determinants of  
DNA-Binding Affinity and Specificity

by

Sarah Ellen Ades

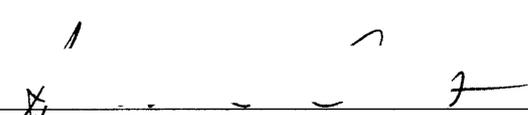
B.S., Molecular Biophysics and Biochemistry  
Yale University  
May, 1988

Submitted to the Department of Biology  
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the  
Massachusetts Institute of Technology  
September 1995

© 1995 by Sarah Ellen Ades. All rights reserved.  
The author hereby grants to MIT permission to reproduce and to  
distribute copies of this thesis document in whole or in part.

  
Signature of Author

Department of Biology

Certified by

  
Robert T. Sauer, Thesis Supervisor

Accepted by

  
Frank Solomon, Chairman, Biology Graduate Committee

MASSACHUSETTS INSTITUTE

OF TECHNOLOGY

AUG 04 1995

~~05/08/1995~~

LIBRARIES

Science

**The Engrailed Homeodomain:  
Determinants of DNA-Binding Affinity and Specificity**

by

**Sarah Ellen Ades**

Submitted to the Department of Biology  
in partial fulfillment of the requirements for the degree of Doctor of Philosophy

**Abstract**

The work presented in this thesis examines the DNA-binding activity of the homeodomain from the *Drosophila* transcription factor engrailed. Mutagenesis studies of both the protein and the DNA are used to address the contribution of protein-DNA contacts to the affinity and specificity of the complex.

Chapter 1 presents an overview of specificity in protein-DNA complexes from both structural and biochemical perspectives.

Chapter 2 describes the role of residue 50 in determining the DNA-binding specificity of the engrailed homeodomain. Binding site selections identify the site TAATTA as the consensus binding site for the wild-type engrailed homeodomain. A single amino acid substitution at position 50 of the homeodomain, glutamine → lysine, changes the binding site preference to TAATCC and increases the affinity and half-life of the homeodomain-DNA complex. The wild-type glutamine makes only a small contribution to the overall binding energy since replacing the glutamine with an alanine (QA50) reduces the binding energy by only two-fold. The QA50 homeodomain is able to discriminate between the TAATTA and TAATCC sites as well as the wild-type Gln50 protein suggesting that the amino acid at position 50 is not the sole determinant of differential specificity. These experiments were published as "Differential DNA-Binding Specificity of the Engrailed Homeodomain: The Role of Residue 50" (Ades, S. E. & Sauer, R. T. (1994) *Biochemistry* 33, 9187-9194).

Chapter 3 describes the relative contributions to specificity and affinity of two parts of the homeodomain-DNA complex: interactions in the minor groove by residues from the flexible N-terminal arm and interactions in the major groove from residues of the recognition helix. An altered-specificity variant of the engrailed homeodomain with lysine at position 50 was used for these studies. Minor groove interactions from the arm are shown to have a comparable contribution to binding affinity but a lower specificity than major groove interactions from the recognition helix. Although the specificity is moderate, the homeodomain can discriminate among bases in the minor groove and shows a base preference at the positions of minor groove interactions. These experiments have been submitted for publication as "Specificity of Minor-Groove and Major-Groove Interactions in a Homeodomain-DNA Complex" (Ades, S. E., & Sauer, R. T. (1995) *submitted*).

Chapter 4 summarizes the results of the experiments presented in chapters 2 and 3 and describes further directions for research into homeodomain-DNA interactions.

Thesis Supervisor: Robert T. Sauer

Title: Whitehead Professor of Biochemistry

*to my grandparents*

*Louis and Frances Ades  
and  
Joseph and Libby Fleishman*

## Acknowledgements

I would like to thank my family for their support and encouragement. My mother for inspiring me to get a Ph.D., my father for his interest in the mysteries of biochemistry, and my sister, the doctor, and my brother, the fireman, for their friendship.

Thanks to Bob Sauer, my advisor, for guidance, advice, and teaching me how to do good science.

Thanks to members of the Sauer lab past and present for their help and insights.

Thanks to the members of my thesis committee - Carl Pabo, Phil Sharp, and Uttam Rajbhandary - for advice and ideas, and to Roger Brent for serving on my thesis defense committee.

Thanks to Carl Pabo for many helpful discussions about homeodomains and science. And thanks to the crystallographers in the Pabo lab for their interest in the homeodomain project and their ongoing efforts.

Finally, thanks to Ken for support, advice, and friendship throughout graduate school.

## Table of Contents

Abstract	2
Dedication	4
Acknowledgements	5
Table of Contents	6
List of Tables	7
List of Figures	8
Chapter 1: Origins of Specificity in Protein-DNA Complexes	10
Chapter 2: Differential DNA-Binding Specificity of the Engrailed Homeodomain: the Role of Residue 50	47
Chapter 3: Specificity of Minor-Groove and Major-Groove Interactions in a Homeodomain-DNA Complex	79
Chapter 4: Homeodomain-DNA Recognition	108

## List of Tables

### Chapter 2

Table 1: Equilibrium and Kinetic DNA-Binding Constants	66
--	----

### Chapter 3

Table 1: Equilibrium and Kinetic DNA-Binding Constants for Alanine Substitution Mutants	100
---	-----

Table 2: Tabulation of Data from Binding Site Selections Using <u>NNATCC</u>	101
--	-----

Table 3: Equilibrium DNA-Binding Constants to Altered Sites	102
---	-----

## List of Figures

### Chapter 1

Figure 1:	Families of DNA-Binding Proteins	14
Figure 2:	Contacts between amino acids and functional groups on the edges of base pairs in the major groove of the DNA	22, 23
Figure 3:	Base contacts in the EcoRI structure	34
Figure 4:	Base contacts in the $\lambda$ repressor structure	36
Figure 5:	Base contacts in the Arc repressor structure on the right half-site	39

### Chapter 2

Figure 1:	Molecular graphics representation of the engrailed homeodomain bound to DNA	67
Figure 2a:	Sequence of the gene constructed to encode the engrailed homeodomain	69
Figure 2b:	Sequences of DNA fragments used for binding assays	69
Figure 2c:	Sequences of synthetic oligonucleotides for binding site selections	69
Figure 3:	CD spectra and thermal denaturation of wild-type, QA50, and QK50 proteins	70
Figure 4:	Gel mobility shift assays from the first and final rounds of binding site selections for the wild-type engrailed homeodomain	72
Figure 5a:	Aligned individual binding sites for the engrailed homeodomain obtained after <i>in vitro</i> selection	74
Figure 5b:	Tabulation of the aligned data	75

Figure 6:	Binding site preferences at positions 5 and 6 for the wild-type, QA50, and QK50 engrailed homeodomains following <i>in vitro</i> selections using <u>TAATNN</u>	76
Figure 7:	Equilibrium binding curve for the wild-type engrailed homeodomain	77
Figure 8:	Dissociation kinetics of complexes	78
Chapter 3		
Figure 1:	Molecular graphics representation of the engrailed homeodomain bound to DNA	103
Figure 2:	Contacts between the altered-specificity homeodomain and DNA	104
Figure 3:	Base pairs used for binding site substitutions	105
Figure 4:	Comparison of specificity inferred from binding site selections and affinity measurements at positions 1, 2, 5, and 6	107

## **Chapter One**

### **Origins of Specificity in Protein-DNA Complexes**

The ability of a DNA-binding protein to bind to a unique DNA site is essential for many cellular functions. A detailed analysis of protein-DNA interactions is important to understand how a protein can recognize a specific base sequence among all of the DNA in the genome. Structural and biochemical studies of protein-DNA interactions provide complementary approaches toward understanding the basis of site-specific recognition. A crystal structure of a protein-DNA complex renders a detailed three-dimensional picture of the interactions between the protein and its binding site, while detailed biochemical analysis gives information about the importance of the contacts observed in the structure to the overall binding energy and specificity of the protein.

Over the last decade the number of cocrystal structures of protein-DNA complexes reported each year has steadily increased. To date, high resolution cocrystal structures have been solved of 37 different sequence-specific DNA binding proteins, including prokaryotic and eukaryotic transcription factors, restriction enzymes, a methylase, and a recombinase (see below for a complete list of the structures cited in this chapter). In addition, the structures of several of these proteins bound to more than one DNA site have been determined: 434 repressor with  $O_{R1}$ ,  $O_{R2}$ , and  $O_{R3}$  (Aggarwal et al., 1988; Rodgers & Harrison, 1993; Shimon & Harrison, 1993); GCN4 with ATF/CREB and AP-1 sites (Ellenberger et al., 1992; Konig & Richmond, 1993); estrogen receptor DNA-binding domain with ERE-CON and ERE-VitB1 sites (Schwabe et al., 1993; Schwabe et al., 1995); and NF- $\kappa$ B with an idealized  $\kappa$ B target and a MHC class I enhancer site (Ghosh et al., 1995; Muller et al., 1995). In contrast, detailed thermodynamic studies evaluating the functional importance of the

observed contacts have been performed for far fewer of these proteins (Sarai & Takeda, 1989; Lesser et al., 1990; Lesser et al., 1993; Brown et al. 1994).

An examination of both biochemical and structural aspects of protein-DNA interactions can lead to a greater understanding of these macromolecular recognition events. Structural studies define the architecture of an interaction while biochemical studies define the functional relevance of the interaction. My work, presented in Chapters 2 and 3 of this thesis, focuses on the determinants of DNA recognition in the engrailed homeodomain-DNA complex. In this chapter I will present an overview of specificity in a larger group protein-DNA complexes from both crystallographic and biochemical perspectives. In the first two sections, I examine the general principles of site-specific recognition from the cocrystal structures focusing first on the surfaces of the proteins which are used to bind to DNA and then on individual contacts in the protein-DNA complexes. In the last section of the chapter, I consider the functional implications of principles of recognition derived from a structural perspective, by examining biochemical studies which assess the importance of contacts observed in the structures through equilibrium binding studies of mutant proteins and DNA sites.

From a brief glance through the cocrystal structures, it is immediately apparent that there are many different ways in which proteins bind to DNA. Proteins contact the DNA in both the major and minor groove using amino acids from all units of secondary structure:  $\alpha$ -helices,  $\beta$ -sheets, loops, turns, and segments of extended polypeptide chain. Many of the proteins can be grouped into families which use a common folded structure to recognize

DNA and dock on the DNA in a similar manner. Among members of a family, base contacts are often formed by residues at comparable positions of the DNA-binding motif. The families of DNA-binding proteins, and their members represented among the existing cocrystal structures, are listed in Figure 1 below. The proteins are grouped first according to the main unit of secondary structure used for DNA recognition, then into superfamilies which share certain structural features, and then into families which have a conserved fold and dock onto the DNA in a conserved fashion. The following section will focus on how these proteins use units of regular and irregular secondary structure to bind to DNA, focusing first on interactions with the major groove and then on interactions with the minor groove.

---

**Figure 1: (following page) Families of DNA-binding proteins.** The cocrystal structures of complexes considered in this chapter are listed. All references to these complexes in the text are from the following citations:  $\lambda$  repressor (Jordan & Pabo, 1988; Clarke et al., 1991; Beamer & Pabo, 1992), 434 repressor ( $O_R1$ -(Aggarwal et al., 1988), 434 cro (Mondragon & Harrison, 1991), CAP (Schultz et al., 1991), Trp repressor (Otwinowski et al., 1988; Lawson & Carey, 1993), PurR (Schumacher et al., 1994), Oct-1 (Klemm et al., 1994), Engrailed homeodomain (Kissinger et al., 1990), Mat- $\alpha 2$  (Wolberger et al., 1991), HNF-3/Forkhead (Clark et al., 1993), Paired Domain (prd, (Xu et al., 1995), Hin recombinase (Feng et al., 1994), Zif268 (Pavletich & Pabo, 1991), Gli (Pavletich & Pabo, 1993), Tramtrack (ttk, Fairall et al., 1993), GAL4 (Marmorstein et al., 1992), PPR1 (Marmorstein & Harrison, 1994), Estrogen Receptor DNA-binding domain (ER-DBD, Schwabe et al., 1993), Glucocorticoid Receptor DNA-binding domain (GR-DBD, Luisi et al., 1991), GCN4 (Ellenberger et al., 1992; Konig & Richmond, 1993), c-fos/c-jun (Glover & Harrison, 1995), MyoD (Ma et al., 1994), USF (Ferré-D'Amaré et al., 1994), Max (Ferré-D'Amaré et al., 1993), E47 (Ellenberger et al., 1994), EcoRI (McClarín et al., 1986; Kim et al., 1990; Rosenberg, 1991), p53 (Cho et al., 1994), bovine papillomavirus E2 protein (E2, Hegde et al., 1992), PvuII (Cheng et al., 1994), Arc repressor (Raumann et al., 1994), MetJ repressor (Somers & Phillips, 1992), TATA Binding Protein (TBP, (Kim et al., 1993b; Kim et al., 1993a), EcoRV (Winkler et al., 1993), HhaI Methyltransferase (Klimasauskas et al., 1994), and NF- $\kappa$ B p50 subunit (Ghosh et al., 1995; Muller et al., 1995).

## Families of DNA-Binding Proteins

### Helices:

#### Helix-turn-Helix

##### 1) Classic Helix-turn-Helix

*λ Repressor, 434 Repressor, 434 Cro, CAP, Trp Repressor, PurR, Oct-1 (POU-specific domain)*

##### 2) Homeodomain

*Engrailed, Mat α2, Oct-1 (homeodomain)*

##### 3) Winged Helix

*HNF-3/Forkhead*

##### 4) Homeodomain-Like

*Paired domain, Hin Recombinase*

#### Metal Binding Domains

##### 1) Cys<sub>2</sub>His<sub>2</sub> Zinc Fingers

*Zif268, Gli, Tramtrack*

##### 2) Zn<sub>2</sub>His<sub>6</sub> Binuclear Cluster

*GAL4, PPR1*

##### 3) Nuclear Hormone Receptor

*Estrogen Receptor, Glucocorticoid Receptor*

#### bZIP and bHLH

##### 1) bZIP

*GCN4, c-fos/c-jun*

##### 2) bHLH and bHLHzip

*MyoD, USF, E47, Max*

#### Others

*EcoRI, p53, E2*

### β-Sheet:

#### Two-Stranded, Antiparallel

*PvuII*

#### Ribbon-Helix-Helix

*Arc Repressor, MetJ Repressor*

#### β-Sheet

*TATA Binding Protein*

### Loops:

*EcoRV, Hha1 Methyltransferase, Nf-κB (p50 subunit)*

*Modes of Recognition in the Major Groove:* The majority of protein-DNA interactions observed in the cocrystal structures occur in the major groove of the DNA and all units of secondary structure can provide a surface for DNA recognition in the wide major groove of B-form DNA.

*$\alpha$ -Helices:*  $\alpha$ -Helices are the most common unit of secondary structure used as a scaffold for protein-DNA interactions in the cocrystal structures solved to date. The size and shape of an  $\alpha$ -helix is well suited for protein-DNA recognition, particularly in the major groove of the DNA as noted by Pabo and Sauer (1992). In the proper orientation, an  $\alpha$ -helix can fill the major groove, and residues from different positions along the helix can contact bases on either strand of the DNA and the sugar-phosphate backbone on either side of the major groove.

Despite the fact that a large number of DNA-binding proteins use helices to bind to DNA, the helices can be positioned in the major groove in quite different fashions. Proteins of the classic and homeodomain-like HTH families, the Cys<sub>2</sub>His<sub>2</sub> family of Zn fingers, and EcoRI, bind to DNA with the N terminus of a helix angled into the major groove. The Zn<sub>2</sub>Cys<sub>6</sub> binuclear cluster proteins, GAL4 and PPR1, bind to DNA with the C terminus of a helix inserted into the major groove. Members of the homeodomain, winged-helix, nuclear receptor, bZIP, and bHLH families and the as yet unclassified E2 and p53 proteins bind to DNA by inserting a helix into the major groove such that the length of the helix runs along the major groove.

$\alpha$ -Helices are not only docked on the DNA in a variety of orientations, but they are also presented to the DNA in a variety of contexts. For most of

the proteins which use helices to recognize DNA, the helices are part of a larger, globular fold of the DNA-binding domain and residues from one face of the helix contribute to the hydrophobic core of the DNA-binding domain while residues on the other face of the helix contact bases. In contrast, proteins of the bZIP/bHLH superfamily have a particularly interesting mode of recognition: the basic region helices, which are responsible for all of the base contacts, are not part of a globular structure and extend through the major groove of the DNA as isolated helices. It has been shown for several members of this family (GCN4, Weiss et al., 1990; USF, Ferré-D'Amaré et al., 1994; MyoD, C. O. Pabo, personal communication), that the basic regions are unstructured in the absence of DNA and become helical upon DNA binding. In effect, these proteins bind to DNA using an induced fit mechanism rather than by docking a stable, pre-formed surface to the DNA.

*β-Sheets:* Although helices are the most common scaffold for major-groove interactions,  $\beta$ -sheets can also be used to recognize DNA as seen in the cocrystal structures of four protein-DNA complexes. Three of the four proteins (MetJ repressor, Arc repressor, and the restriction endonuclease PvuII) use a two-stranded antiparallel  $\beta$ -sheet as a scaffold for major-groove interactions. The two-stranded sheet is inserted into the major groove and residues on the side of the ribbons facing the floor of the groove contact bases. Raumann et al. (1994) noted that such  $\beta$ -sheets do not fill the major groove to the same extent as  $\alpha$ -helices, therefore contacts to the sugar-phosphate backbone may be especially important to hold the sheet in the proper orientation in the major groove. The fourth protein using  $\beta$ -sheets for DNA recognition, TATA binding protein (TBP), contacts the DNA solely in the minor groove and is discussed below.

*Loops, Turns and Extended Chain:* Loops, turns and extended regions of polypeptide chain also provide surfaces for major-groove recognition. In three protein-DNA structures, those of NF- $\kappa$ B, HhaI methyltransferase, and EcoRV, all of the major-groove contacts are formed with residues from loop regions of the protein which form irregular structures. In several other complexes, these irregular units of secondary structure are used in conjunction with helices or sheets to provide additional DNA interactions.  $\lambda$  repressor, the homeodomain proteins, the paired domain and Hin recombinase all contain N-terminal or C-terminal regions of polypeptide chains which adopt an extended structure upon DNA binding and provide additional base contacts. EcoRI forms base contacts in the major groove with side chains from a region of extended polypeptide chain in addition to forming contacts with side chains from two  $\alpha$ -helices. Finally, both the HNF-3/forkhead domain and p53 use a  $\alpha$ -helix and a loop to contact bases in the binding site.

*Modes of Recognition in the Minor Groove:* While the majority of protein-DNA interactions involve the major groove, several proteins also interact with the minor groove. Interactions between the protein and the minor groove are seen in fourteen cocrystal structures and these interactions are generally coupled to interactions with the major groove. The minor groove is the sole source of protein-DNA interactions in only one of the complexes, that of TBP. Although  $\alpha$ -helices are the most common scaffold for interactions in the major groove; loops, turns and extended regions provide the main source of amino acid contacts in the minor groove. The narrower and deeper minor groove of standard B-form DNA cannot easily

accommodate an  $\alpha$ -helix or  $\beta$ -sheet and, in the structures of proteins which do use such surfaces to contact bases in the minor groove, the DNA is severely distorted.

*Loops, Turns and Extended Chain:* Residues from loops, turns, and extended polypeptide chain interact with bases in the minor groove without causing any major changes in DNA structure. The DNA binding domains of 434 repressor, 434-cro, HNF-3, p53, PvuII and EcoRV all show similar minor groove-protein interactions in which a loop or turn traverses the minor groove and inserts a side chain into the minor groove. The homeodomains and the homeodomain-like proteins make more extensive contacts with the minor groove. The N-terminal arms of these proteins (with the exception of the paired domain) lie in the minor groove allowing side chains to contact the edges of base pairs in the minor groove. In the Paired domain complex, residues from a  $\beta$ -turn are inserted into the minor groove. In addition, both Hin recombinase and the Paired domain have C-terminal tails that lie in the minor groove and use the polypeptide backbone rather than the side chains to contact the floor of the groove.

*$\alpha$ -Helices and  $\beta$ -Sheets:* In contrast to the interactions mentioned above which do not alter the structure of the DNA, minor groove interactions by  $\alpha$ -helices and  $\beta$ -sheets, as seen in the PurR and the two TBP structures, are accompanied by significant distortions of the DNA. For both PurR and TBP, hydrophobic side chains intercalate between base pairs of the DNA duplex, causing a kink in the DNA and a bend away from the protein toward the major groove. As a result, the DNA is underwound and the minor groove becomes wider and shallower allowing an  $\alpha$ -helix or  $\beta$ -sheet to

be accommodated. In the PurR complex, a pair of leucines pry apart the central base pairs of the binding site, thereby distorting the DNA so that the two hinge helices can lie side-by-side in the minor groove. In the TBP complex, two pairs of phenylalanines intercalate between base pairs at either end of the binding site resulting in an almost total unwinding of the DNA helix. All of the protein-DNA contacts are made in this widened minor groove by an eight stranded  $\beta$ -sheet which lies over the DNA.

### *Direct Protein-DNA Contacts*

The complementarity of the binding surfaces of the protein and DNA provides the basis for specific recognition which can be divided into two primary components: direct read-out and indirect read-out. Direct read-out refers to contacts between protein side chains and functional groups along the edges of base pairs of the DNA and is discussed in detail below for both major-groove and minor-groove interactions. Indirect read-out refers to the sequence-dependent alterations in DNA structure which allow a protein to form optimal contacts with a given binding site. This is often reflected in the observation that bases which are not directly contacted by the protein are important for binding specificity and affinity (Bell & Koudelka, 1993; Bell & Koudelka, 1995).

Although direct contacts probably form the main determinants of specific recognition, for many proteins additional specificity appears to be garnered from modest deformations of the DNA structure which allow optimal docking of the protein and DNA. In other cases, DNA binding is accompanied by more dramatic structural changes. The DNA in the CAP,

EcoRI, and EcoRV structures is bent or kinked. As discussed above, the DNA in the PurR and TBP structures is even more drastically distorted. In these cases the DNA structure itself, which is a function of the base composition of the binding site, can contribute to specificity as different sequences can have different structural properties.

#### *Direct Read-Out in the Major Groove:*

Direct interactions between protein side chains and functional groups on the edges of base pairs on the DNA form the foundation of site-specific DNA recognition. To examine the nature of these contacts in greater detail, I have compiled a list of interactions seen in the cocrystal structures between amino acids and functional groups on bases in the major groove (Figure 2a, b). There are fewer examples of interactions in the minor groove and they will be discussed separately. Earlier compilations of side chain-base contacts using fewer structures suggested that there is no obvious base-recognition code (Pabo & Sauer, 1992). Today, significantly more cocrystal structures have been solved and it is worth re-examining the contacts seen in the structures to find if any general trends of recognition emerge with a larger database. The contacts are compiled from the authors' descriptions of the cocrystal structures. For proteins which bind to DNA as oligomers and form symmetric contacts with the DNA, contacts were only counted from one of the half-sites. Although water-mediated contacts appear to contribute to specificity in some cases (for Trp repressor, in particular), they were not included in the list as many of the structures are not of sufficiently high resolution to assign waters unambiguously. The structures from which the contacts were compiled, and the abbreviations used in the table, are listed in the legend for Fig. 1.

**Figure 2: (following pages) Contacts between amino acids and functional groups on the edges of base pairs in the major groove of the DNA.** Amino acids are referred to by the protein and residue number. Residues which form more than one hydrogen bond are in bold-face type and each contact is listed. Residues marked with (') indicate contacts from the second monomer. As noted in the text, symmetric contacts from oligomeric proteins were only listed for one half-site. Diagonal stripes filling a section of the chart indicate that the interaction is not allowed chemically. Hydrophobic interactions with the 5-methyl group of thymine are only cited when they are explicitly noted by the authors of the respective paper as van der Waals interactions. (a) Contacts by Arg, Lys, Asn, and Gln. A column has been included for residues which form pairs of hydrogen bonds with the N7 and O6 of a single guanine base and with the N7 and N6 of a single adenine base. (b) Contacts formed by the remaining amino acids and the polypeptide backbone.

Figure 2a:

	Guanine			Adenine			Thymine		Cytosine
	N7	O6	N7 and O6	N7	N6	N7 and N6	5 Me	O4	
<b>Arg</b>	Hin 178 ER 33 GII 149 E47 346 Max 36 USF 212 MyoD 111	Arc 13 <b>Oct1p 49</b>	TTK 152 TTK 128 zif 18 zif 74 zif 24 zif 80 zif 46 Hha 240 GCN4 243 GR 466	Cap 180 Cap 185 PurR 26 Oct1p 49 Mat $\alpha$ 2 54 Arc 13' NF-KB 57 NF-KB 59 p53 280	Arc 13' (I) E47 344 EcoRI 145 EcoRI 145			Cap 185	
<b>Lys</b>	$\lambda$ 3 PPR1 41 Gal4 18 NF-KB 244 Pvd 52 GR 461 ER 32	$\lambda$ 3 PPR1 41 Gal4 18 NF-KB 244 $\lambda$ 4 $\lambda$ 4 Met J 23' ER 28	MetJ 23 E2 339 p53 120 GII 150					NF-KB 244 ER32	
<b>Gln</b>		434Cro 29 Hha 237	434R 29			434R 28 $\lambda$ 44 434Cro 28 Oct1p 44 Arc 9' Arc 9	En 50	434R 33	
<b>Asn</b>	$\lambda$ 55 E47 341		PvuII 141	E2 336	EcoRI 141 EcoRI 141	Oct1hd 51 HNF3 165 En 51 Mat $\alpha$ 2 51 TTK 155 TTK 125 PvuII 140 EcoRV 185		Arc 11' GCN4 235 fos/jun	Arc 11 GCN4 235 fos/jun E2 336

Figure 2b:

	Guanine		Adenine		Thymine		Cytosine	
	N7	O6	N7	N6	5 Me	O4	N4	
<b>Glu</b>				<b>E47 345</b> <b>Max 32</b> <b>MyoD 118</b>			<b>Cap 181</b> <b>ER 25</b> <b>USF 208</b> <b>E47 345</b> <b>Max 32</b> <b>MyoD 118</b> <b>NF-KB 63</b> <b>NF-KB 63</b>	
<b>Asp</b>							<b>TTK 154</b> <b>Gli 116</b> <b>Gli 144</b> <b>Gli 144</b>	
<b>His</b>	<b>NF-KB 67</b> <b>Max 28</b>	<b>Prd 47</b> <b>USF 204</b> <b>Pvull 84</b> <b>Zif 49</b>			<b>HNF3 3</b>			
<b>Ser</b>	$\lambda$ 45 <b>Pvull 81</b>	<b>Gli 115</b> <b>Gli 146</b>	<b>Hin 174</b> <b>Gli 147</b>		<b>Prd 46</b>	<b>Mato2 50</b> <b>TTK 124</b>		
<b>Cys</b>		<b>E2 340</b>		<b>E2 340</b>	<b>Oct1hd 50</b> <b>Oct1hd 50</b> <b>Prd 49</b>		<b>p53 277</b>	
<b>Thr</b>			<b>MetJ 25</b> <b>MetJ 25'</b>	<b>PurR 16</b>	<b>MyoD 115</b>	<b>Oct1p 45</b> <b>PurR 16</b> <b>EcoRV 186</b>	<b>Oct1p 45</b>	
<b>Ile</b>					<b>PurR 4</b> <b>PurR 4</b> <b>En 47</b> <b>PPR1 42</b>			
<b>Val</b>					<b>Oct1hd 47</b> <b>GR 462</b>			
<b>Ala</b>					<b>Gli 114</b> <b>GCN4 238</b> <b>GCN4 239</b> <b>fos/jun</b> <b>fos/jun</b>			
<b>Phe</b>					<b>E2 343</b>			
<b>Peptide Backbone</b>								
<b>NH</b>	<b>Hha 257</b>	<b>EcoRV 184</b>				<b>EcoRI 142</b>		
<b>CO</b>							<b>PPR1 40</b> <b>Hha 237</b> <b>PPR1 41</b> <b>Hha 254</b> <b>PPR1 41</b> <b>EcoRV 182</b> <b>Gal4 17</b> <b>EcoRI 138</b> <b>Gal4 18</b> $\lambda$ 45 <b>Gal4 18</b>	

*Arginine:* Arginine is by far the most common amino acid found to interact with the DNA. In the 37 cocrystal structures, 30 arginines from 20 different proteins form a total of 52 hydrogen bonds with DNA bases (Fig. 2a). The arginine side chain is a hydrogen bond donor and, in principle, can interact with the guanine N7 and O6 groups, the adenine N7 group, and the thymine O4 group. In fact, each of the chemically allowable interactions is observed in at least one of the structures. The most prevalent contacts, however, are between arginine and guanine (26 examples) and these contacts are found in the complexes of proteins belonging to several different families of DNA-binding proteins which use different units of secondary structure as binding surfaces. In contrast, only three interactions are seen in the cocrystal structures between arginine and the N7 position of adenine and only one with the O6 of thymine.

A single arginine side chain is frequently observed to form more than one hydrogen bond with the DNA. A particularly favored interaction is one in which the guanidinium group of an arginine side chain donates hydrogen bonds to the N7 and O6 groups of the guanine base, 19 of the arginines tabulated have this conformation (Fig. 2a). This interaction was predicted by Seeman et al. (1976) to play an important role in specific recognition, as no other base can interact with an arginine side chain in this way. When the arginine side chain forms only a single direct hydrogen bond with a base, the remaining hydrogen bond donors of the guanidinium group are frequently involved in a variety of other interactions. For example, in instances in which the side chain donates a hydrogen bond to only the N7 of guanine, additional hydrogen bonds formed include a water-mediated contact to the O6 of the same guanine (ER-DBD and MyoD), a salt bridge with the adjacent

phosphate oxygen (MyoD, E47 and Max), or hydrogen bonds with other side chains at the protein-DNA interface.

*Lysine:* Lysines are also observed frequently at the protein-DNA interface: in the structures examined, 15 lysine side chains form a total of 27 hydrogen bonds (Fig. 2a). As with arginine, the lysine side chain can only donate hydrogen bonds and the vast majority of the contacts seen are with guanines. There are only two lysine-thymine contacts and no contacts were observed with the N7 of adenine. This does not mean that lysine is prohibited from contacting an adenine and could be an artifact of the sample size, but it is intriguing. As with arginine, a lysine side chain often forms several hydrogen bonds at the protein-DNA interface. Lysine donates multiple hydrogen bonds to acceptors on the same base, on successive bases on the same strand, or on successive bases on opposite strands of the DNA. In addition, when only a single direct hydrogen bond is observed between a lysine and a base, additional contacts are formed such as water-mediated hydrogen bonds to bases (GR-DBD and MetJ), hydrogen bonds to other side chains at the protein-DNA interface (ER-DBD), and salt bridges to phosphate oxygens (Prd).

*Asparagine and Glutamine:* Both asparagine and glutamine are commonly used for base contacts. The amide group of the side chains can both donate and accept hydrogen bonds and therefore can interact with each of the bases. In the cocrystal structures 17 asparagines form 30 hydrogen bonds and 10 glutamines form 17 hydrogen bonds (Fig. 2a). The most frequent interactions seen are pairs of hydrogen bonds between the side chain amide groups and the N7 and N6 groups of adenine. These interactions were

also predicted by Seeman et al. (1976) as a means of uniquely specifying an adenine. Again, these amino acids often form more than one hydrogen bond at the protein-DNA interface. In several instances, the amide group of a single side chain interacts with hydrogen bond acceptors and donors on successive base pairs (GCN4, c-fos/c-jun, and E2). The side chain NH<sub>2</sub> or O are also seen to individually donate or accept, respectively, two hydrogen bonds from groups on bases (434R, PvuII, and EcoRI). In addition to forming base contacts, asparagines and glutamines often participate in extensive hydrogen bonding networks at the protein-DNA interface, as seen in the Arc complex.

*Aspartate and Glutamate:* The carboxylate group of the aspartate and glutamate side chains can only accept hydrogen bonds. Therefore DNA contacts by these side chains are limited to the N6 group of adenine and the N4 group of cytosine. In the structures, 3 aspartates and 7 glutamates were found making 4 and 11 hydrogen bonds respectively (Fig. 2b). The majority of these contacts are with the N4 of cytosine. In two complexes, successive cytosines are bridged by a single aspartate or glutamate (NF-κB and Gli). The N6 of adenine and N4 of cytosine are bridged by a glutamate in other structures (E47, Max, and MyoD).

*Histidine:* Six histidines in the cocrystal structures form hydrogen bonds with bases (Fig. 2b). Although histidine is chemically capable of donating and receiving hydrogen bonds, the only interactions observed are of the histidine side chain donating a hydrogen bond to the N7 or O6 groups of guanine.

*Serine, Threonine, Tyrosine and Cysteine:* The hydroxyl and sulfhydryl groups of these side chains can both receive and donate hydrogen bonds and therefore can potentially interact with each base. Of the four amino acids, serine is found most often at the protein-DNA interface (Fig. 2b). Eight serines form 8 hydrogen bonds with bases, all of which are donated by the side chain hydroxyl group to hydrogen bond acceptors on bases. Five threonines are observed, of which 3 donate hydrogen bonds to a base (MetJ and EcoRV) and 2 both donate and accept hydrogen bonds from successive bases (PurR and Oct-1). Two cysteines are found, one receives a hydrogen bond from the N4 of cytosine (p53) and the other donates a hydrogen bond to a guanine O6 and receives a hydrogen bond from the neighboring adenine N6 on the opposite strand (E2). Tyrosine is observed in one structure accepting a hydrogen bond from the N4 of a cytosine (Gli).

*Peptide Backbone:* Hydrogen bonding interactions between the protein and bases of the DNA are not limited solely to interactions from the functional groups of side chains, but can also be formed by the NH and CO groups of the peptide backbone (Fig. 2b). These interactions are highly dependent on the docking of the protein with the DNA as they require a close approach of the polypeptide backbone and the DNA. Hydrogen bonds are donated by the NH group of the peptide backbone to hydrogen bond acceptors of guanine and thymine (HhaI, EcoRV, and EcoRI). Although the carbonyl group of the peptide backbone can accept hydrogen bonds from either the N6 group of adenine or the N4 group of cytosine, contacts are only seen with the N4 of cytosine. In several cases (Gal4, PPR1,  $\lambda$ , and HhaI), both the peptide amide and the side chain functional groups of a single amino acid simultaneously interact with bases.

*Hydrophobic Interactions:* Hydrophobic interactions also play a role in site-specific recognition and primarily involve the 5-methyl group of thymine (occasional contacts have been reported to the 5-C of cytosine). van der Waals interactions are observed between the thymine methyl group and the hydrophobic side chains of isoleucine, alanine, valine, and phenylalanine, the methyl group of threonine, and the side chains of serine, cysteine, histidine, and glutamine (Fig. 2a, b). In addition, many authors describe hydrophobic patches surrounding thymine methyl groups at protein-DNA interfaces. These patches are formed by many of the amino acids listed above as well as the aliphatic portions of lysine, arginine, asparagine, methionine, leucine, and glycine.

***Direct Read-Out in the Minor Groove:***

Minor-groove recognition differs from major-groove recognition in that the minor groove of B-form DNA is narrower and deeper than the major groove, so the bases are less accessible, and there are fewer ways to distinguish among base pairs based on the patterns of hydrogen bond donors and acceptors (Seeman et al., 1976). Contacts are observed between the protein and bases in the minor groove in only fourteen of the cocrystal structures. The following discussion focuses on contacts formed in the minor groove of binding site in which the DNA retains a general B-form character. Contacts formed in the PurR and TBP structures will not be included here since the structure of the minor groove is so deformed that the contacts are not readily comparable to contacts seen with the less accessible minor groove of B-form DNA.

*Arginine:* As seen for major-groove contacts, arginine is also the amino acid found to interact most often with bases in the minor groove. Arginine can donate hydrogen bonds to the N3 of adenine and the O2 of thymine in the minor groove. In the cocrystal structures of 434 repressor, 434 cro, HNF-3, and p53, an arginine from a loop inserts into the minor groove at an A:T rich region and the guanidinium group of the side chain packs against the sugar phosphate backbone forming direct and water-mediated contacts with bases. The minor groove is compressed in these regions and it has been suggested that the arginine reduces the repulsion between phosphates across the minor groove (Shimon & Harrison, 1993). In the engrailed, Mat $\alpha$ 2, Oct-1, and Hin recombinase structures, arginine side chains from the N-terminal arm project into the minor groove and form hydrogen bonds with the O2 and/or N3 groups of thymine and adenine bases. In the structure of the paired domain complex, an arginine side chain from the  $\beta$ -turn also forms a water-mediated hydrogen bond to a thymine O2 group.

*Other amino acids:* Asparagine and aspartate are the only other amino acids which contact bases in the minor groove in the cocrystal structures. An asparagine from the  $\beta$ -turn of paired accepts hydrogen bonds from the N2 of a guanine base. An asparagine from EcoRV forms a direct hydrogen bond and an aspartate from PvuII forms a water-mediated contact with cytosine O2 groups in the minor grooves of their respective binding sites.

*Peptide Backbone Amide:* Both the NH and CO groups of the peptide backbone are observed to interact with bases in the minor groove. The N-terminal arm of Hin recombinase approaches the minor groove closely, and both the side chain and the backbone NH group of an arginine donate

hydrogen bonds to the N3 of adenine and O2 of thymine, respectively. In the Paired complex, a peptide carbonyl from the  $\beta$ -turn accepts a hydrogen bond from the N2 group of a guanine base. In addition, the C-terminal tails of the paired domain and Hin recombinase run along the minor groove of the DNA with polypeptide backbone forming hydrogen bonds with bases while the side chains project out of the groove. The minor groove interactions by the tail of Hin recombinase resemble those of the minor groove binding drugs netropsin and Hoescht 33258 (Kopka et al., 1985; Pjura et al., 1987). A similar type of minor groove interaction has been proposed for the SPKK motifs of sea urchin spermatogenesis histones, H1 and H2B (Suzuki, 1989).

#### *General Principles:*

Even with a larger database of cocrystal structures, no simple recognition code for protein-DNA interactions emerges that can be used to predict a binding site for a given protein or to design a protein to bind to a particular site. As shown in Figure 2, nearly every allowable hydrogen bonding interaction between amino acids and functional groups on the edges of bases is seen in at least one of the cocrystal structures. However, despite the wide variety in the type of contacts seen at the protein-DNA interface, certain interactions between amino acids and bases seem to be preferred. These include: arginine and lysine with guanine, asparagine and glutamine with adenine, and glutamate, aspartate and the peptide backbone carbonyl with cytosine. This may indicate that these interactions are particularly favorable. For example, lysines and arginines interact primarily with guanines and rarely with other bases, which is probably due to the stronger electronegativity of a guanine base.

Another interesting observation is that amino acids in many of the structures form multiple hydrogen bonds with a single base, with successive bases on either strand of the DNA, with a base and a phosphate oxygen, or with a base and other amino acids at the protein-DNA interface. Seeman et al. (1976) proposed that the formation of two hydrogen bonds to the same base could be used to discriminate effectively among base pairs and would provide more specificity than the formation of a single hydrogen bond to a base. They predicted that arginine could form such an interaction with guanine by donating hydrogen bonds to both the N7 and O6 positions and that glutamine or asparagine could form such an interaction with adenine by donating a hydrogen bond to the N7 position and receiving one from the N6 position. All three of these interactions are seen quite often in the cocrystal structures (Fig. 2a). In a similar manner, hydrogen bonds from a single amino acid to successive bases can be used to discriminate among pairs of bases. These interactions are also seen quite frequently in the cocrystal structures. Additionally in many of the complexes, amino acids which form hydrogen bonds with bases also form hydrogen bonds with the phosphate backbone or with other amino acids at the protein-DNA interface. These networks of hydrogen bonds appear to help to position the side chains precisely on the DNA and may enhance the overall specificity and affinity of binding.

Far less diversity is seen in protein-minor groove interactions than in protein-major groove interactions and this may be due to the relatively small number of such interactions. However, arginine seems to be particularly well suited to contact A:T base pairs in the minor groove. The arginine side chain is long and flexible and can project into the minor groove to interact with A:T base pairs without significant distortion of the DNA. Little is known about

the contribution of minor-groove interactions to protein-DNA recognition. In several cases the minor-groove interactions have been shown to be crucial for DNA-binding activity (Chapter 3, Xu et al., 1995). However, the modeling studies of Seeman et al. (1976) suggest that there are fewer ways to distinguish among different base pairs in the minor groove and therefore such interactions may be less specific. These issues are addressed directly in Chapter 3.

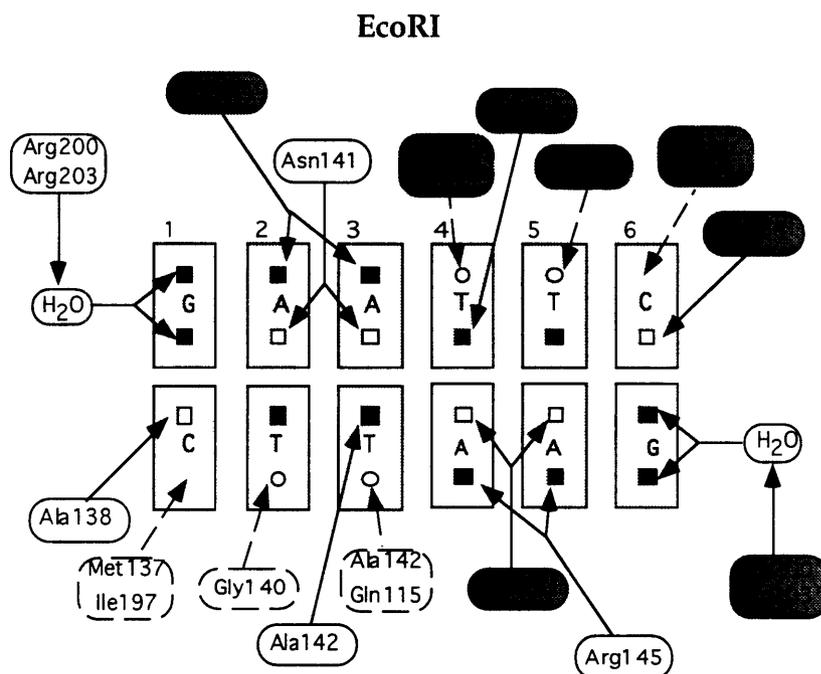
### *Thermodynamic Studies*

The examination of the known cocrystal structures reveals the variety of ways in which proteins bind to DNA and raises several questions about how specificity is achieved. There is considerable degeneracy in the recognition code, *i. e.* one amino acid can interact with several bases. Is this degeneracy reflected in a lower binding specificity, for example can an serine interact equally well with the N7 of guanine and the N7 of adenine? Is specificity increased when an amino acid side chain forms multiple hydrogen bonds with the DNA? To what extent do the extensive hydrogen bonding networks seen in some of the cocrystal structures among amino acid side chains, bases of the binding site, and the phosphodiester backbone increase specificity? How do major-groove contacts compare to minor-groove contacts in terms of binding specificity and energy? Are contacts from a flexible region of secondary structure less specific than those from an explicit binding surface? Systematic biochemical studies addressing the importance of interactions seen in a cocrystal structure can provide answers to such questions. For many of the proteins discussed in this chapter, only genetic or crude biochemical experiments have been used to address the importance of

an amino acid or base pair in the binding site (Neuberg et al., 1989; He et al., 1992; Hughes et al., 1992; Freeman et al., 1994). While such experiments have provided valuable qualitative information, there are fewer studies which provide quantitative information through systematic analyses of equilibrium reactions. The last section of this chapter will review such data for three protein-DNA complexes, those of EcoRI, Arc repressor, and  $\lambda$  repressor. Biochemical studies of the DNA-binding activity of the engrailed homeodomain are presented in Chapters 2 and 3.

#### *EcoRI:*

The type II restriction endonuclease, EcoRI, binds to its recognition site, GAATTC, as a dimer and forms symmetric contacts with each half-site. Contacts with the DNA are formed by residues from two  $\alpha$ -helices and a region of extended chain which lies along the pyrimidine-rich strand of the DNA. The EcoRI complex is unusual in that nearly every functional group of the bases in the major groove is involved in interactions with the protein, as shown in Figure 3 below. The complex also provides an example of the use of a single amino acid to contact two base pairs in the binding site, as was discussed above. Both Arg145 and Asn141 form two hydrogen bonds with the N7 and N6 groups respectively of successive adenines (Fig. 3).



**Figure 3: Base contacts in the EcoRI structure.** Hydrogen bond acceptors on bases are indicated by filled boxes, donors are indicated by open boxes, and thymine methyl groups are indicated by open circles. Solid lines refer to hydrogen bonds and dashed lines refer to hydrophobic interactions. Bases which are contacted by the protein are lightly shaded. Residues from the second monomer are noted by darkly shaded boxes. Arg200, Arg203, and Ile47 are from the outer  $\alpha$ -helix. Arg145 is from the inner  $\alpha$ -helix. Met137, Ala138, Gly140, and Ala142 are from the extended chain.

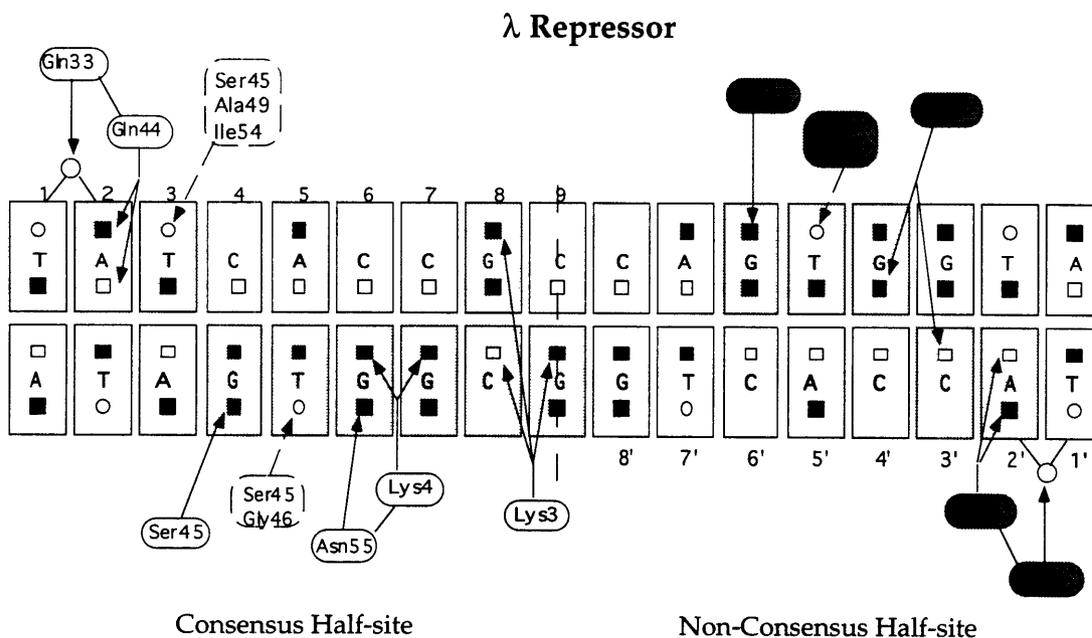
Lesser et al. (1990, 1993) carried out an exhaustive study on the effects of substitutions of natural bases and base analogues on both the free energy of the association reaction and the transition state free energy of the cleavage reaction. The discussion here will focus on the free energy of binding to mutant binding sites, *i. e.* the ability of EcoRI to discriminate among binding sites in the association step. Single base substitutions at each position in the recognition site resulted in a 4-6 kcal/mol loss in the free energy of binding. The double substitutions tested and a complete reversal of the recognition site also resulted in a 4-6 kcal/mol loss in the free energy of binding indicating

significant non-additivity. Lesser et al. (1990) make several points which are important to consider when evaluating such results in terms of specificity. First, base substitutions can perturb several aspects of a protein-DNA complex by not only removing a particular contact but also introducing unfavorable interactions. Second, a base substitution may change the conformational flexibility of the DNA. The binding site in the EcoRI complex is kinked and the effect of base substitutions on binding affinity will reflect the energetic cost of kinking the DNA in the altered site. Finally, base substitutions may also perturb interactions of EcoRI with phosphate groups of the DNA backbone, and the authors did observe that certain interactions with phosphate groups were altered in the substituted sites. As a result, the effect of substitutions using natural bases reflects the overall specificity at a position in the binding site which derives from both direct and indirect read-out.

In contrast, the effects of substitutions with isosteric base analogues, which remove a functional group on the bases, mainly reflect the direct read-out component of recognition, *i. e.* the contribution of a specific interaction to binding affinity. The EcoRI phosphate contacts were unaltered in the sites with isosteric base substitutions and in general these substitutions did not appear to alter the structure of the DNA. All but one of the interactions probed by the base analogues (hydrogen bonds by charged and uncharged side chains, water-mediated hydrogen bonds, and hydrophobic interactions with thymine methyl groups) were found to be of comparable energy, 1-2 kcal/mol. The one striking exception is the substitution of A3 with purine which actually increased the binding affinity, presumably by removing a steric hindrance which is created when the DNA is kinked (Lesser et al., 1993).

### *Lambda Repressor:*

$\lambda$  repressor binds to DNA as a dimer and the structure of the N-terminal domain bound to the operator site,  $O_{L1}$ , has been solved. Specific base contacts are formed by residues from the N-terminal arm (Lys3, Lys4), the recognition helix (Gln44, Ser45), and the turn following the recognition helix (Asn55) as outlined in Figure 4. The  $O_{L1}$  site is comprised of two asymmetric half-sites: a consensus half-site, the sequence of which is conserved among the  $\lambda$  operator sites, and a nonconsensus half-site, which differs at certain positions from the consensus half-site. The  $\lambda$  repressor complex is particularly interesting because the N-terminal arm is ordered only in the consensus half-site of the operator and only a subset of the contacts in between the half-sites are symmetric.



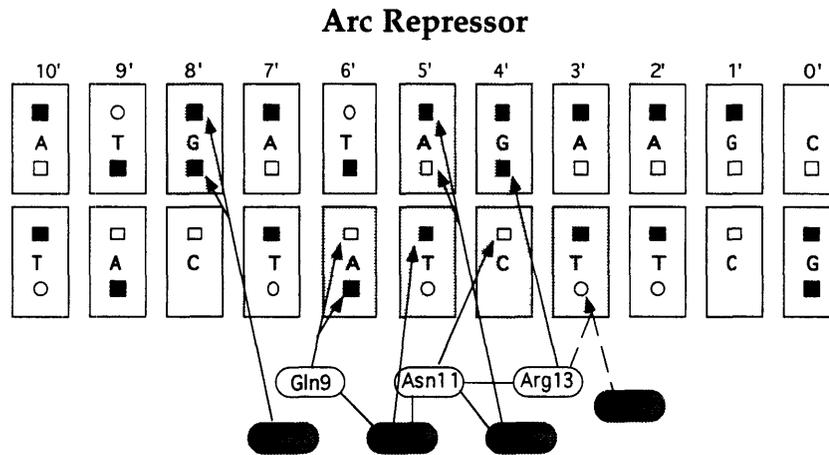
**Figure 4: Base contacts in the  $\lambda$  repressor structure.** Hydrogen bond acceptors on bases are indicated by filled boxes, donors are indicated by open boxes, and thymine methyl groups are indicated by open circles. Solid lines refer to hydrogen bonds and dashed lines refer to hydrophobic interactions. Bases which are contacted by the protein are lightly shaded. Residues from the second monomer are noted by darkly shaded boxes.

Sarai and Takeda (1989) examined the effects of base pair substitutions in  $O_{R1}$ .  $O_{R1}$  differs from  $O_{L1}$ , the operator site used in cocrystal structure, by a single T:A  $\rightarrow$  A:T change at position 5 in the nonconsensus half-site. Substitutions at each base pair involved in hydrogen bonds with the protein decreased binding affinity, with changes at base-pairs 6 through 9 on the consensus half-site causing the largest effects. Base pairs at these positions form hydrogen bonds with Lys3 and Lys4 of the N-terminal arm and Asn55 of the loop following the recognition helix. These results show that the arm can contribute interactions which are highly specific despite its flexibility. On the nonconsensus half-site at the same positions, only the transversion mutations at base-pair 6 and the base-pair reversal at position 7 cause large decreases in affinity supporting the observation from the cocrystal structure that the arm of the subunit bound at the nonconsensus half-site is disordered and does not interact with the DNA. Functional asymmetry is also seen at base-pairs 2 and 2' which are contacted by Gln44 of the recognition helix. Substitutions at base-pair 2 of the consensus half-site decrease binding more than those at base-pair 2' of the nonconsensus half-site. Finally, substitutions at base-pair 4 of both half-sites reduce binding affinity, but to a lesser extent than at other positions of major-groove contacts. Therefore among the positions which are involved in hydrogen bonds with the protein, there is a range of tolerance to base substitutions, *i. e.* some positions are more specific than others. This contrasts with the results from EcoRI in which each base had a relatively equal contribution to binding specificity and may reflect the fact that in the EcoRI complex each base is involved in multiple interactions with the protein.

Sarai and Takeda (1989) also addressed the role of hydrophobic interactions with thymine methyl groups in the complex. At two positions on the consensus half-site, hydrophobic pockets are formed by the protein around thymine methyl groups. Removal of the thymine methyl group by substituting the A:T base pair at position 5 in the consensus half-site with A:U, results in a large decrease in binding affinity suggesting that the hydrophobic interactions at this position contribute to binding specificity and affinity. In fact, substitutions at this position have as large an effect on the binding energy as substitutions at base pairs which form hydrogen bonds with the protein. In contrast, natural base substitutions at base-pair 3 of the consensus half-site have little effect on binding affinity and substitution of thymine with uracil actually increases binding affinity. Thus, the hydrophobic interactions observed in the structure at base-pair 3 do not contribute to specificity or affinity and may even cause slight steric hindrance upon binding.

*P22 Arc Repressor:*

Dimers of the Arc repressor protein bind cooperatively to each half-site of the operator yielding a DNA-bound tetramer (Brown & Sauer, 1993). Base contacts are formed on each half-site by amino acids from a two stranded  $\beta$ -sheet (Figure 5). In addition to direct contacts with bases of the operator, there is an extensive hydrogen bonding network at the protein-DNA interface involving side chains that contact bases and phosphate groups. The N-terminal arm of the protein, which is disordered in the absence of DNA, becomes structured upon DNA binding and amino acids from the arm contribute phosphate contacts and hydrophobic interactions.



**Figure 5: Base contacts in the Arc repressor structure on the right half-site.** Hydrogen bond acceptors on bases are indicated by filled boxes, donors are indicated by open boxes, and thymine methyl groups are indicated by open circles. Solid lines refer to hydrogen bonds and dashed lines refer to hydrophobic interactions. Bases which are contacted by the protein are lightly shaded. Residues from the second monomer are noted by darkly shaded boxes. Contacts with the left half-site are the same with the exception of Arg13' which donates a hydrogen bond to the N7 of adenine at base pair 8.

Brown et al. (1994) used alanine scanning mutagenesis to assess the importance of amino acids in the protein to DNA-binding affinity. As expected, each amino acid involved in a direct contact with the DNA in the cocrystal structure makes a significant contribution to both half-site binding affinity and whole-site binding affinity. Among the side chains which contact the DNA, there is a range of effects with Arg13 making the largest contribution. A particularly interesting class of mutants were those that do not make direct base contacts but instead link different parts of the protein which contact the DNA. For example, the side chain of Asn34 from helix B forms hydrogen bonds with the peptide amide of Arg13 from the  $\beta$ -sheet, while the peptide NH of Asn34 forms a hydrogen bond with a phosphate oxygen. When this network, which connects the major-groove interactions of the  $\beta$ -sheet with the phosphate contacts mediated by helix B, is disrupted,

binding affinity is decreased to the same extent as deleting a side chain which forms direct base contacts.

Residues from the N-terminal arm of Arc, which is disordered in the absence of DNA, are also important for binding affinity. Residues from the arm form hydrophobic interactions with the DNA, phosphate contacts, and linkage contacts. These amino acids all make significant contributions to the overall binding affinity despite the entropic cost of ordering the arm.

### *Conclusions*

Several conclusions can be drawn about the origins of site-specific recognition from the discussion of specificity in protein-DNA interactions presented here.

1) Most hydrogen bonding interactions between the protein and the DNA inferred from the cocrystal structures are important biochemically. However, the actual contributions to binding specificity and affinity of the contacts vary and there is no obvious correlation between the type of interaction and its contribution.

2) The role of hydrophobic interactions in protein-DNA recognition is context dependent. In the EcoRI complex and at certain positions in the  $\lambda$  repressor complex, hydrophobic interactions appear to contribute to recognition. However, at other positions in the  $\lambda$  repressor complex, where hydrophobic interactions are observed in the structure, these interactions are not important.

3) When a base pair in the binding site is involved in more than one interaction with the protein, the specificity at that position is enhanced. This is demonstrated in the EcoRI complex where nearly every functional group on each base pair in the binding site is contacted by the protein. The specificity at each position of the binding site (as defined by the effect of base substitutions on affinity) is high. In addition, in the  $\lambda$  repressor complex, multiple contacts are made by the protein at base-pairs 6 through 9 and these positions are among the most sensitive to base substitutions.

4) Hydrogen bonding networks among amino acids at the protein-DNA interface can make a significant contribution to binding affinity. As was demonstrated with Arc repressor, contacts of the linkage class, which connect interactions with bases in the major groove to other regions of the protein, can contribute as much to binding energy as a direct base contact.

6) Contacts to the DNA from regions of polypeptide chain which are flexible and disordered in the absence of DNA can make a significant contribution to DNA recognition, despite the entropic cost paid by ordering the polypeptide chain upon DNA binding. The N-terminal arms of both  $\lambda$  repressor and Arc repressor are necessary for high affinity DNA binding and, in the case of  $\lambda$  repressor, form specific interactions with the DNA.

The studies in the following two chapters provide additional insights into the biochemical basis of specific recognition by examining the contributions of bases in the binding site and amino acids in the protein to the binding specificity and affinity of the engrailed homeodomain and of an

altered-specificity mutant (Gln50 → Lys) of the engrailed homeodomain. The structures of both proteins bound to their optimal binding sites have been determined (Kissinger et al., 1990; Tucker-Kellogg et al., 1995). The proteins bind to DNA as monomers and residues from the third  $\alpha$ -helix interact with bases in the major groove and residues from the N-terminal arm, which is disordered in the free protein, interact with bases in the minor groove. In contrast to the structures discussed above, there are no hydrogen bonding networks between side chains at the protein-DNA interface and each base pair in the binding site is contacted by only one amino acid.

The work presented in Chapter two focuses on the contribution of a single amino acid in the homeodomain, the residue at position 50, to binding specificity. By changing the amino acid at position 50 of the engrailed homeodomain, the binding specificity of the homeodomain can be altered (Ades & Sauer, 1994). The work presented in Chapter three addresses the contributions to binding specificity and affinity of interactions between bases in the minor groove with residues of the N-terminal arm and interactions between bases in the major groove with residues of  $\alpha$ -helix three. The arm of engrailed, like that of Arc repressor and  $\lambda$  repressor, has a significant contribution to binding affinity. However, unlike the N-terminal arm of  $\lambda$  repressor which contributes as much to specificity as the recognition helix, interactions by residues of the arm in engrailed have a lower specificity than interactions formed by residues from the third  $\alpha$ -helix (Ades & Sauer, submitted).

## References

- Ades, S. E., & Sauer, R. T. (1994) *Biochemistry* 33, 9187-9194.
- Aggarwal, A. K., Rodgers, D. W., Drottar, M., Ptashne, M., & Harrison, S. C. (1988) *Science* 242, 899-907.
- Beamer, L. J., & Pabo, C. O. (1992) *J. Mol. Biol.* 227, 177-196.
- Bell, A. C., & Koudelka, G. B. (1993) *J. Mol. Biol.* 234, 542-553.
- Bell, A. C., & Koudelka, G. B. (1995) *J. Biol. Chem.* 270, 1205-1212.
- Brown, B. M., & Sauer, R. T. (1993) *Biochemistry* 32, 1354-1363.
- Brown, B. M., Milla, M. E., Smith, T. L., & Sauer, R. T. (1994) *Nature Struct. Biol.* 1, 164-168.
- Cheng, X., Balendiran, K., Schildkraut, I., & Anderson, J. E. (1994) *Embo J.* 13, 3927-3935.
- Cho, Y., Gorina, S., Jeffrey, P. D., & Pavletich, N. P. (1994) *Science* 265, 346-355.
- Clark, K. L., Halay, E. D., Lai, E., & Burley, S. K. (1993) *Nature* 364, 412-420.
- Clarke, N. D., Beamer, L. J., Goldberg, H. R., Berkower, C., & Pabo, C. O. (1991) *Science* 254, 267-270.
- Ellenberger, T., Fass, D., Arnaud, M., & Harrison, S. C. (1994) *Genes Dev.* 8, 970-980.
- Ellenberger, T. E., Brandl, C. J., Struhl, K., & Harrison, S. C. (1992) *Cell* 71, 1223-1237.
- Fairall, L., Schwabe, J. W., Chapman, L., Finch, J. T., & Rhodes, D. (1993) *Nature* 366, 483-487.
- Feng, J. A., Johnson, R. C., & Dickerson, R. E. (1994) *Science* 263, 348-355.
- Ferré-D'Amaré, D. A., Pognonec, P., Roeder, R. G., & Burley, S. K. (1994) *Embo J.* 13, 180-189.
- Ferré-D'Amaré, D. A., Prendergast, G. C., Ziff, E. B., & Burley, S. K. (1993) *Nature* 363, 38-45.

- Freeman, J., Schmidt, A., Scharer, E., & Iggo, R. (1994) *EMBO J.* 13, 5393-5400.
- Ghosh, G., van, D. G., Ghosh, S., & Sigler, P. B. (1995) *Nature* 373, 303-310.
- Glover, J. N., & Harrison, S. C. (1995) *Nature* 373, 257-261.
- He, Y., McNally, T., Manfield, I., Navratil, O., Old, I. G., Phillips, S. E. V., Saint-Girons, I., & Stockley, P. G. (1992) *Nature* 359, 431-433.
- Hegde, R. S., Grossman, S. R., Laimins, L. A., & Sigler, P. B. (1992) *Nature* 359, 505-512.
- Hughes, K. T., Gaines, P. C. W., Karlinsey, J. E., Vinayak, R., & Simon, M. I. (1992) *EMBO J.* 11, 2695-2705.
- Jordan, S. R., & Pabo, C. O. (1988) *Science* 242, 893-899.
- Kim, J. L., Nikolov, D. B., & Burley, S. K. (1993a) *Nature* 365, 520-527.
- Kim, Y., Geiger, J. H., Hahn, S., & Sigler, P. B. (1993b) *Nature* 365, 512-520.
- Kim, Y. C., Grable, J. C., Love, R., Greene, P. J., & Rosenberg, J. M. (1990) *Science* 249, 1307-1309.
- Kissinger, C. R., Liu, B. S., Martin, B. E., Kornberg, T. B., & Pabo, C. O. (1990) *Cell* 63, 579-590.
- Klemm, J. D., Rould, M. A., Aurora, R., Herr, W., & Pabo, C. O. (1994) *Cell* 77, 21-32.
- Klimasauskas, S., Kumar, S., Roberts, R. J., & Cheng, X. (1994) *Cell* 76, 357-369.
- Konig, P., & Richmond, T. J. (1993) *J. Mol. Biol.* 233, 139-154.
- Kopka, M. L., Yoon, C., Goodsell, D., Pjura, P., & Dickerson, R. E. (1985) *J. Mol. Biol.* 183, 553-563.
- Lawson, C. L., & Carey, J. (1993) *Nature* 366, 178-182.
- Lesser, D. R., Kurpiewski, M. R., & Jen-Jacobson, L. (1990) *Science* 250, 776-86.
- Lesser, D. R., Kurpiewski, M. R., Waters, T., Connolly, B. A., & Jen-Jacobson, L. (1993) *Proc. Natl. Acad. Sci. U. S. A.* 90, 7548-7552.

- Luisi, B. F., Xu, W. X., Otwinowski, Z., Freedman, L. P., Yamamoto, K. R., & Sigler, P. B. (1991) *Nature* 352, 497-505.
- Ma, P. C., Rould, M. A., Weintraub, H., & Pabo, C. O. (1994) *Cell* 77, 451-459.
- Marmorstein, R., Carey, M., Ptashne, M., & Harrison, S. C. (1992) *Nature* 356, 408-414.
- Marmorstein, R., & Harrison, S. C. (1994) *Genes Dev.* 8, 2504-2512.
- McClarín, J. A., Frederick, C. A., Wang, B. C., Greene, P., Boyer, H. W., Grable, J., & Rosenberg, J. M. (1986) *Science* 234, 1526-1541.
- Mondragon, A., & Harrison, S. C. (1991) *J. Mol. Biol.* 219, 321-334.
- Muller, C. W., Rey, F. A., Sodeoka, M., Verdine, G. L., & Harrison, S. C. (1995) *Nature* 373, 311-317.
- Neuberg, M., Schuermann, M., Hunter, J. B., & Müller, R. (1989) *Nature* 338, 589-590.
- Otwinowski, Z., Schevitz, R. W., Zhang, R. G., Lawson, C. L., Joachimiak, A., Marmorstein, R. Q., Luisi, B. F., & Sigler, P. B. (1988) *Nature* 335, 321-329.
- Pabo, C. O., & Sauer, R. T. (1992) *Annu. Rev. Biochem.* 61, 1053-1095.
- Pavletich, N. P., & Pabo, C. O. (1991) *Science* 252, 809-817.
- Pavletich, N. P., & Pabo, C. O. (1993) *Science* 261, 1701-1707.
- Pjura, P. E., Grzeskowiak, K., & Dickerson, R. E. (1987) *J. Mol. Biol.* 197, 257-271.
- Raumann, B. E., Rould, M. A., Pabo, C. O., & Sauer, R. T. (1994) *Nature* 367, 754-757.
- Raumann, B. E., Brown, B. M., & Sauer, R. T. (1994) *Curr. Opin. Struct. Biol.* 4, 36-43.
- Rodgers, D. W., & Harrison, S. C. (1993) *Structure* 1, 227-240.
- Rosenberg, J. M. (1991) *Curr. Opin. Struct. Biol.* 1, 104-113.
- Sarai, A., & Takeda, Y. (1989) *Proc. Natl. Acad. Sci. U. S. A.* 86, 6513-6517.
- Schultz, S. C., Shields, G. C., & Steitz, T. A. (1991) *Science* 253, 1001-1007.

- Schumacher, M. A., Choi, K. Y., Zalkin, H., & Brennan, R. G. (1994) *Science* 266, 763-770.
- Schwabe, J. W. R., Chapman, L., Finch, J. T., & Rhodes, D. (1993) *Cell* 75, 567-578.
- Schwabe, J. W. R., Chapman, L., & Rhodes, D. (1995) *Structure* 3, 201-213.
- Seeman, N. C., Rosenberg, J. M., & Rich, A. (1976) *Proc. Natl. Acad. Sci. U. S. A.* 73, 804-808.
- Shimon, L. J., & Harrison, S. C. (1993) *J. Mol. Biol.* 232, 826-838.
- Somers, W. S., & Phillips, S. E. (1992) *Nature* 359, 387-393.
- Suzuki, M. (1989) *Embo J.* 8, 797-804.
- Tucker-Kellogg, L., Rould, M. A., Chambers, K. A., Ades, S. E., Sauer, R. T., & Pabo, C. O. (1995) *Manuscript in preparation.*
- Weiss, M. A., Ellenberger, T., Wobbe, C. R., Lee, J. P., Harrison, S. C., & Struhl, K. (1990) *Nature* 347, 575-578.
- Winkler, F. K., Banner, D. W., Oefner, C., Tsernoglou, D., Brown, R. S., Heathman, S. P., Bryan, R. K., Martin, P. D., Petratos, K., & Wilson, K. S. (1993) *Embo J.* 12, 1781-1795.
- Wolberger, C., Vershon, A. K., Liu, B., Johnson, A. D., & Pabo, C. O. (1991) *Cell* 67, 517-528.
- Xu, W., Rould, M. A., Jun, S., Desplan, C., & Pabo, C. O. (1995) *Cell* 80, 639-650.

## **Chapter 2**

### **Differential DNA-Binding Specificity of the Engrailed Homeodomain: the Role of Residue 50**

## *Introduction*

Understanding the determinants of binding specificity is one of the central challenges in the study of protein-DNA interactions. The homeodomain, a sixty residue DNA binding motif, provides an attractive system in which to study this problem because different homeodomains are structurally similar and bind to DNA in a similar manner, but often have distinct DNA-binding specificities. The cocrystal structure of the homeodomain from the *Drosophila* transcription factor engrailed is known (Kissinger et al., 1990), and provides a basis for understanding homeodomain-DNA recognition. Many homeodomains, including engrailed, bind to DNA sites containing the core sequence TAAT (Laughon, 1991). In the engrailed cocrystal structure, these bases are contacted by Arg3 and Arg5 in the N-terminal arm and by Ile47 and Asn51 in the third  $\alpha$ -helix (Figure 1). Not surprisingly, the identity or general chemical character of these four amino acids are conserved in homeodomains that bind to sites containing TAAT, suggesting that recognition of this core sequence occurs in a similar manner in these homeodomains (Laughon, 1991).

The base pairs following the TAAT core sequence differ in the binding sites of many homeodomains, and interactions between the side chain at position 50 (the ninth residue in  $\alpha$ -helix 3) and these bases appear to play a key role in determining differential DNA-binding specificity (Hanes & Brent, 1989; Treisman et al., 1989; Percival-Smith et al., 1990; Hanes & Brent, 1991). Thus, when the lysine at position 50 of the *Drosophila* bicoid homeodomain is replaced with a glutamine as found in antennapedia class of homeodomains, the mutant bicoid homeodomain now recognizes the antennapedia class binding site TAATTG rather than the bicoid binding site TAATCC (Hanes & Brent, 1991). In the

engrailed cocrystal structure, Gln50 projects into the major groove and forms a van der Waals interaction with the thymine methyl group of the final A:T base pair of the binding site TAATTA (Figure 1; Kissinger et al., 1990). Since van der Waals contacts are not generally thought to be critical determinants of binding specificity, this result raised a number of questions. Does engrailed discriminate among binding sites in the same manner as other homeodomains using Gln50 as the prime determinant of differential specificity? If Gln50 is important, is the contact seen in the cocrystal structure relevant or does crystal packing prevent or distort another contact? Is the TAATTA site to which the engrailed homeodomain is bound in the cocrystal structure a high affinity binding site? This last question arises because the natural binding site for engrailed is not known, and the protein was co-crystallized with a DNA fragment containing a binding site for another homeodomain which by chance also contained a TAATTA site (Kissinger et al., 1990). To address these questions, we have examined both the DNA binding site preferences and the energetics of binding for the wild type engrailed homeodomain and for variants with lysine or alanine at position 50.

### ***Materials and Methods:***

*Oligonucleotides:* The oligodeoxyribonucleotides used for these studies were synthesized on an Applied Biosystems model 381A DNA synthesizer and are listed in Figure 2. Double stranded DNA fragments used for equilibrium binding studies were purified by chromatography on a Pharmacia MonoQ anion exchange column. All other oligonucleotides were gel purified as needed.

*Construction of the synthetic gene:* A gene encoding the sixty amino acid homeodomain from the *Drosophila* engrailed protein (see Figure 2a) was

constructed by ligating four double-stranded oligonucleotide cassettes. Several unique restriction sites were incorporated in the coding sequence and a methionine was added to allow expression in *Escherichia coli*. The gene was cloned between the NdeI and ClaI sites of the T7 expression phagemid pAED4 to create the plasmid pSEA100. pAED4 (a gift from Don Doering) contains the pUC19 backbone and f1 intergenic region, and the T7 polymerase promoter, ribosome binding site, and transcription termination sequences derived from pET3a (Studier et al., 1990). Genes encoding the mutant engrailed homeodomains, QK50 and QA50, were constructed by cloning the appropriate synthetic oligonucleotides between the BglII and BssHIII sites of pSEA100. The sequences of the synthetic gene and both variants were verified by dideoxy sequencing (Sanger et al., 1977).

*Expression and purification of proteins:* The wild-type and mutant engrailed homeodomains were purified from *E. coli* strains BL21(DE3)/pLysS/pSEA100 and X90(DE3)/pSEA100, respectively. Cells were grown with aeration at 37 °C in 1 liter of LB broth plus 150 µg/ml of ampicillin to an OD<sub>600</sub> of 0.7-1.0, and transcription from the T7 promoter was induced by the addition of IPTG to 0.4 mM. After three hours, cells were harvested by centrifugation and resuspended in 5 volumes of lysis buffer (100 mM Tris-HCl [pH 8.0], 200 mM KCl, 1 mM EDTA, 2 mM CaCl<sub>2</sub>, 10 mM MgCl<sub>2</sub>, 2 mM NaN<sub>3</sub>, 1.4 mM beta-mercaptoethanol, and 50% glycerol). 10 µl of phenylmethylsulfonylfluoride (100 mM in ethanol) was added per liter of cell culture. The purification was monitored at each step by electrophoresing samples on Tris-tricine polyacrylamide gels (Schagger & von Jagow, 1987) followed by staining with Coomassie blue. Cells were lysed by sonication and the nucleic acids were precipitated with 0.5% polyethyleneimine. After centrifugation, proteins in the supernatant were precipitated by the

addition of solid ammonium sulfate to 95% saturation. The ammonium sulfate pellet was collected by centrifugation and resuspended in column buffer (25 mM Tris-HCl [pH 7.5], 0.1 mM EDTA, and 1.4 mM 2-mercaptoethanol) containing 100 mM NaCl. Following extensive dialysis against the same buffer, the material was loaded onto an 8 ml DEAE Sephacel column (Pharmacia) and the flow-through fraction and the first column volume of wash were collected. These fractions were combined and loaded onto a 12 ml CM-Sephadex C-50 column (Sigma) which was eluted with steps of column buffer with increasing concentrations of NaCl. The fractions containing the engrailed homeodomain (400 - 500 mM NaCl) were concentrated by ultrafiltration and loaded onto a C<sub>18</sub> reverse phase column which was eluted with a gradient from 35% reverse phase buffer A (0.1% trifluoroacetic acid (TFA) in HPLC grade water) to 45% reverse phase buffer B (0.1% TFA, 80% acetonitrile in HPLC grade water). The fractions containing the pure engrailed homeodomain were pooled and lyophilized. For storage, the protein was resuspended in column buffer with 100 mM NaCl.

Protein concentrations were determined using an extinction coefficient at 280 nm of  $6758 \text{ M}^{-1} \text{ cm}^{-1}$ . The sequence of the first seven amino acids of the purified wild-type engrailed homeodomain was determined by sequential Edman degradation using an Applied Biosystems Model 477A Protein Sequencer with on-line Model 120 PTH Amino Acid Analyzer. Protein sequencing and determination of the amino acid composition were performed by the MIT Biopolymers Laboratory. Circular dichroism was used to monitor the folding of the wild-type and mutant engrailed homeodomains. All experiments were performed using an AVIV 60DS spectropolarimeter fitted with a Hewlett-Packard temperature controller. Spectra from 200 and 300 nm were collected at 20 °C in 1 nm steps with an averaging time of 1 s and averaged over 5 repeats.

Samples contained 25 µg/ml (wild type and QK50) or 18 µg/ml (QA50) of protein in 50 mM potassium phosphate [pH 7.0], 100 mM KCl. Protein stability was assessed by following the ellipticity at 222 nm as a function of temperature. Ellipticity was measured at 1 °C intervals from 15 to 90 °C with an equilibration time of 1 min and a 30 s averaging time. Thermal denaturation data for two-state denaturation were fit by a nonlinear least squares procedure using a Macintosh version of the program NonLin.

*Equilibrium and Kinetic Assays of DNA Binding: Binding site*

oligonucleotides for mobility shift assays (Figure 2b) were 5'-end labelled with  $\gamma$ -<sup>32</sup>P-ATP and T4 polynucleotide kinase using standard protocols (Sambrook et al., 1989). After one strand had been end-labelled, the complementary oligonucleotide was added, the mixture was heated to 90 °C, and annealing was performed by cooling slowly to room temperature. Unincorporated nucleotides were removed using a G25 Sephadex Quick Spin column (Boehringer Mannheim). All equilibrium and kinetic assays were performed at 20 °C in a buffer containing 10 mM Tris-HCl [pH 7.5], 1 mM EDTA, 5% glycerol, 50 mM NaCl, 50 µg/ml bovine serum albumin, and 0.02% NP-40.

For equilibrium gel shift assays radiolabelled DNA fragments (1-10 pM) were incubated with increasing amounts of the engrailed homeodomain for a minimum of 2 h. 30 µl of each binding reaction were loaded onto a 0.5X TBE, 10% polyacrylamide gel running at 300V and after the samples had entered the gel, the voltage was reduced to 155V. It was necessary to load running gels to obtain consistent results. Prior to loading, gels were prerun for a minimum of 45 min at 300V. Tracking dyes were not added to the samples but were loaded in the outside lanes of the gel instead. After electrophoresis, the gels were dried

and exposed to film overnight at -70 °C with an intensifying screen. Binding assays were quantified by scanning densitometry. Because the rate of protein-DNA dissociation is generally fast for the engrailed homeodomain proteins, some complexes dissociate while the gel is running and thus the bound band tends to be diffuse. For this reason, the loss of the free band was used to calculate  $\theta$ , the fraction of bound DNA. Equilibrium dissociation constants ( $K_d$ ) were determined by linear regression using the Scatchard equation:

$$\frac{\theta}{[P]} = \frac{1}{K_d} - \frac{\theta}{K_d}$$

where [P] represents the free protein concentration. Because the DNA concentration used in our binding experiments was well below the  $K_d$ , the free protein concentration was approximated by the total protein concentration.

For the wild-type and QA50 proteins stable gel shifts were obtained with DNA fragments containing the TAATTA site but could not be obtained with DNA fragments containing the TAATCC site. In these cases, equilibrium constants were determined by a competition assay. Sufficient protein was mixed with the radiolabelled TAATTA site to give roughly 80-90% binding, and aliquots were dispensed into tubes with 0.5 nM- 1.0  $\mu$ M concentrations of unlabelled DNA containing the TAATCC site. After equilibration, samples were loaded onto gels and electrophoresed as described above. The equilibrium dissociation constant ( $K_I$ ) for the TAATCC site was determined as a function of  $\theta$  (the fraction of bound radiolabelled DNA calculated as described above), the equilibrium dissociation constant ( $K_d$ ) for the radiolabelled binding site, the total concentration of the unlabelled competitor DNA ( $I_T$ ), and the total concentration of protein ( $P_T$ ) by fitting data to the following equation.

$$K_I = \frac{[P][I]}{[PI]} = \frac{[P][I_T - P_T + P]}{[P_T - P]} = \frac{\left(\frac{\theta \times K_d}{(1-\theta)}\right) \left(I_T - P_T + \frac{\theta \times K_d}{(1-\theta)}\right)}{P_T - \frac{\theta \times K_d}{(1-\theta)}}$$

where [P] and [I] are the concentration of free protein and cold competitor DNA, respectively, and [PI] is the concentration of the complex between the two. The above equation is valid as long as  $P_T \approx P + PI$  (*i. e.* the concentration of radiolabelled DNA is small compared to  $P_T$  and  $I_T$ ).

To measure dissociation rates, sufficient protein was equilibrated with radiolabelled DNA to give roughly 80-90% binding, unlabelled competitor DNA was added to a final concentration of 0.1  $\mu$ M, and at different times 20  $\mu$ l aliquots were loaded directly onto a 0.5X TBE, 10% polyacrylamide gel running at 300V. Gels were electrophoresed and processed as described above. The dissociation rate constant ( $k_d$ ) was determined by fitting the data to the rate equation

$$\ln\left(\frac{\theta}{\theta_0}\right) = -k_d t$$

where  $\theta$  represents the fraction of DNA bound at time  $t$  and  $\theta_0$  represents the fraction bound at time zero.

*Binding Site Selection:* The DNA oligonucleotide N<sub>9</sub> (where N represents an equal mixture of G, A, T, and C) contains nine randomized base positions at its center (Figure 2c). Prior to the first round of binding site selection, a 4-fold molar excess of primer A was annealed to N<sub>9</sub> and extended for 1 h at 37 °C with sequenase v2.0 (USB) in the presence of unlabelled nucleotides and a small amount of  $\alpha$ -<sup>32</sup>P-dATP. Unincorporated nucleotides were removed using a G25

Sephadex Quick Spin column (Boehringer Mannheim) and the labelled duplex DNA was purified on a 1X TBE, 10% polyacrylamide gel.

High affinity binding sites for the engrailed homeodomain were selected using a gel retardation assay. Roughly 0.1 nM of labelled randomized DNA (N<sub>9</sub>; Figure 2c) was equilibrated in 50 µl of binding buffer with 0.1 nM, 1 nM, 10nM, 100 nM, or 1 µM of the engrailed homeodomain for at least 2 h. At this time, 30 µl from each reaction were loaded on a 0.5X TBE, 10% polyacrylamide gel as described above. The gels were dried and exposed to film overnight at -70 °C with an intensifying screen. In each round of the selection, DNA was isolated from the binding reaction containing the lowest concentration of protein for which a bound band was visible by excising the band from the dried gel and soaking it for 3-4 h at 37 °C in elution buffer (0.5 M ammonium acetate, 10 mM MgCl<sub>2</sub>, 1 mM EDTA and 0.1% SDS). After soaking, the buffer was removed from the gel slice and extracted twice with phenol:chloroform (1:1) and then precipitated with ethanol using 1 µg glycogen as a carrier.

The bound DNA fragments were amplified by the polymerase chain reaction (PCR) using one-fourth of the eluted DNA as template. The 100 µl reaction contained 5 mM MgCl<sub>2</sub>, PCR reaction buffer (Perkin Elmer Cetus GeneAmp kit), 20 pmol end-labelled primer A, 20 pmol primer B, 1 mM dNTP's, and 1 U Amplitaq (Perkin Elmer Cetus). The reaction was layered with 60 µl mineral oil and amplified by 20 cycles of 94 °C x 30 s, 55 °C x 30 s and 72 °C x 40 s followed by a final extension at 72 °C x 10 min using a Perkin Elmer Cetus model 480 thermacycler. Amplified DNA was purified on 1X TBE, 10% polyacrylamide gels and used for the next round of selection. As a negative control, a blank slice of the gel was excised after each round and was treated in the same manner as

the bound band. No PCR product was detected from this control indicating that there was no contaminating template. After the final round of selection, the eluted DNA was amplified as before using unlabelled primers. The resulting DNA was extracted twice with phenol:chloroform (1:1), ethanol precipitated twice, and cloned into the vector pBluescript/KS+ (Stratagene). Individual clones were sequenced from single stranded DNA.

Binding site selections using the N<sub>2</sub> oligonucleotide (Figure 2c), which contains the sequence TAATNN, were performed essentially as described above but with the following changes. To select the tightest binding sequences from the 16 possible sequences, a molar excess of DNA over protein was used in the binding reactions after the first round of selection. After the final round of selection and amplification, at least 30 selected binding sites were cloned and sequenced for each protein.

## ***Results***

*Expression, Purification, and Properties of the Engrailed Homeodomain:* The engrailed homeodomain was overproduced in *E. coli* using the T7 expression system from a synthetic gene which encodes the entire 60 amino acid homeodomain and an additional N-terminal methionine. The resulting 61 residue protein was purified to homogeneity using a combination of ion-exchange and reverse phase column chromatography, and its primary structure, including the presence of the N-terminal methionine, was verified by amino acid analysis and N-terminal sequencing.

In thermal denaturation experiments monitored by CD spectroscopy, the engrailed homeodomain undergoes a reversible unfolding transition with a  $t_m$  of

55 °C (Figure 3b) . The CD spectrum at 20 °C (Figure 3a), where the protein is fully folded, is basically that expected for an  $\alpha$ -helical protein but the signal at 222 nm is about two-thirds of the value expected for a protein like the engrailed homeodomain which contains ~60%  $\alpha$ -helix (assuming a value of -33,000 for 100% helix). We were initially concerned that the aberrant CD spectra might indicate that the solution structure of the protein alone differed from that seen in the crystal structure of the DNA-bound complex or indicate chemical or structural heterogeneity in our purified protein. However, several observations argue against these possibilities: (i) the X-ray structure of the protein alone has recently been solved and, with the exception of the N-terminal arm which is disordered, the fold is nearly identical to that seen in the cocrystal structure (N. Clarke, personal communication); (ii) in two-dimensional NMR experiments using our purified protein, we were able to account for all of the  $\alpha$ -helical dNN NOE's expected (not shown); and (iii) additional steps of purification failed to reveal any heterogeneity and protein purified by a variety of methods gave identical CD spectra. It seems likely that the reduced negative ellipticity at 222 nm in the CD spectrum results from positive contributions from one or more of the five aromatic groups in the protein (Woody, 1978; Chakrabarty et al., 1993).

*Selection of High Affinity Binding Sites:* To identify strong binding sites for the engrailed homeodomain, a gel shift selection and PCR amplification procedure based on that of Blackwell and Weintraub (1990) was performed. The oligonucleotide N<sub>9</sub> which contains 9 randomized base positions at its core was used for the binding site selection (Figure 2c). To prevent the sequences flanking the random core from influencing the selection, we did not include any TAAT sequences in these regions and avoided using T or A at the junctions. In the first round of selection, less than 10% of the DNA was bound at a concentration of 100

nM protein (Figure 4). After four rounds of selection and amplification, a bound band was visible at a concentration of 0.1 nM protein (Figure 4) indicating that the pool of DNA was enriched with high affinity sites. At this point, the pool of DNA was cloned and individual clones were sequenced.

Of the 74 binding sites sequenced, 69 could be aligned with the sequence TAAT (Figure 5a). When a clone contained two TAAT sequences or TAAT sequences on both strands of the binding site, each individual occurrence of the sequence was included in the alignment since it was not possible to determine which site had been selected. As a result, a total of 106 individual sequences were included in the alignment. By tabulating the occurrence of each base at a particular position in the binding site, the consensus binding sequence, TAATTA, was determined (Figure 5b). The bases of the core sequence were almost fully restricted to TAAT: at the first position T is preferred in 93% of the sequences, A's are found exclusively at positions 2 and 3, and T is found in 98% of the sequences at position 4. At the fifth position of the six base sequence, T is preferred in 90% of the sequences. At the sixth position A is the preferred base, occurring in 64% of the sequences, but there is a secondary preference for G which is found in 24% of the sequences, and C is notable in its exclusion. The base preference at positions 5 and 6 was confirmed in a second binding site selection using the sequence TAATNN, (N<sub>2</sub>, Figure 2c). As shown in Figure 6, there is a clear preference for T at position 5 and for A at position 6 of the binding site.

*Substitutions at Position 50:* To assess the role of position 50 in the binding of the engrailed homeodomain, we constructed and purified mutant engrailed homeodomains containing an alanine (QA50) and a lysine (QK50) at position 50.

The CD spectra and thermal denaturation profiles of these mutant proteins were very similar to those of the wild-type engrailed homeodomain (Figure 3a, b) indicating that there are no gross structural changes upon mutation and that the mutant proteins, like the wild-type protein, are fully folded at 20 °C, the temperature at which DNA binding was assayed.

To determine whether the alanine and lysine substitutions at position 50 affected the DNA-binding specificity, the preference of the QA50 and QK50 proteins for bases at positions 5 and 6 of the binding site was evaluated by a binding site selection using the TAATNN sequence (Figure 6). The QA50 protein showed a modest preference for T at position 5 but only weak preferences for T or A at position 6. The QK50 protein showed a strong preference for C at position 6, and a modest preference for C at position 5.

*DNA Binding to the TAATTA and TAATCC Sites:* DNA fragments containing the TAATTA and TAATCC sites were synthesized (Figure 2b) and used to determine equilibrium and kinetic constants (Table 1). The half-life of the complex between the wild-type engrailed homeodomain and the TAATTA site was very short (see Figure 8) with over 85% of the complexes dissociating within 4 s ( $k_d \approx 0.28 \text{ s}^{-1}$ ). This rapid dissociation reaction made it technically difficult to obtain consistent equilibrium binding data, but we were able to minimize this problem by loading running gels and by performing a minimum of four repetitions for each experiment. Despite the rapid dissociation reaction, the engrailed homeodomain binds quite strongly to the TAATTA site with a  $K_d$  of  $7.3 \times 10^{-11} \text{ M}$  (Figure 7). This suggests that the association reaction must be close to the diffusion limit for bimolecular reactions (calculated  $k_a \approx 3 \times 10^9 \text{ M}^{-1} \text{ s}^{-1}$ ). The equilibrium binding of the wild-type engrailed homeodomain to the

TAATCC fragment ( $K_d \approx 2.1 \times 10^{-9}$  M) was reduced approximately 25-fold compared with binding to the TAATTA fragment (Table 1). As shown in Table 1, the DNA-binding properties of the QA50 protein are quite similar to those of the wild-type protein. The QA50 protein binds the TAATTA site only 2.4-fold less strongly than wild type, and, like wild type, shows significantly reduced binding to the TAATCC site.

The QK50 protein binds to the TAATTA site only 4-fold less well than wild type, but binds to the TAATCC site roughly 250-fold more strongly than the wild-type engrailed homeodomain (Table 1). This significant increase in the affinity of the QK50 protein for the TAATCC site is also accompanied by kinetic stabilization of the protein-DNA complex. The half-life of the complex of QK50 with TAATCC is 288 s, an increase of more than 100-fold in comparison with the half-life of the complex of wild type with TAATTA (Figure 8).

#### *Discussion:*

An understanding of structure-function relationships requires information at many levels. Although the cocrystal structure of the engrailed homeodomain has been solved (Kissinger et al., 1990), relatively little biochemical or mutagenic information has been available for this system although many such studies have been performed for related homeodomains (Affolter et al., 1990; Percival-Smith et al., 1990; Ekker et al., 1991; Florence et al., 1991; Wilson et al., 1993). The work presented here establishes some of the basic biochemical properties of the engrailed homeodomain and clarifies the role played by amino acid 50 in determining differential DNA-binding specificity.

Using binding site selection experiments, we determined the consensus binding site for the engrailed homeodomain to be TAATTA, the same sequence to which the protein is bound in the cocrystal structure (Kissinger et al., 1990). A consensus sequence obtained from DNaseI footprinting studies of the full-length engrailed protein bound to DNA upstream of the *engrailed* gene includes the sequence TAATTG (Hoey & Levine, 1988). In our selections, TAATTG was the second most favored sequence, suggesting that the binding specificity of the isolated homeodomain is close to that of the full-length protein. We note that the sequence TAATTA was not present in the DNA fragment used for the footprinting experiments.

The paired and fushi tarazu homeodomains contain serine and glutamine, respectively, at position 50 (Scott et al., 1989). It has been shown that when these residues are replaced by lysine, the residue found at position 50 of the bicoid homeodomain, the binding specificity of the variant paired and fushi tarazu proteins is changed to that of the bicoid homeodomain (Treisman et al., 1989; Percival-Smith et al., 1990). Furthermore, when the lysine at position 50 of the bicoid homeodomain is changed to a glutamine as in the antennapedia homeodomain, the binding specificity is changed to that of the antennapedia homeodomain (Hanes & Brent, 1991). Our results confirm that position 50 of the engrailed homeodomain also plays an important role in establishing binding specificity. Specifically, when the glutamine at position 50 is replaced with a lysine, the binding specificity changes from TAATTA to TAATCC. Compared with the binding of the wild-type engrailed homeodomain to the TAATTA site, the QK50 protein has higher affinity for and a longer half-life with the TAATCC binding site. This suggests that the lysine forms a more favorable interaction with the TAATCC site than does the glutamine with the TAATTA site.

Structural studies will be required to establish how the lysine interacts with the CC sequence, but it is tempting to speculate that it may form hydrogen bonds with one or both base-pairs.

Although our results confirm that position 50 of the engrailed homeodomain is important for establishing differential binding specificity, the wild type glutamine at this position does not appear to contribute significantly to the overall energy of DNA binding. When the glutamine at position 50 is replaced by an alanine, the affinity of the protein for the TAATTA site is reduced only 2.4-fold, corresponding to a change in the free energy of binding of 0.5 kcal/mole. This small effect seems consistent with the loss of the van der Waals interaction observed between the Gln50 side chain and the thymine methyl of base-pair 6 in the cocrystal structure (Kissinger et al., 1990). This contact also explains the preference of the wild-type engrailed homeodomain for an A:T base pair at position 6 of the binding site. The QA50 protein does not show a strong base preference at position 6 and examination of the cocrystal structure shows that the C $\beta$  of an alanine at position 50 could not make a van der Waals interaction with the thymine methyl without significant structural changes in the complex. Thus, the crystallographic results and our biochemical results are consistent.

Our results raise the question of why the engrailed homeodomain binds so poorly to the TAATCC binding site. The free energy of binding to the TAATTA site is 1.9 kcal/mol more favorable than the free energy of binding to the TAATCC site and yet the contact made by Gln50 appears to contribute no more than 0.5 kcal/mole to this discrimination: both the wild-type and QA50 homeodomains bind 17-fold to 25-fold more tightly to the TAATTA site than to

the TAATCC site. This suggests that the contact made by Gln50 is not the major determinant of differential specificity between these two DNA sites. One explanation for these observations is that the presence of C:G base-pairs at positions 5 and/or 6 causes conformational changes that weaken interactions made by other homeodomain residues. In this case, we would need to postulate that the favorable interactions between Lys50 and the C:G base pairs at positions 5 and/or 6 are more than sufficient to offset any unfavorable interactions elsewhere in the complex. In the cocrystal structure, the major groove of the DNA is unusually wide and deep around the bound protein compared to the major groove of canonical B-DNA (Nekludova & Pabo, 1994). It will be important to determine the crystal structure of the QK50 engrailed protein bound to the TAATCC site to ask if any significant changes in conformation are observed relative to the wild-type cocrystal.

Another question raised by our studies concerns the structural basis for the preference of the engrailed homeodomain for base-pair 5 of the binding site. In two binding site selection experiments, we observed a strong preference of the engrailed homeodomain for a T:A base-pair at base-pair 5, and yet there are no contacts with this base-pair in the cocrystal structure. Because the QA50 protein also shows some preference for a T:A base-pair at position 5, the differential specificity at this position may depend on other determinants in addition to residue 50 and also involve indirect effects mediated by DNA conformation.

**Acknowledgements:** We thank Bronwen Brown, Neil Clarke, Carl Pabo, and Brenda Schulman for helpful discussions, advice and assistance, and communication of unpublished results.

### References

- Affolter, M., Percival-Smith, A., Müller, M., Leupin, W., & Gehring, W. J. (1990) *Proc. Natl. Acad. Sci. U.S.A.* 87, 4093-4097.
- Blackwell, T. K., & Weintraub, H. (1990) *Science* 250, 1104-1110.
- Chakrabartty, A., Kortemme, T., Padmanabhan, S., & Baldwin, R. L. (1993) *Biochemistry* 32, 5560-5565.
- Ekker, S. C., Young, K. E., von Kessler, D. P., & Beachy, P. A. (1991) *EMBO J.* 10, 1179-1186.
- Florence, B., Handrow, R., & Laughon, A. (1991) *Mol. Cell. Biol.* 11, 3613-3623.
- Hanes, S. D., & Brent, R. (1989) *Cell* 57, 1275-1283.
- Hanes, S. D., & Brent, R. (1991) *Science* 251, 426-430.
- Hoey, T., & Levine, M. (1988) *Nature* 332, 858-861.
- Kissinger, C. R., Liu, B., Martin-Blanco, E., Kornberg, T. B., & Pabo, C. O. (1990) *Cell* 63, 579-590.
- Laughon, A. (1991) *Biochemistry* 30, 11357-11367.
- Nekludova, L., & Pabo, C. O. (1994) *in press*.
- Percival-Smith, A., Müller, M., Affolter, M., & Gehring, W. J. (1990) *EMBO J.* 9, 3967-3974.
- Qian, Y. Q., Billeter, M., Otting, G., Müller, M., Gehring, W. J., & Wüthrich, K. (1989) *Cell* 59, 573-580.
- Sanger, F., Nicklen, S., & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A.* 74, 5463-5467.
- Schagger, H., & von Jagow, G. (1987) *Anal. Biochem.* 166, 368-379.
- Scott, M. P., Tamkun, J. W., & Hartzell, G. W. (1989) *Biochim. Biophys. Acta* 989, 25-48.
- Studier, F. W., Rosenberg, A. H., Dunn, J. J., & Dubendorff, J. W. (1990) *Methods Enzymol.* 185, 60-89.

Treisman, J., Gönczy, P., Vashishtha, M., Harris, E., & Desplan, C. (1989) *Cell* 59, 553-562.

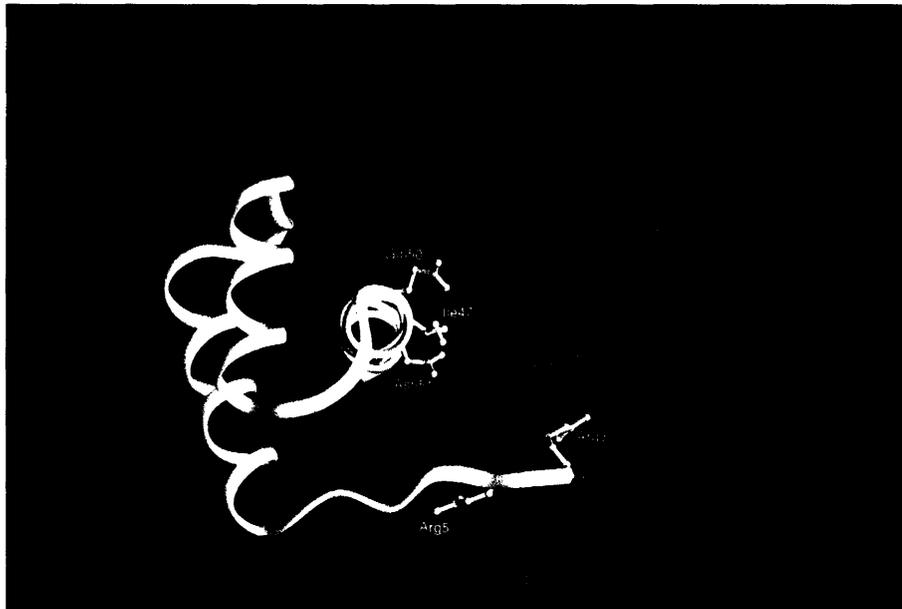
Wilson, D., Sheng, G., Lecuit, T., Dostatni, N., & Desplan, C. (1993) *Genes Dev.* 7, 2120-21334.

Woody, R. W. (1978) *Biopolymers* 17, 1451-1467.

<b>Protein</b>	<b>Binding Site</b>	<b>Kd</b>	<b>Half-Life</b>
WT	TAATTA	$7.9 (\pm 2.3) \times 10^{-11}$ M	$\leq 2.5$ s
QA50	TAATTA	$1.9 (\pm 0.5) \times 10^{-10}$ M	$< 2.5$ s
QK50	TAATTA	$3.2 (\pm 1.6) \times 10^{-10}$ M	$< 2.5$ s
WT	TAATCC	$2.1 (\pm 0.8) \times 10^{-9}$ M	n. d. <sup>a</sup>
QA50	TAATCC	$3.4 (\pm 2.6) \times 10^{-9}$ M	n. d. <sup>a</sup>
QK50	TAATCC	$8.8 (\pm 4.7) \times 10^{-12}$ M	289 s

**Table 1: Equilibrium and Kinetic DNA-Binding Constants**

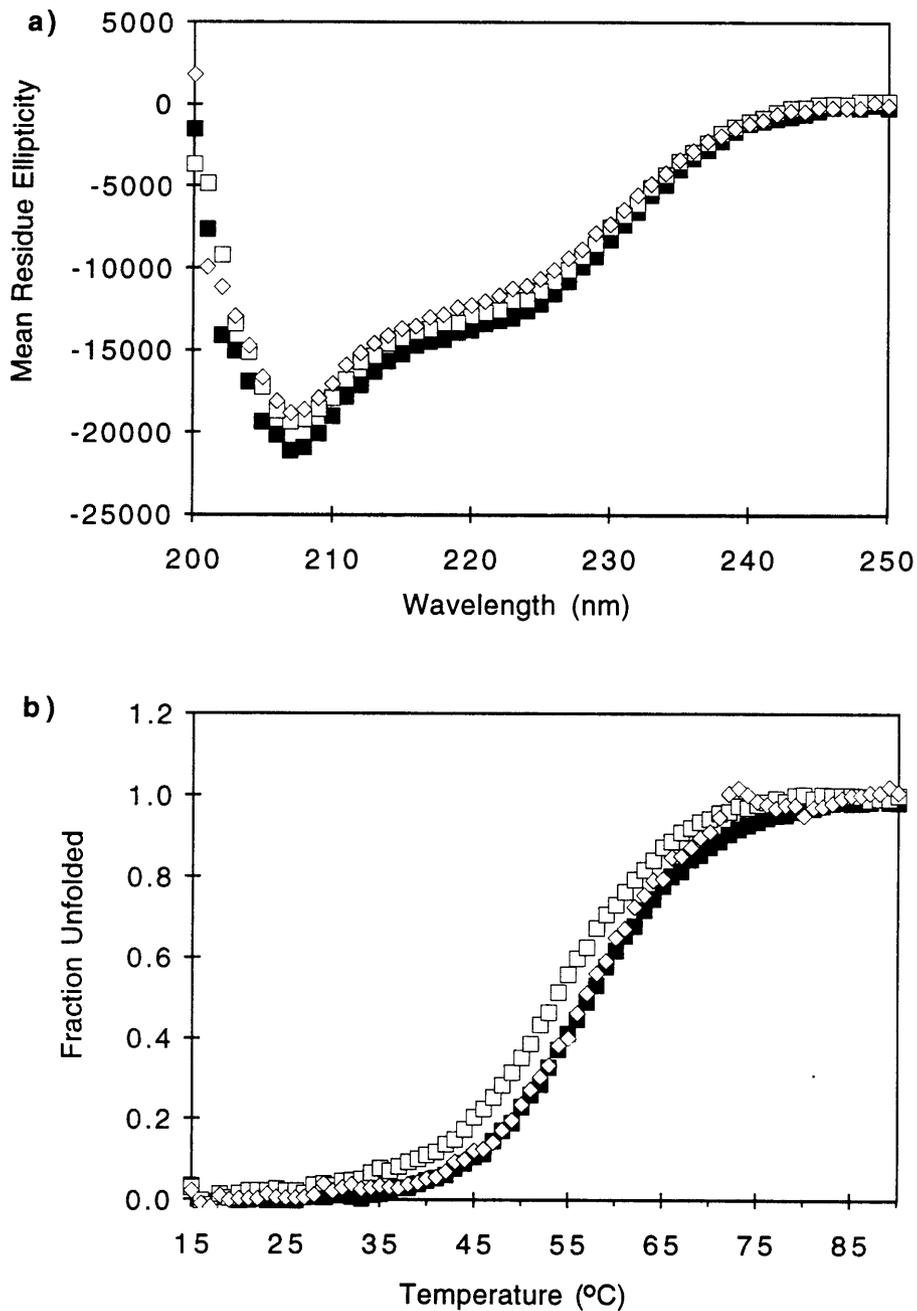
<sup>a</sup> It was not possible to determine the half-life of the WT and QA50 proteins with the TAATCC site because they do not give a stable gel shift with this binding site.



**Figure 1: Molecular graphics representation of the engrailed homeodomain bound to DNA** (Kissinger et al., 1990). The protein backbone is shown as a ribbon trace and the five side chains that make base contacts are shown in ball-and-stick representation. The thymine methyl group which interacts with Gln50 is marked by a van der Waals surface.

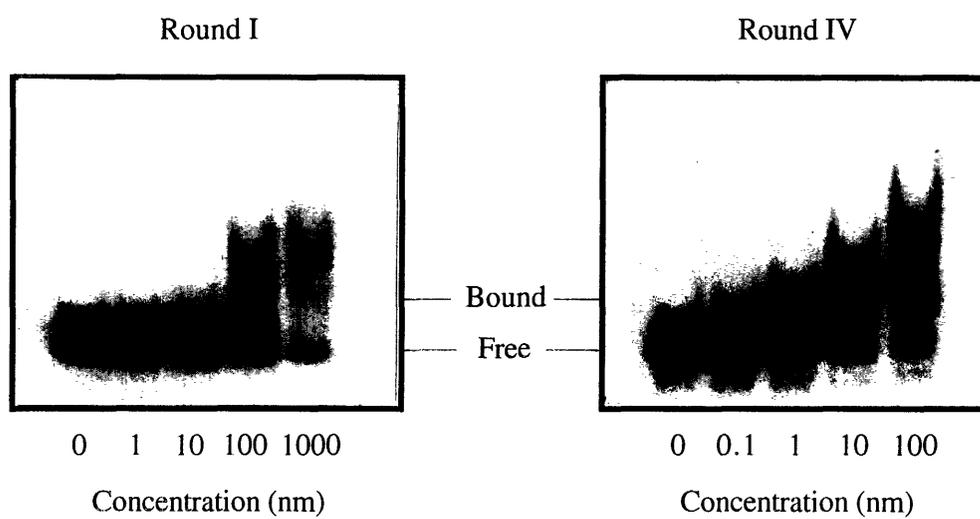
**Figure 2:** *(following page)* **a) Sequence of the gene constructed to encode the engrailed homeodomain.** Unique restriction enzyme sites are indicated. The amino acid numbering is according to Qian et al. (1989) to maintain consistency with other homeodomains. **b) Sequences of DNA fragments used for binding assays.** **c) Sequences of synthetic oligonucleotides used for binding site selections.** Locations of primers for PCR are indicated. *N* refers to an equimolar combination of all four nucleotides.





**Figure 3:** a) CD spectra of the proteins: wild type (3.3  $\mu\text{M}$ , filled boxes), QA50 (2.4  $\mu\text{M}$ , open diamonds), and QK50 (3.3  $\mu\text{M}$ , open boxes). b) Thermal denaturation of the wild type, QA50, and QK50 proteins (symbols and concentrations as in panel A). Fitting of the denaturation curves using nonlinear least squares methods yields the following values: wild type,  $t_m = 55.5\text{ }^\circ\text{C}$ ,  $\Delta H = 35.9\text{ kcal/mole}$ ; QA50,  $t_m = 56.8\text{ }^\circ\text{C}$ ,  $\Delta H = 39.0\text{ kcal/mole}$ ; QK50,  $t_m = 53.8\text{ }^\circ\text{C}$ ,  $\Delta H = 34.3\text{ kcal/mole}$ .

**Figure 4: (following page) Gel mobility shift assays from the first and final rounds of binding site selection for the wild-type engrailed homeodomain.** The concentration of the engrailed homeodomain in each lane is indicated. The left-most lane of each gel is a no protein control.

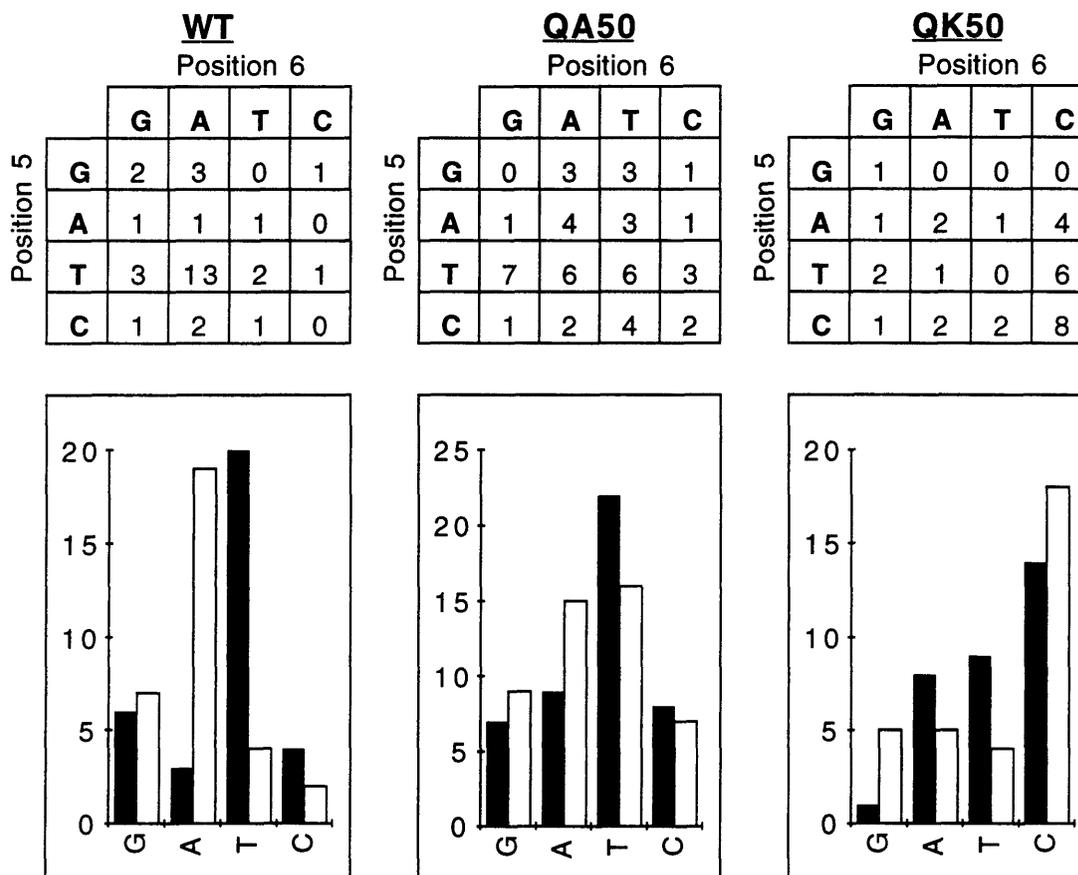


**Figure 5: (following pages) a) Aligned individual binding sites for the engrailed homeodomain obtained after *in vitro* selection. In the designation of each clone, T and B refer to the top and bottom strands of the clone with respect to the sequencing primer. b) Tabulation of the aligned data.**

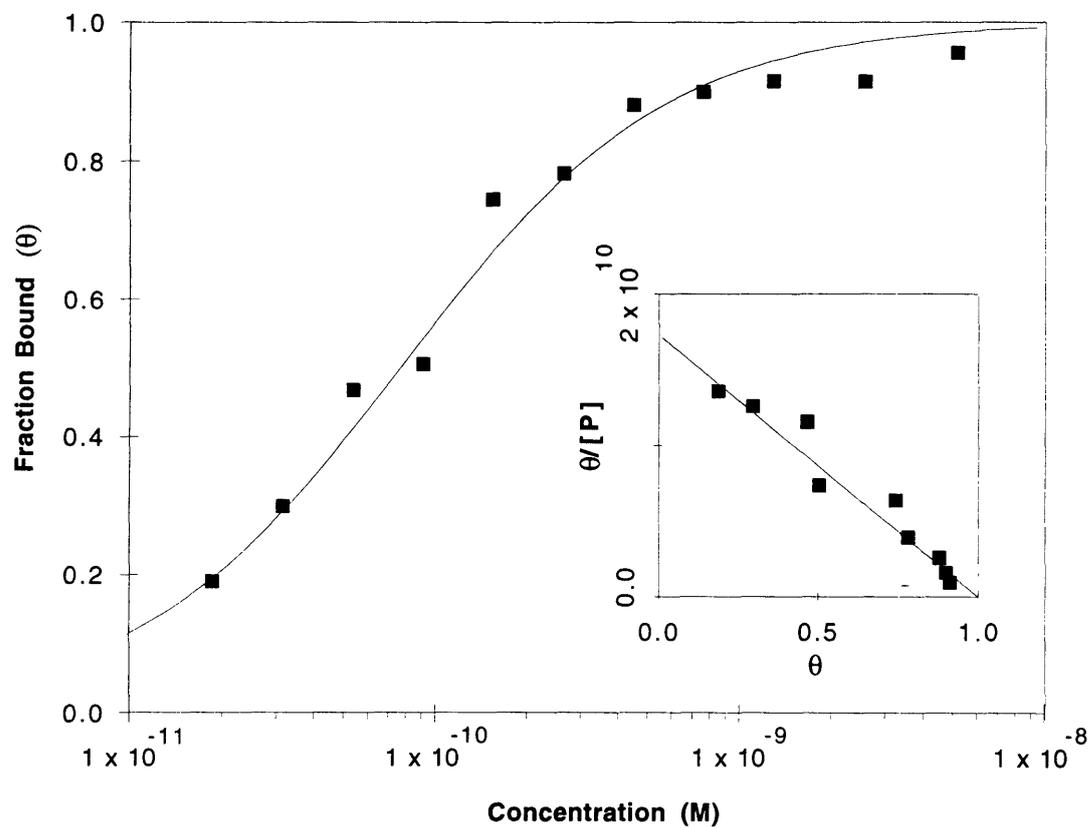
23-1-T	A C T A A T T G	36-T	T A A T T A T T A
23-2-B	C T T A A T T G	36-B1	T A A T A A T T A
23-3-T	A T G A T T A A T	36-B2	T A A T A A T T A
23-3-B	A T T A A T C A T	40-T1	A A T C A A T G A
24-B	C C G A T A A T T	40-T2	A A T C A A T G A
22-T	G T A A A T T G T	41-T	C A T T A A T T A
22-B	A C A A T T T A C	41-B1	T A A T T A A T G
21-B	T A A T T T A G G	41-B2	T A A T T A A T G
20-1-T	T T T A A T T A G	42-T	A G A T A A T T A
20-1-B	C T A A T T A A A	42-B	T A A T T A T C T
20-2-T	T A T A A T T A	43-3-T	A A T T A A T T A
20-2-B	T A A T T A T A	43-3-B1	T A A T T A A T T
20-3-B	C A T T T A A T T	43-3-B2	T A A T T A A T T
19-T	C T T A A T G A T	45-B	C A C T A A T T G
18-T	G C A T A A T T A	46-T	A A T T A A T A G
18-B	T A A T T A T G C	46-B	C T A T T A A T T
17-B	T G G T A A T T G	37-1-T	T T T A A T T A
16-T	G C T A A T T A C	37-1-B	T A A T T A A A A
16-B	G T A A T T A G C	37-3-B	A A T A A T G A G
15-T	A A T T T A T C G	38-1-T	A A C T A A T T A
15-B	C G A T A A A T T	38-1-B	T A A T T A G T T
11-T	T A A T T G C T G	38-3-B	C T A A T A T A C
10-B	T C A C T A A T T	47-B	C G T G T A A T T
9-T	A A A T T A A T T	48-1-B	T A T A A T T G A
9-B	A A T T A A T T T	48-3-T	C A G T A A T T A
8-T	A A A T T A A T T	48-3-B	T A A T T A C T G
8-B	A A T T A A T T T	53-T	A T A T A A T T A
7-2-B	C T G T A A T T G	53-B	T A A T T A T A T
7-3-T	T T A A T T A C A T T	54-B	A G T T A A T T G
7-3-B	A A T G T A A T T A A	56-B	T A A T T G A T T
6-B	T T C C T A A T T	57-T	G T A A T T G G C
5-T	A A C T A A T G A	58-T	A A T T T A C A C
4-T	G C T T A A T G A	58-B	G T G T A A A T T
3-B	C A T A A T T G G	59-T	A G T A A T T A T
1-T	A A G T A A T T A	59-B	A T A A T T A C T
1-B	T A A T T A C T T	39-1-T	A C T A A T T A A
35-B	G A G T T A A T T	39-1-B	T T A A T T A G T
34-T	A A T T A A T T T	39-2-T	T A A T T A G T
34-B	A A A T T A A T T	39-2-B	A C T A A T T A
33-T	C T A A T T A G C	71-T	C A C T A A T T A
33-B	G C T A A T T A G	71-B	T A A T T A G T G
32-B	A G A T A A T G A	70-T	C C T A A T T A C
31-T	A T A A T T A G G	70-B	G T A A T T A G G
31-B	C C T A A T T A T	69-T	A G T A A T T G A
30-B	C G C T A A T T G	67-T	G A T A A T T G C
29-1-T	T A C T A A T T G	65-T	C A T A A T T A C
28-1-T	A T A A T T G T A	65-B	G T A A T T A T G
28-3-T	T T T A A T T A G	64-B	T G G G C T A A A
28-3-B	C T A A T T A A A	63-B	T G C A T A A T T
27-T	T A A T T A G C T	62-T	C A A T T T A C G
27-B	A G C T A A T T A	62-B	C G T A A A T T G
25-B	G A G T T A A T T	72-B	C G T A A A T T G
		61-T	T A T A A T T G G
		60-B	A C A T A A T T G
UNALIGNED:		48-T-2	G T T G A T T G
49-T	C A T A T A G A A	31-T	A T C A A A G G G
50-T	T G A G T C T A A	64-T	A C T G T C G G C

Figure 5b

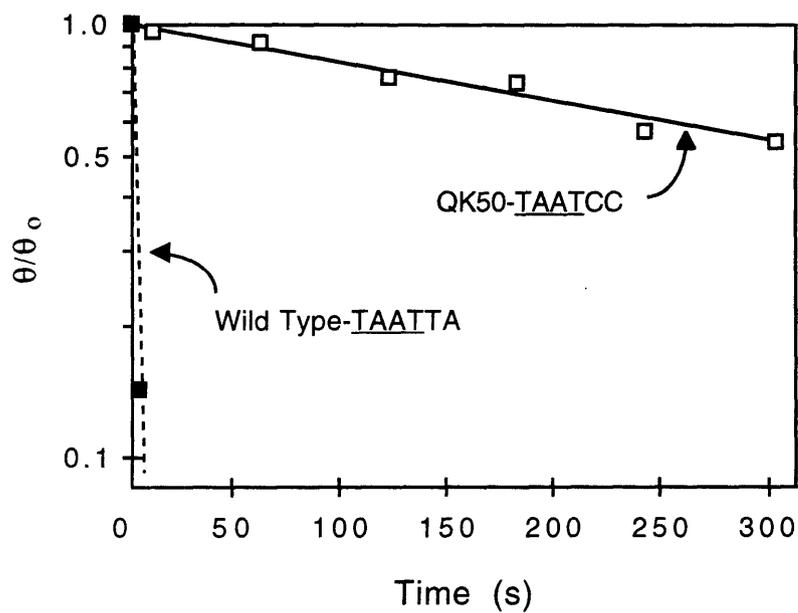
position:	<b>T</b>	<b>A</b>	<b>A</b>	<b>T</b>	<b>5</b>	<b>6</b>
G:	0	0	0	0	7	21
A:	4	<b>106</b>	<b>106</b>	2	3	<b>57</b>
T:	<b>96</b>	0	0	<b>104</b>	<b>92</b>	10
C:	3	0	0	0	2	0
total:	103	106	106	106	104	88
Consensus:	<b>T</b>	<b>A</b>	<b>A</b>	<b>T</b>	<b>T</b>	<b>A</b>



**Figure 6: Binding site preferences at positions 5 and 6, for the wild-type, QA50, and QK50 engrailed homeodomains following *in vitro* selections using TAATNN.** Individual sequences are listed in the tables. The charts present the data as the number of occurrences of each base at position 5 (filled bars) and position 6 (open bars).



**Figure 7: Equilibrium binding curve for the wild-type engrailed homeodomain binding to a DNA fragment containing the TAATTA site. The solid line represents a theoretical curve with  $K_d = 7 \times 10^{-11}$  M. In the inset the same data is plotted in Scatchard form.**



**Figure 8: Dissociation kinetics of complexes of the wild-type engrailed homeodomain and the TAATTA site (filled boxes) and the QK50 homeodomain and the TAATCC site (open boxes).** Time is measured from the addition of cold competitor DNA. The half-life of the wild-type complex is  $\sim 2.5$  s and the half-life of the QK50 complex is  $\sim 300$  s.

## **Chapter 3**

### **Specificity of Minor-Groove and Major-Groove Interactions in a Homeodomain-DNA Complex**

## *Introduction*

Specificity is one of the hallmarks of the binding of transcription factors to their DNA recognition sites. Interactions mediated by the protein must be capable of providing both high-affinity binding to the proper site and permitting discrimination against closely related DNA sites. The structures of a large number of protein-DNA complexes have now been solved, providing a detailed molecular view of the interactions with the DNA bases and sugar-phosphate backbone that stabilize the correct protein-DNA complex. In these complexes, the majority of interactions with the DNA bases are localized to the major groove. There are fewer examples of interactions with bases in the minor groove and less is known about the importance of these interactions. Modelling studies have suggested that minor-groove interactions may be less specific than major-groove interactions because there are fewer ways to uniquely distinguish among the hydrogen-bond acceptors and donors on the edges of bases in the minor groove (Seeman et al., 1976).

The homeodomain provides a simple model system in which to study the relative contributions of minor-groove and major-groove interactions to the specificity and stability of a protein-DNA complex. In each of the known homeodomain-DNA structures, residues from  $\alpha$ -helix 3 mediate a set of major-groove contacts, while residues from an extended N-terminal arm, which is unstructured in the absence of DNA, mediate a set of minor-groove contacts (Kissinger et al., 1990; Otting et al., 1990; Woldberger et al., 1991; Billeter et al., 1993; Klemm et al., 1994). The overall architecture of the engrailed homeodomain-DNA complex is illustrated in Fig. 1. In the cocrystal structure, residues from the N-terminal arm of engrailed contact the

minor-groove edges of first two bases of the core sequence TAAT (Kissinger et al., 1990). These two base-pairs are not contacted in the major groove, and yet are strongly conserved in binding site selections *in vitro* (Ades & Sauer, 1994). Moreover, the optimal binding sites of many homeodomains contain the same TAAT core sequence (Müller et al., 1988; Ekker et al., 1991; Florence et al., 1991; Regulski et al., 1991; Ekker et al., 1992; Catron et al., 1993). These findings suggest that homeodomains can discriminate among potential binding sites on the basis of minor-groove interactions but do not provide a quantitative analysis of the specificity afforded by these interactions. Here, we use binding site selections and mutagenesis of a homeodomain and its DNA site to gain a better understanding of the role of minor-groove and major-groove interactions in determining binding energy and specificity. The wild-type engrailed homeodomain can bind in two symmetric orientations to its preferred site, TAATTA, which complicates mutational studies. In the studies presented here, we use an altered-specificity mutant of the engrailed homeodomain containing lysine at position 50 which binds in a unique orientation to the DNA site TAATCC (Ades & Sauer, 1994). We refer to this protein as the altered-specificity homeodomain. The cocrystal structure of the altered-specificity homeodomain complexed with the TAATCC site has recently been determined (Tucker-Kellogg, L., Rould, M. A., Chambers, K. A., Ades, S. E., Sauer, R. T., & Pabo, C. O., *manuscript in preparation*).

### ***Materials and Methods***

*Oligonucleotides:* Oligonucleotides were synthesized on an Applied Biosystems Model 381A DNA synthesizer and gel purified by standard

methods. The oligonucleotide, 5'-cgcagtg**TAATCC**cctcgac-3', and its complement, with an additional 5' overhang for end-filling purposes, were synthesized for binding studies. The altered-specificity binding site is indicated in bold-face type and all binding site mutations studied were in this background. The sequence of the oligonucleotide used for binding site selections, N<sub>2</sub>, is 5'-ccgcaggcaactcgagcttacgtcg**NNATCC**gctgcagtcatgctctccgtct-3' (where N refers to an equimolar mixture of A, T, G, and C). Primers to N<sub>2</sub> were synthesized for PCR and second-strand synthesis.

*Site Directed Mutagenesis:* All proteins used in this work are derivatives of the altered-specificity mutant of the engrailed homeodomain which contains Lys50 in place of the wild-type Gln50 residue (Ades & Sauer, 1994). Mutants encoding the RA3, RA5, IA47, and NA51 substitutions were constructed by cloning synthetic oligonucleotide cassettes encoding the substitutions between the appropriate restriction sites of plasmid pSEA100-QK50, which expresses the altered-specificity protein from the T7 promoter. The altered-specificity homeodomain and a variant containing Ala50 (KA50) were available from a previous study (Ades & Sauer, 1994).

*Expression and Purification of Proteins:* Proteins were purified from *Escherichia coli* strain BL21(DE3)/pLysS transformed with the appropriate derivatives of pSEA100-QK50 essentially as described (Ades & Sauer, 1994). Cells were grown in 400 mL of LB broth supplemented with 150 µg/mL ampicillin, and transcription from the T7 promoter was induced by the addition of IPTG to 0.4 mM. After cells were harvested by centrifugation, the cell pellet was resuspended in 15 mL lysis buffer [100 mM Tris-HCl (pH 8.0), 200 mM KCl, 1 mM EDTA, 2 mM CaCl<sub>2</sub>, 10 mM MgCl<sub>2</sub>, 2 mM NaN<sub>3</sub>, and 50%

glycerol], 10  $\mu$ L of a fresh 100 mM solution of phenylmethanesulfonyl fluoride in ethanol was added to inhibit proteolysis, and cells were lysed by sonication. Nucleic acids were precipitated by the addition of 0.5% polyethyleneimine, and proteins were precipitated from the resulting supernatant by the addition of solid ammonium sulfate to 95% saturation. The ammonium sulfate pellet was resuspended in column buffer [25 mM Tris-HCl (pH7.5), 0.1 mM EDTA, and 1.4 mM 2-mercaptoethanol] plus 100 mM NaCl, dialyzed extensively against the same buffer, and loaded onto a 5 mL DEAE Sephacel column. The flow-through fraction and first column volume of wash from the DEAE column were collected and loaded directly onto a 10 mL Affi-Gel Blue column (Bio-Rad; 100-200 mesh) equilibrated in column buffer plus 100 mM NaCl. The protein was eluted with successive washes of column buffer containing increasing concentrations of NaCl. The altered-specificity homeodomain and IA47 and NA51 variants eluted in the 0.7 - 0.8 M NaCl washes while the RA3 and RA5 variants eluted slightly earlier, in the 0.6 - 0.7 M NaCl washes, consistent with the removal of a positively charged arginine side chain. Each of these proteins binds tightly to the Affi-Gel Blue resin and is ~99% pure upon elution as judged by Coomassie Blue staining of Tris-tricine polyacrylamide gels (Schagger & von Jagow, 1987). The fractions containing pure protein were dialyzed into column buffer plus 100 mM NaCl, concentrated by ultrafiltration, and stored at 4 °C.

Circular dichroism experiments were conducted to monitor the folding and stability of the homeodomains. Spectra of samples containing protein at 3  $\mu$ M in 50 mM potassium phosphate (pH 7.0) and 100 mM KCl were obtained by averaging five scans, each collected at 20 °C in 1 nM steps with a 1 s

averaging time. The thermal stabilities of the proteins were determined by measuring the ellipticity at 222 nm at 1 °C intervals from 15 °C to 90 °C with a 1 min equilibration time and 30 s averaging time. The CD spectra (data not shown) and the thermal stabilities of the purified mutants were very similar to that of the altered-specificity homeodomain, suggesting that none of the mutations affects the overall fold or stability of the homeodomain. Fitting the denaturation curves using nonlinear least squares methods yields the following values: altered-specificity homeodomain,  $t_m=53.8$  °C,  $\Delta H=34.3$  kcal/mol; RA3,  $t_m=55.2$  °C,  $\Delta H=33.3$  kcal/mol; RA5,  $t_m=54.9$  °C,  $\Delta H=32.4$  kcal/mol; IA47,  $t_m=56.6$  °C,  $\Delta H=33.4$  kcal/mol; KA50,  $t_m=56.8$  °C,  $\Delta H=39.0$  kcal/mol; and NA51,  $t_m=50.1$  °C,  $\Delta H=25.6$  kcal/mol. All of the proteins are greater than 99% folded at 20 °C, the temperature at which DNA-binding affinities were measured.

*Equilibrium and Kinetic Assays of DNA Binding:* When necessary, double-stranded binding site oligonucleotides used in gel mobility shift assays were labeled by end-filling in a reaction containing 1 picomole DNA in sequenase reaction buffer, 1 U sequenase v2.0 (United States Biochemicals), and 30  $\mu\text{Ci}$  [ $\alpha^{32}\text{P}$ ]dATP (6000 Ci/mMole) for 30 - 60 min at room temperature. The reactions were extracted with phenol: chloroform (1:1) and unincorporated nucleotides were removed using a G-25 Sephadex quick-spin column (Boehringer Manneheim). Equilibrium and kinetic constants were determined using gel mobility shift assays performed at 20 °C in binding buffer containing 10 mM Tris-HCl (pH 7.5), 0.1 mM EDTA, 50 mM NaCl, 0.02% NP-40, 50  $\mu\text{g/ml}$  bovine serum albumin, and 5% glycerol.

Equilibrium gel mobility shift assays were conducted as previously described (Ades & Sauer, 1994). Briefly, varying concentrations of protein were incubated with radiolabeled DNA fragments ( $\leq 5$  pM) for 2 h in a 50  $\mu$ L reaction and then 30  $\mu$ l were loaded onto 0.5X TBE, 10% polyacrylamide gels (pre-run for >30 min at 300 V) running at 300 V. The voltage was reduced to 150 V after the samples had entered the gel. Tracking dyes were loaded in the outer lanes of the gel and were not included in the samples. After electrophoresis, gels were dried and exposed to film at -70 °C with an intensifying screen. Binding assays were quantified by scanning densitometry and the loss of the free band was used to determine the fraction of bound DNA. Equilibrium dissociation constants were determined by linear regression using the Scatchard equation. Three or more gel mobility shift assays were conducted for each binding constant determined.

Equilibrium constants for dissociation of the altered-specificity homeodomain protein from variant binding sites were determined using a competition gel mobility shift assay. Sufficient protein was added to bind 80-90% of a radiolabeled TAATCC fragment at a concentration of 1 pM. Aliquots of this mixture were added to tubes containing 0.02 nM- 250 nM competitor DNA's containing the variant binding sites. After equilibration, samples were loaded onto gels and electrophoresed as described above. The free bands were quantified and equilibrium dissociation constants for the competitor DNA fragments were calculated as described in Ades & Sauer (1994). Again three or more assays were conducted for each binding constant determined.

Dissociation rates were measured by assaying the increase in free radiolabeled DNA as a function of time after the addition of unlabeled

competitor DNA. Sufficient protein was equilibrated with radiolabeled binding site oligonucleotides to bind 80-90% of the DNA. An excess of unlabeled competitor DNA was added and aliquots were loaded at the appropriate times onto a 0.5X TBE, 10% polyacrylamide gel running at 300 V. Gels were electrophoresed and processed as described above. Dissociation rate constants were determined by fitting the data to a first-order rate equation.

*Binding Site Selections:* The base preferences of the altered-specificity homeodomain and the RA3 and RA5 variants at the first two positions of the DNA binding site were determined in selections using the N<sub>2</sub> oligonucleotide which contains the sequence NNATCC (where N represents an equal mixture of A, T, G, and C). The starting pool of DNA was generated by annealing a primer to N<sub>2</sub> and extending with sequenase v2.0 in the presence of unlabeled nucleotides and a small amount of [ $\alpha$ <sup>32</sup>P]dATP. In the first round of selection, protein at several different concentrations, from 0.001 nM to 1  $\mu$ M depending on the variant, was incubated with roughly 0.5 nM randomized DNA in a 50  $\mu$ l reaction. Bound DNA was separated from free DNA using a gel mobility shift assay as described above. The bound DNA was eluted from dried gels from the lane containing the lowest concentration of protein for which a bound band was visible: 0.1 nM for the altered-specificity homeodomain, 1 nM for the RA3 variant, and 100 nM for the RA5 variant. The eluted DNA was then amplified by the polymerase chain reaction using a <sup>32</sup>P end-labeled primer and subjected to three more rounds of selection and amplification. In these latter rounds of selection with the altered-specificity homeodomain and RA3 variant, an excess of DNA over protein was used in the binding reactions, 0.05 - 0.07 nM altered-specificity homeodomain and 0.3 - 0.6 nM RA3 protein were equilibrated with roughly 1 - 2 nM amplified DNA. In the

remaining rounds of selection with the RA5 variant, roughly equimolar quantities of protein and DNA (~ 10 nM) were used in binding reactions. After four rounds of selection and amplification, the selected pools of binding sites were cloned between the XhoI and PstI sites of pBluescript/KS+ (Stratagene) and individual clones were sequenced.

## *Results*

*Contributions of Minor-Groove and Major-Groove Contacts Probed By Alanine Mutations:* To determine the contribution of side chain-base contacts to the overall DNA-binding energy of the altered-specificity homeodomain, we constructed alanine substitution mutants for each residue involved in a base contact in the cocrystal structure (Fig. 2; Kissinger et al., 1990; Tucker-Kellogg et al., *in preparation*). The five mutant proteins (RA3, RA5, IA47, KA50, and NA51) were purified, and the binding of each variant to the TAATCC site was probed by gel mobility shift assays. As shown in Table 1, the affinities of the RA3 and IA47 proteins for the TAATCC site were reduced by roughly 10-20 fold and the dissociation rates of the protein-DNA complexes were increased by roughly 20-fold. The RA5, KA50, and NA51 mutations reduced binding to the point where stable, quantifiable gel shifts with the TAATCC site were not observed. Based on the faint gel shifts that were observed, it appears that the affinities of these mutant proteins for the TAATCC site are reduced by at least two orders of magnitude. The affinity of the KA50 mutant for the TAATCC site was measured in a previous study by a competition method and found to be reduced by roughly 400-fold (Ades & Sauer, 1994).

*Binding Site Selections to Probe the Base Preferences at Positions 1 & 2:*

Based on the cocrystal structures (Kissinger et al., 1990; Tucker-Kellogg et al., *in preparation*), the minor-groove edges of the first two base pairs of the TAATCC binding site are expected to be contacted by Arg3 and Arg5 from the homeodomain's N-terminal arm (Figs. 1 & 2). To assess the base preferences at these sites of minor-groove interactions, binding site selections were conducted with the altered-specificity homeodomain, the RA3 mutant, and the RA5 mutant using a population of binding sites in which the first two positions were randomized (NNATCC). Although the RA5 protein does not give a shifted band that is sufficiently stable for quantification, it does give a faint, shifted band at high concentrations of protein. After four rounds of selection and amplification, the pools of DNA enriched for tightly binding sequences were cloned and sequenced. The results are shown in Table 2. The altered-specificity homeodomain shows a marked preference for the expected bases, T (80%) at position 1 and A (90%) at position 2. The binding site preferences of the RA3 mutant are broader. At the first position, T (58%) is still the preferred base but there is a secondary preference for A (36%). At the second position, A (60%) is the preferred base with weaker preferences for G (25%) and T (11%). For the RA5 mutant, the significant preferences seem to be against C at position 1 and against C and T at position 2. It is also notable that C:G base pairs were rarely recovered at either position in any of the three selections, even though sequencing of randomized but unselected oligonucleotides showed that C:G base pairs were present at reasonable frequencies in the starting pool (Table 2).

*Affinity for Binding Sites with Substitutions at Positions 1 & 2:* To evaluate the ability of the altered-specificity homeodomain to discriminate

among binding sites in a more quantitative fashion, equilibrium dissociation constants for a set of binding sites containing natural base-pair substitutions at positions 1 and 2 were determined (Table 3). The affinities for binding sites with C:I and I:C substitutions were also measured to help distinguish between minor-groove and major-groove effects. Inosine lacks the exocyclic N2 amino group of guanine and thus a C:I base pair resembles a T:A base pair in the minor groove and a C:G base pair in the major groove (Fig. 3). The equilibrium dissociation constants for each of the altered sites were determined using an assay in which an unlabeled variant site competed for binding of the altered-specificity homeodomain to a labeled TAATCC site.

*Position 1:* Compared to the preferred TAATCC site, the altered-specificity homeodomain shows modestly reduced affinity for each of the position 1 variants tested, including those with I:C or C:I (Table 3). The largest loss of affinity, about 6-fold, occurs when the wild-type T:A base pair is replaced by a C:G base pair. The C:I substitution, which differs from the C:G base pair only in the minor groove, reduces affinity 3 to 4-fold. Sites bearing the A:T, G:C, or I:C transversion substitutions also have affinities reduced by approximately 3 to 4-fold. In the cocrystal structure, the side chain of Arg5 contacts the base pair at position 1 (Fig. 2). It is important to note that the reductions in affinity caused by the base-substitution mutations at position 1 are small when compared with the greater than 100-fold reduction caused by the RA5 mutation.

*Position 2:* The affinities of the altered-specificity homeodomain for sites with substitutions at position 2 (TAATCC) fall into two classes (Table 3). Transition mutations (A:T to G:C or I:C) have little effect on affinity (reduced

2 to 3-fold for G:C; unchanged for I:C). Transversion mutations (A:T to T:A, C:G or C:I) have larger effects ranging from 7 to 28-fold. Surprisingly, affinity is reduced 7-fold for the T:A mutation but reduced 21-fold for the C:I mutation. Since both base pairs have identical functional groups in the minor groove, it seems likely that interactions mediated by the major-groove edge of the base pairs must be responsible for the observed difference.

Because DNA sites with a purine on the top-strand at position 2 have the highest affinities for the altered-specificity homeodomain, we reasoned that the N7 position of the purine might be important for binding. To test this idea, we determined the affinity for a binding site with a  $^{7\text{C}}$ A:T base pair (where  $^{7\text{C}}$ A represents N7-deazaadenine; see Fig. 3) at position 2. The affinity for this  $^{7\text{C}}$ A:T site was reduced 5-fold, consistent with the idea that the N7 position in the major groove does influence binding affinity in some fashion. To provide a comparison, we also synthesized a DNA site with a  $^{7\text{C}}$ A:T base pair replacing the normal A:T base pair at position 3. The side chain of Asn51 forms a bidentate hydrogen bond to the N7 and N6 positions of this adenine in the protein-DNA complex. When adenine 3 is changed to N7-deazaadenine, the affinity of the altered-specificity homeodomain is reduced approximately 100-fold (Table 3).

*Binding of RA3 to Sites Altered at Position 2:* In principle, the RA3 mutation should remove interactions with the minor-groove edge of base-pair 2 but should not directly affect any major-groove interactions. Hence, the results described above suggest that the RA3 mutant should retain some sensitivity to position 2 alterations to the extent that the effects of these alterations are mediated through the major groove. To test this, the affinity of the RA3 mutant was determined for several position 2 mutants (Table 3).

In general, the relative affinity of RA3 for each mutant DNA site is reduced compared with the relative affinity of the parent protein for that site. However, the RA3 protein still binds more strongly to the TAATCC site than to any of the position 2 variant sites. These two findings are consistent with idea that the base pair at position 2 affects affinity both via minor-groove interactions mediated by Arg3 and through major-groove interactions.

*Effects of Base-Pair Substitutions at Positions 5 & 6:* As mentioned above, most homeodomain proteins, including engrailed, bind to sites containing the conserved core sequence TAAT. The differential specificity of homeodomains, however, is frequently determined by the identity of residue 50 in the protein and the identities of positions 5 and 6 in the DNA site (Hanes & Brent, 1989; Treisman et al., 1989; Percival-Smith et al., 1990; Hanes & Brent, 1991). To evaluate the contributions of base-pairs 5 and 6 to the affinity and specificity of binding by the altered-specificity homeodomain, we measured the affinity of the protein for binding sites with all natural, single base-pair substitutions at these positions (Table 3). At position 5, each base-pair substitution reduces binding 10 to 13-fold. At position 6, each substitution reduces binding 9 to 20-fold.

*Interactions Between DNA Positions:* In the binding site selections, specificity is broadened at both the first and second position when either Arg3 (which is thought to contact position 2) or Arg5 (which contacts position 1) are changed to alanine (Table 2). This finding suggests that interactions with the first two base pairs of the binding site may be coupled. If these interactions are coupled, then the effects of mutations at positions 1 and 2 of the DNA site should not be additive in terms of binding energies. To test

this, the affinities of the altered-specificity homeodomain for the ATATCC and CCATCC sites were measured (Table 3). Binding to the ATATCC site is reduced by about 1.1 ( $\pm 0.3$ ) kcal/mol, whereas a 2.0 ( $\pm 0.5$ ) kcal/mol reduction would be expected if the effects of each mutation were independent. Binding to the CCATCC site is reduced by 2.4 ( $\pm 0.4$ ) kcal/mol, a value within error of the 3.0 ( $\pm 0.5$ ) kcal/mol reduction expected on the basis of independent mutant effects. Hence, interactions of the homeodomain with base-pairs 1 and 2 seem to be energetically coupled for binding to some sites but not others.

The preferred binding site of the wild-type engrailed homeodomain (TAATTA) differs from the preferred site for the altered-specificity protein (TAATCC) at both positions 5 and 6. The affinity of the altered-specificity protein for the TAATTA site is reduced 2.1 ( $\pm 0.4$ ) kcal/mol (Table 3; Ades & Sauer, 1994) compared to the preferred site, whereas a reduction of 3.0 ( $\pm 0.5$ ) kcal/mol would be expected if the base-substitution effects were independent. This result suggests a small energetic coupling between interactions at base-pairs 5 & 6.

### *Discussion*

Cocrystal structures of protein-DNA complexes provide a three dimensional map of molecular interactions, while biochemical studies provide a way to address the importance of interactions to binding affinity and specificity. Structures have been solved for both the engrailed homeodomain/TAATTA complex (Kissinger et al., 1990) and the engrailed altered-specificity/TAATCC complex (Tucker-Kellogg et al., *in preparation*)

providing a basis for interpreting the functional studies presented here. In particular, we have probed the contribution to affinity and specificity of two parts of the altered-specificity complex: interactions between the protein's N-terminal arm and the minor groove and interactions of the lysine at position 50 with bases in the major groove.

Several issues need to be considered in evaluating the results presented here. First, in considering the effects of mutations at base positions 1 & 2 with those at 5 & 6, minor-groove interactions are being compared to major-groove interactions but interactions from a flexible region of protein are also being compared with those from a relatively rigid unit of secondary structure. Second, in addition to perturbing the expected base contacts, a mutation may perturb backbone or base contacts at other positions via effects on the overall DNA or protein structure. Finally, a base change introduces new functional groups which may permit new interactions with the protein. Structures of each mutant complex would be needed to know with certainty whether significant conformational changes occur or new contacts are made. Nevertheless, comparing the observed functional effects of mutations with the simplest expectations based upon the known protein-DNA structures is still worthwhile. In cases, where this fails to provide a satisfactory explanation, more complex mechanisms probably contribute to the observed effects and structural studies are indicated.

Three side chains of the altered-specificity homeodomain (Ile47, Lys50, Asn51) make base contacts in the major groove and two side chains (Arg3, Arg5) make base contacts in the minor groove. As measured by the effects of alanine substitution mutations, each of these side chains contributes to

binding affinity, albeit at different levels. Mutations of Arg3 or Ile47 reduce affinity modestly (1.3-1.7 kcal/mol), while mutations of Arg5, Asn51, or Lys50 have larger effects (> 2.7 kcal/mol). At a general level, these results show that the overall energetic contributions of the minor-groove interactions made by the flexible N-terminal arm are comparable to those of the major-groove interactions made by the recognition helix.

To compare the contributions to binding specificity of the base pairs at positions 1, 2, 5, and 6 of the binding site, we calculated the specificity index ( $I_{\text{spec}}$ ) as defined by Stormo et al. (1991). For each position,  $I_{\text{spec}}$  is calculated from the relative affinities of the three mutant sites with natural base substitutions and ranges from 0 bits of information (no specificity) to 2 bits of information (maximum specificity). For the altered-specificity homeodomain the results are as follows: position 1,  $I_{\text{spec}}=0.41$ ; position 2,  $I_{\text{spec}}=0.63$ ; position 5,  $I_{\text{spec}}=0.97$ ; and position 6,  $I_{\text{spec}}=1.02$ . By this measure, base-pairs 5 and 6 of the binding site have a higher information content, *i.e.* greater specificity, than base-pairs 1 and 2. Hence, the major-groove interactions mediated by residue 50 of the recognition helix are more specific than those formed in the minor groove by the N-terminal arm.

Several general points are worth noting with respect to specificity and affinity. First, the minor-groove interactions made by the flexible N-terminal arm do contribute to binding specificity, even if the effect is modest. The modelling studies of Seeman et al. (1976) correctly suggested that minor-groove interactions would have lower specificity than major-groove interactions but also indicated that proteins would not be able to differentiate between T:A and A:T base pairs in the minor groove. However, the altered-

specificity homeodomain differentiates between T:A and A:T at position 1 as well as it differentiates between other base substitutions. Finally, there is no simple correlation between the affinities suggested for particular interactions by the alanine mutations and the specificities inferred for these contacts. For example, Arg5 and Lys50 contribute approximately equally to affinity but the interactions mediated by Arg5 show significantly lower specificity than those mediated by Lys50. We assume that this occurs because Arg5, to a greater extent than Lys50, is able to make alternative contacts with either the mutant bases or the sugar-phosphate backbone of the mutant DNA. Both the arginine and lysine side chains should have comparable flexibility, but contacts from the N-terminal arm are presumably more easily rearranged than those from the recognition  $\alpha$ -helix.

Binding site selection experiments provide an additional probe of binding specificity which can be compared to results from affinity measurements. Selections for positions 1 and 2 were performed here and selections for positions 5 and 6 were described previously (Ades & Sauer, 1994). As shown in Fig. 4, although both methods identify the same preferred bases (T:A at position 1, A:T at position 2, C:G at positions 5 & 6), the binding site selections can overestimate or underestimate the degree of specificity. This is not surprising. First, because many rounds of site selection and amplification are performed, there is no reason that the results should be strictly proportional to thermodynamic stability. Second, the binding site selections were performed following randomization of several base pairs. If there are cooperative interactions between base positions (as appears to be the case both for positions 1 & 2, and 5 & 6; see below), this will affect the selections in a manner not mirrored by single-site affinity studies.

The structural and mutational analyses of the interaction between Arg3 and base-pair 2 provide examples of some of the complexities that can emerge in such studies. In the cocrystal structures, Arg3 of the wild-type engrailed homeodomain is positioned to contact the minor-groove face of base-pair 2 (Kissinger et al., 1990) but Arg3 in the altered-specificity homeodomain is poorly ordered (Tucker-Kellogg et al., *in preparation*). Nevertheless, our mutational studies indicate that Arg3 does contribute to binding affinity of the altered-specificity homeodomain and show that interactions mediated by base-pair 2 contribute to binding specificity. However, transition mutations at this position have a much smaller effect than transversion mutations indicating that purines are favored on the sense strand of the binding site by the altered-specificity homeodomain. Several lines of evidence suggest that this effect is mediated, at least in part, through the major groove of the DNA. First, substitution of the preferred A:T base pair with C:I reduces binding to a greater extent than with T:A, even though both C:I and T:A have similar functional groups in the minor groove. Second, substitution of A:T with <sup>7</sup>C A:T also reduces affinity, even though this substitution only affects the major groove. Third, the RA3 mutant, which should no longer interact with the minor groove at position 2, is still sensitive to base substitutions at this position. In the altered-specificity complex, there are no contacts between side chains and base-pair 2 in the major groove, although there is an ordered water close to the N7 of adenine 2 (Tucker-Kellogg et al., *in preparation*). This latter interaction might contribute to specificity, although it seems more likely that transversions affect the structure of the DNA to some extent and thereby perturb other contacts in the complex.

The non-additivity of some mutational effects suggests that several interactions between the homeodomain and DNA are energetically coupled. Coupling is most simply explained by entropic considerations when interactions help to stabilize each other and is commonly observed when the mutations probe functional groups that are close or interact directly in the structure (Wells, 1990). At positions 1 & 2, both the results of binding site selections and the non-additivity of mutations (for the ATATCC site, in particular) suggest linkage. The Arg3 and Arg5 side chains do not appear to interact with each other in the cocrystal structure, but interactions of these amino acids with the DNA could serve to fix the position of the otherwise flexible arm and help position the second amino acid for its contact. At positions 5 & 6, linkage can be explained in a simple fashion since the  $\epsilon$ -amino group of Lys50 is positioned to form hydrogen bonds with both bases.

*Acknowledgements:* We thank Kristen Chambers, Carl Pabo, Mark Rould, and Lisa Tucker-Kellogg for helpful discussions, advice, and assistance.

## References

- Ades, S. E., & Sauer, R. T. (1994) *Biochemistry* 33, 9187-9194.
- Billeter, M., Qian, Y. Q., Otting, G., Müller, M., Gehring, W. J., & Wüthrich, K. (1993) *J. Mol. Biol.* 234, 1084-1097.
- Catron, K. M., Iler, N., & Abate, C. (1993) *Mol. Cell. Biol.* 13, 2354-2365.
- Ekker, S. C., Young, K. E., von Kessler, D. P., & Beachy, P. A. (1991) *EMBO J.* 10, 1179-1186.
- Ekker, S. C., von Kessler, D. P., & Beachy, P. A. (1992) *EMBO J.* 11, 4059-4072.
- Florence, B., Handrow, R., & Laughon, A. (1991) *Mol. Cell. Biol.* 11, 3613-3623.
- Hanes, S. D., & Brent, R. (1989) *Cell* 57, 1275-1283.
- Hanes, S. D., & Brent, R. (1991) *Science* 251, 426-430.
- Kissinger, C. R., Liu, B., Martin-Blanco, E., Kornberg, T. B., & Pabo, C. O. (1990) *Cell* 63, 579-590.
- Klemm, J. D., Rould, M. A., Aurora, R., Herr, W., & Pabo, C. O. (1994) *Cell* 77, 21-32.
- Müller, M., Affolter, M., Leupin, W., Otting, G., Wüthrich, K., & Gehring, W. J. (1988) *Embo J.* 7, 4299-4304.
- Percival-Smith, A., Müller, M., Affolter, M., & Gehring, W. J. (1990) *EMBO J.* 9, 3967-3974.
- Otting, G., Qian, Y. Q., Billeter, M., Müller, M., Affolter, M., Gehring, W. J., & Wüthrich, K. (1990) *EMBO J.* 9, 3085-3092.
- Regulski, M., Dessain, S., McGinnis, N., & McGinnis, W. (1991) *Genes Dev.* 5, 278-286.
- Schagger, H., & von Jagow, G. (1987) *Anal. Biochem.* 166, 368-379.
- Seeman, N. C., Rosenberg, J. M., & Rich, A. (1976) *Proc. Natl. Acad. Sci. U. S. A.* 73:, 804-808.

Stormo, G. D., & Yoshioka, M. (1991) *Proc. Natl. Sci. U. S. A.* 88, 5699-5703.

Treisman, J., Gönczy, P., Vashishtha, M., Harris, E., & Desplan, C. (1989) *Cell* 59, 553-562.

Wells, J. A. (1990) *Biochemistry* 29, 8509-8517.

Wolberger, C., Vershon, A. K., Liu, B., Johnson, A. D., & Pabo, C. O. (1991) *Cell* 67, 517-528.

Amino Acid Substitution <sup>a</sup>	DNA Contact	K <sub>d</sub> (M)	Relative Affinity (K <sub>d-mut</sub> /K <sub>d-wt</sub> ) <sup>b</sup>	Relative k <sub>off</sub> (k <sub>off-mut</sub> /k <sub>off-wt</sub> ) <sup>b</sup>
RA3 (arm)	Minor Groove Position 2	8.8 (± 4.3) x 10 <sup>-11</sup>	10	20
RA5 (arm)	Minor Groove Position 1	n.d.	>100	n.d.
IA47 (helix 3)	Major Groove Position 4	1.9 (± 0.7) x 10 <sup>-10</sup>	20	17
KA50 (helix 3)	Major Groove Position 5 & 6	n.d. <sup>c</sup>	>100	n.d.
NA51 (helix 3)	Major Groove Position 3	n.d.	>100	n.d.

**Table 1: Equilibrium and Kinetic DNA-Binding Constants for Alanine Substitution Mutants**

n.d.: It was not possible to determine the equilibrium and kinetic constants for these proteins because they do not give a stable gel shift.

*a* As noted in the text, all mutations are in the altered-specificity (Lys50) background of the engrailed homeodomain.

*b* K<sub>d-wt</sub> = 8.9 (± 4.0) x 10<sup>-12</sup> M and k<sub>off-wt</sub> = 0.003 sec<sup>-1</sup>, the equilibrium and kinetic constants for the altered-specificity homeodomain binding to the TAATCC site at 20 °C in binding buffer.

*c* The affinity of the KA50 mutant for the TAATCC binding site was measured in a previous study using a different method and found to be 3.4 (± 2.6) x 10<sup>-9</sup> M (Ades & Sauer, 1994).

	Altered-Specificity Homeodomain		RA3 Mutant		RA5 Mutant		Unselected	
	Position 1	Position 2	Position 1	Position 2	Position 1	Position 2	Position 1	Position 2
Base	1	2	1	2	1	2	1	2
G	1	2	2	9	8	9	10	7
A	5	28	13	22	5	8	7	5
T	25	1	21	4	7	2	8	11
C	0	0	0	1	0	1	5	7
Total	31	31	36	36	20	20	30	30

**Table 2: Tabulation of Data from Binding Site Selections Using NNATCC**

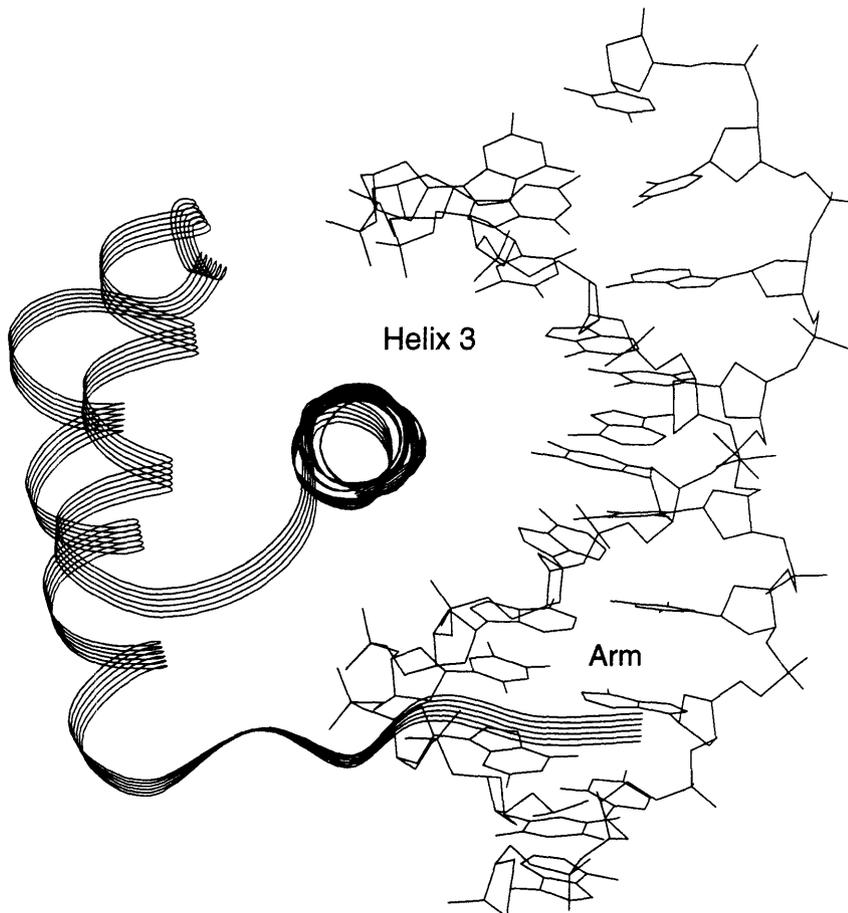
The recovery of individual bases at positions 1 and 2 in sites after binding site selections using the altered-specificity homeodomain, the RA3 variant of the altered-specificity homeodomain, and the RA5 variant of the altered-specificity homeodomain is tabulated. The last column shows the recovery of individual bases at positions 1 and 2 in sites from the starting pool of oligonucleotides which were not subjected to binding site selection. The totals refer to the number of binding sites sequenced.

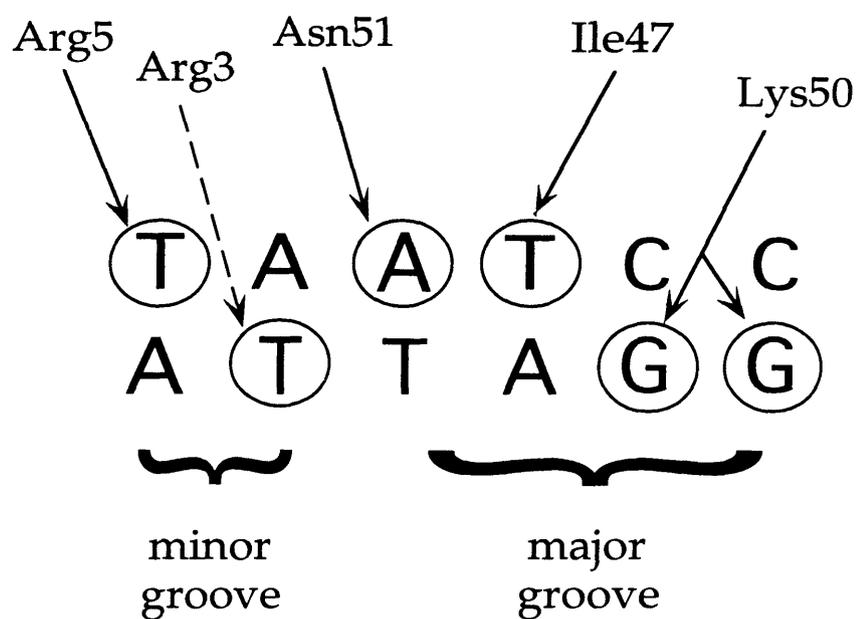
Binding Site						Altered-Specificity Homeodomain			RA3 Mutant		
						Kd (pM)	rel. aff. <sup>a</sup>	ΔΔG (kcal/mol)	Kd (pM)	rel. aff. <sup>a</sup>	ΔΔG (kcal/mol)
T	A	A	T	C	C	8.9 (±4)	1.0	-	88 (±43)	1.0	0.0
A	T	T	A	G	G						
C	-	-	-	-	-	56 (±29)	6.3	1.1			
G	-	-	-	-	-						
C	-	-	-	-	-	31 (±15)	3.5	0.7			
I	-	-	-	-	-						
A	-	-	-	-	-	39 (±15)	4.4	0.9	150 (±50)	1.7	0.3
T	-	-	-	-	-						
G	-	-	-	-	-	31 (±15)	3.5	0.7			
C	-	-	-	-	-						
I	-	-	-	-	-	25 (±12)	2.8	0.6			
C	-	-	-	-	-						
-	G	-	-	-	-	22 (±9)	2.5	0.5			
-	C	-	-	-	-						
-	I	-	-	-	-	13 (±6)	1.5	0.2			
-	C	-	-	-	-						
-	T	-	-	-	-	60 (±18)	6.7	1.1	250 (±100)	2.8	0.6
-	A	-	-	-	-						
-	C	-	-	-	-	250 (±70)	28.1	1.9	1100 (±500)	12.5	1.5
-	G	-	-	-	-						
-	C	-	-	-	-	190 (±90)	21.3	1.8	1600 (±500)	18.2	1.7
-	I	-	-	-	-						
-	<sup>70</sup> C	A	-	-	-	45 (±14)	5.1	0.9			
-	T	-	-	-	-						
C	C	-	-	-	-	570 (±240)	64.0	2.4			
G	G	-	-	-	-						
A	T	-	-	-	-	60 (±23)	6.7	1.1	170 (±60)	1.9	0.4
T	A	-	-	-	-						
-	-	<sup>70</sup> C	A	-	-	960 (±340)	107.9	2.7			
-	-	T	-	-	-						
-	-	-	-	T	-	120 (±43)	13.5	1.5			
-	-	-	-	A	-						
-	-	-	-	A	-	92 (±46)	10.3	1.4			
-	-	-	-	T	-						
-	-	-	-	G	-	120 (±49)	13.5	1.5			
-	-	-	-	C	-						
-	-	-	-	-	T	83 (±39)	9.3	1.3			
-	-	-	-	-	A						
-	-	-	-	-	A	120 (±60)	13.5	1.5			
-	-	-	-	-	T						
-	-	-	-	-	G	180 (±50)	20.2	1.7			
-	-	-	-	-	C						
-	-	-	-	T	A	320 (±160)	36.0	2.1			
-	-	-	-	A	T						

**Table 3: Equilibrium DNA-Binding Constants to Altered Sites**

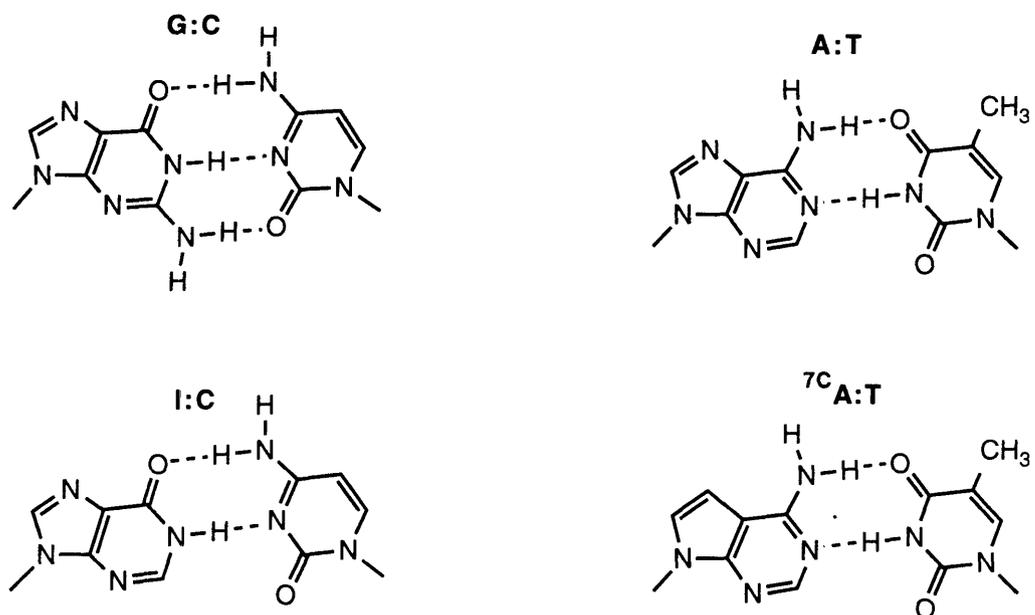
<sup>a</sup> Relative Affinity ( $K_{d\text{-site}}/K_{d\text{-TAAATCC}}$ )

**Figure 1: Molecular graphics representation of the engrailed homeodomain bound to DNA (Kissinger et al., 1990).** The polypeptide backbone of the protein is represented by a ribbon. The view is down  $\alpha$ -helix 3 which lies in the major groove. The N-terminal arm of the protein lies in the minor groove.



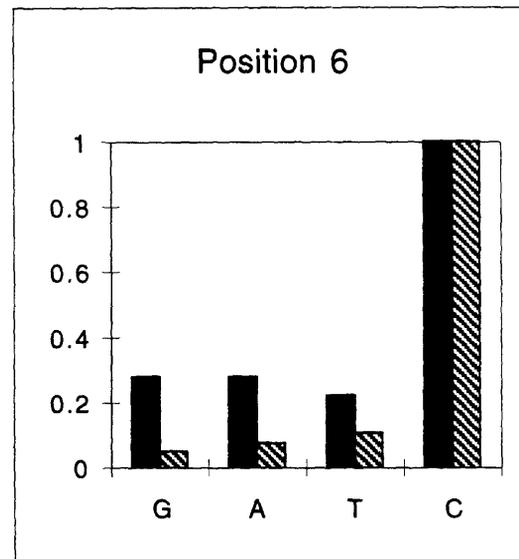
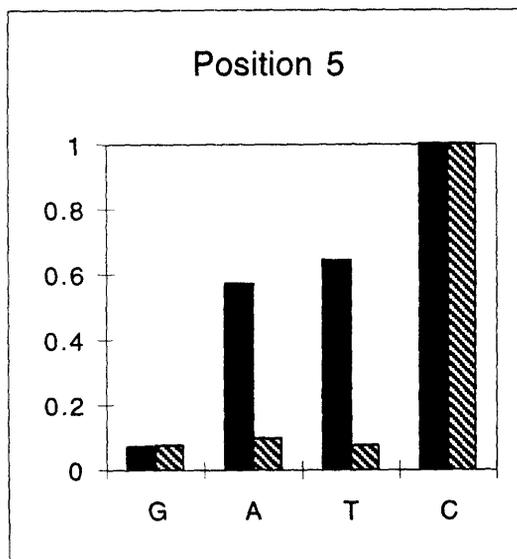
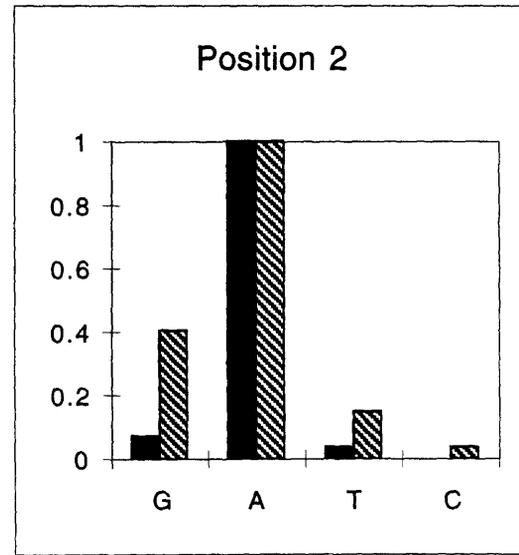
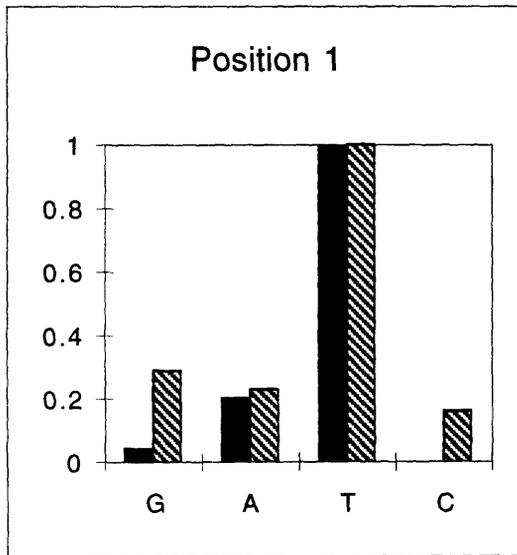


**Figure 2: Contacts between the altered-specificity homeodomain and DNA.** Solid lines indicate contacts from the cocrystal structure (Tucker-Kellogg et al., *in preparation*). Dashed lines indicate structurally plausible contacts inferred from the mutational studies presented in this chapter. Contacts with the major-groove and minor-groove edges of base pairs are indicated.



**Figure 3: Base pairs used for binding site substitutions in these studies. <sup>7C</sup>A is N7-deazaadenine.**

**Figure 4: (following page) Comparison of specificity inferred from binding site selections (solid bars) and affinity measurements (striped bars) at positions 1, 2, 5, and 6.** All measurements are normalized to the preferred base at that position: (frequency of base/frequency of preferred base) for binding site selections and ( $K_d$ -preferred site/ $K_d$ -mutant site) for affinity measurements. Binding site selection experiments for positions 5 and 6 are from Ades & Sauer (1994).



■ Selections  
▨ Affinity

## **Chapter 4**

### **Homeodomain-DNA Recognition**

The work in this thesis examines determinants of DNA-binding specificity and affinity of the homeodomain from the *Drosophila* transcription factor engrailed. The engrailed protein is required for establishing and maintaining posterior compartment identity in the developing embryo (Lawrence & Morata, 1976; Kornberg, 1981) and appears to act as a repressor of transcription (Jaynes & O'Farrell, 1988; Ohkuma et al., 1990). DNA binding is mediated by the 60 amino acid homeodomain near the C terminus of the protein (Poole et al., 1985). The target sites of the engrailed protein *in vivo* are not known. As such, I have focused on how the homeodomain recognizes a biochemically defined, optimal binding site.

Changing one amino acid in the engrailed homeodomain can alter its DNA-binding specificity (Chapter 2, Ades & Sauer, 1994). The isolated engrailed homeodomain binds as a monomer with high affinity to the six base-pair consensus site TAATTA determined in binding site selections. By changing the glutamine at residue 50 of engrailed to a lysine, the binding site preference changes to TAATCC. The change in site preference is accompanied by an increase in the affinity and stability of the complex, indicative of the addition of DNA contacts. A variant of the homeodomain with an alanine at residue 50 can bind to the wild-type site and discriminate between the wild-type and altered-specificity sites nearly as well as the wild-type Gln50 homeodomain. This result implies that, for the engrailed homeodomain, determinants other than Gln50 are involved in determining differential specificity.

Homeodomains bind to DNA with two surfaces: the N-terminal arm in the minor groove contacts the first two base pairs of the binding site, and  $\alpha$ -

helix 3 in the major groove contacts the remaining four base pairs of the binding site (Kissinger et al., 1990; Wolberger et al., 1991; Billeter et al., 1993; Klemm et al., 1994). The work presented in Chapter three illustrates the role each part of the complex plays in DNA recognition by the altered-specificity, (Lys50), engrailed homeodomain. Briefly, both the interactions of the arm in the minor groove and  $\alpha$ -helix 3 in the major groove contribute to the overall binding energy to roughly the same extent. However, although the homeodomain does exhibit base preferences at the site of minor-groove interactions by the arm, the specificity of these interactions is less than that of the interactions between Lys50 of  $\alpha$ -helix 3 and the DNA.

### *Future Directions*

Homeodomains are structurally related and bind to DNA in a similar manner but often have distinct DNA-binding properties. These qualities make the homeodomain a useful system in which to study site-specific DNA recognition. The work in this thesis describes the basic components of recognition for the engrailed homeodomain. Future avenues for research concerning both homeodomain-DNA interactions, in particular, and protein-DNA recognition, in general, are outlined below.

*Differential Specificity:* A large class of homeodomains bind to sites containing the sequence TAAT (Laughon, 1991). Recognition of the core sequence is accomplished by residues which are highly conserved among these homeodomains: residue 3 is generally Arg or Lys, residue 5 is nearly always Arg, residue 47 is generally Ile or Val, and residue 51 is an invariant Asn (Scott et al., 1989). The binding sites for these homeodomains vary

outside of the TAAT core and many experiments have shown that the amino acid at position 50 of the homeodomain is largely responsible for differential specificity at positions immediately 3' to the core sequence (Hanes & Brent, 1989; Treisman et al., 1989; Percival-Smith et al., 1990; Hanes & Brent, 1991). Several questions still remain about how differential specificity at these positions is achieved.

1) The altered-specificity experiments conducted to date have focused on the role of glutamine and lysine at position 50 in determining differential DNA-binding specificity. When the wild-type amino acid at position 50 of the engrailed (Gln50), fushi tarazu (Gln50), or paired (Ser50) homeodomain is replaced by lysine, the residue found at position 50 of the bicoid homeodomain, the altered-specificity homeodomains bind to the bicoid binding site, TAATCC (Treisman et al., 1989; Percival-Smith et al., 1990; Ades & Sauer, 1994). In addition, when the amino acid at position 50 of the paired (Ser50) or bicoid (Lys50) homeodomain is changed to glutamine, as found in the antennapedia homeodomain, the variant homeodomains now bind to the antennapedia binding site, TAATTG (Hanes & Brent, 1989; Treisman et al., 1989; Hanes & Brent, 1991). Thus, for the glutamine and lysine substitutions, the base preferences at positions following the core, TAATNN, are determined by the amino acid at position 50. However, it is not known whether this rule will hold true for other amino acid substitutions at position 50. Would a variant of the engrailed homeodomain with a serine at position 50, as found in the paired homeodomain, recognize the paired binding site, TAATCG (Treisman et al., 1992)? Could any amino acid, even those not naturally found at position 50 in homeodomains, function in the homeodomain context? If this is the case, a variety of specificities could be

achieved by changing the identity of the amino acid at position 50. The answer to this question will provide insights into the evolution of new binding specificities and the tolerance of a DNA-binding motif to variation.

2) As discussed in chapter 2 for the engrailed homeodomain, determinants other than the glutamine at position 50 may be involved in determining differential specificity at the two positions following the TAAT core. A variant of the engrailed homeodomain with an alanine at position 50 can discriminate between the TAATTA site and the TAATCC site almost as well as the wild-type (Gln50) homeodomain, *i. e.* both homeodomains bind tightly to the TAATTA site and poorly to the TAATCC site. Similar results were obtained for the bicoid homeodomain using genetic assays in *Drosophila* embryos (Hanes et al., 1994). Hanes et al. (1994) observed that the bicoid protein activates transcription of a reporter gene with the bicoid binding site, TAATCC, in the upstream regulatory region. A variant of bicoid with a glutamine at position 50 of the homeodomain only activated reporter genes with TAATGA or TAATTA sites but not the TAATCC site. When an alanine was placed at position 50, the Ala50 bicoid protein activated reporter genes with the TAATTA site but not the TAATGA or TAATCC sites. These results are unexpected since the alanine substitution effectively truncates the side chain and should remove any interactions between residue 50 and the DNA, thereby reducing the specificity and affinity of the protein.

Structural studies of the engrailed QA50 variant bound to both the TAATTA and TAATCC sites would be of particular interest in understanding how differential specificity is achieved for the Ala50 homeodomains. (The engrailed QA50 variant does bind to the TAATCC site, albeit with reduced

affinity, so it should be possible to obtain crystals of the complex.) According to the cocrystal structure of the wild-type engrailed homeodomain bound to TAATTA (Kissinger et al., 1990), an alanine side chain should be too far from the DNA to form direct interactions with the bases. It will be interesting to see whether there are rearrangements in the QA50-DNA complex which would allow the alanine to interact with the binding site.

3) A large number of homeodomains have glutamine at position 50 yet show differences in their binding-site preferences at positions following the TAAT core. For instance, both the engrailed homeodomain and the ultrabithorax homeodomain have glutamine at position 50, yet engrailed prefers a TAATTA site (Ades & Sauer, 1994) while ultrabithorax prefers a TAATGG site (Ekker et al., 1991). Other parts of the complex may contribute to DNA-binding specificity at these positions. One possible source for different binding-site preferences is the amino acid at position 54 of the homeodomain which projects into the major groove and can potentially influence binding specificity in this region. In the NMR structure of the antennapedia homeodomain bound to DNA, the methionine at position 54 contacts the DNA at position 5, TAATNN (Billeter et al., 1993); and in the cocrystal structure of the  $\alpha 2$  homeodomain bound to DNA, the arginine at position 54 forms hydrogen bonds with a guanine at position 4 (Wolberger et al., 1991). In the engrailed homeodomain-DNA complex, the alanine at position 54 is too far from the DNA to contact bases in the binding site (Kissinger et al., 1990). Thus, the engrailed homeodomain would be a good system in which to investigate the influence of the amino acid at position 54 on differential DNA-binding specificity by replacing the alanine in engrailed

with amino acids found in other homeodomains at this position: methionine, arginine, serine, and glutamine.

*Phosphate contacts:* Seven amino acids in the engrailed homeodomain contact the sugar-phosphate backbone on either side of the major groove (Kissinger et al., 1990). Several of these amino acids are conserved and Arg53 is invariant among homeodomain sequences (Scott et al., 1989). In addition, in the cocrystal structures of the engrailed and  $\alpha 2$  homeodomains bound to DNA, 6 of the 8 backbone contacts are conserved suggesting that these contacts are needed to position the homeodomain correctly on the DNA (Kissinger et al., 1990). The role of phosphate contacts in specific recognition has been evaluated in a systematic manner for relatively few protein-DNA complexes. For the tetrameric Arc repressor-DNA complex, phosphate contacts were shown to generally have a smaller contribution to the overall binding affinity than direct base contacts (Brown et al., 1994). However, the engrailed homeodomain binds to DNA as a monomer and there are fewer overall interactions between the protein and DNA than for an oligomeric protein. Therefore do the phosphate contacts play a larger role in complex stability for monomeric DNA-binding proteins like the engrailed homeodomain? Are some phosphate contacts more important than others?

*N-terminal Arm:* The N-terminal arm of the homeodomain plays a crucial role in DNA recognition. Deletion of the N-terminal arm of the fushi tarazu homeodomain severely reduces DNA binding affinity (Percival-Smith et al., 1990). The work in Chapter 3 of this thesis demonstrates the importance of Arg3 and particularly Arg5 for engrailed homeodomain-DNA

recognition. In addition to contributing to the overall affinity of the complex, the arm plays a role in determining differential specificity for several classes of homeodomains (Lin & McGinnis, 1992; Zeng et al., 1993; Ekker et al., 1994). The amino acid sequences of the N-terminal arms vary considerably among classes of homeodomain proteins (as defined by Scott et al., 1989), but are conserved among individual members of a class of homeodomain proteins.

Experiments with several homeodomains suggest that residues which appear to position the arm in the minor groove may play a significant role in determining the binding specificity of the arm (Lin & McGinnis, 1992; Zeng et al., 1993; Ekker et al., 1994). The Abdominal-B (Abd-B) homeodomain binds preferentially to the core sequence TTAT rather than TAAT. When three residues in the arm of the ultrabithorax (Ubx) homeodomain, which binds preferentially to a TAAT core, are replaced by those found in Abd-B, the RK3/QK6/TP7 Ubx homeodomain now shows a preference for a TTAT core sequence. Mutation of only Arg3 in Ubx to lysine is not sufficient to alter the binding preference of the protein and additional mutations at position 6 and 7 of the arm are required (Ekker et al., 1994). By analogy to the structure of the engrailed complex, the amino acid at position 3 of Ubx is predicted to contact the DNA directly and the amino acid at position 6 of Ubx is predicted to interact with the DNA backbone (Kissinger et al., 1990). These results imply that residues 6 and 7 could influence specificity by affecting the position of the arm in the minor groove. Do the corresponding amino acids influence binding specificity of the engrailed homeodomain as well? Preliminary experiments with a variant of the engrailed homeodomain in which Thr6 is replaced by alanine show that the backbone contact by Thr6 does not make a substantial contribution to binding affinity (affinity is reduced only 2-3 fold).

Although the affinity is not affected, is the binding specificity reduced?

Would mutating only the amino acids at positions 6 and 7 to those found in Abd-B be sufficient to alter specificity or is it necessary to mutate the amino acid at position 3 to lysine as well?

On a broader level one could ask what sequence determinants in the N-terminal arm of the engrailed homeodomain are necessary for DNA recognition by randomly mutating residues of the arm with oligonucleotide cassette mutagenesis and selecting for variants which are still able to bind to DNA. By randomizing the entire arm, the issue of whether the homeodomain could bind to DNA with significantly different arm sequences could be addressed. Amino acids in the arm could also be randomized individually to assess the functional significance of a particular residue.

The N-terminal arm of the homeodomain can also be used as a model system in which to address issues of minor-groove recognition. As discussed in the first chapter of this thesis, arginine is found to interact with bases in the minor groove more often than any other amino acid. The engrailed homeodomain is no exception, a pair of arginines contact bases in the minor groove (Kissinger et al., 1990). Of the two arginines, Arg5 has the largest contribution to DNA recognition. Arginine is found at position 5 in nearly every non-yeast homeodomain (Scott et al., 1989). Why is this arginine so conserved? Could lysine, which is a hydrogen bond donor, and long and flexible like arginine, substitute for arginine in this interaction? Could altered-specificity mutants that prefer a CAAT core sequence, for example, be generated by changing Arg5 to another amino acid, such as aspartate or glutamate, which can interact with guanines in the minor groove?

## References

- Ades, S. E., & Sauer, R. T. (1994) *Biochemistry* 33, 9187-9194.
- Billeter, M., Qian, Y. Q., Otting, G., Müller, M., Gehring, W., & Wüthrich, K. (1993) *J. Mol. Biol.* 234, 1084-1097.
- Brown, B. M., Milla, M. E., Smith, T. L., & Sauer, R. T. (1994) *Nature Struct. Biol.* 1, 164-168.
- Ekker, S. C., Jackson, D. G., von Kessler, D. P., Sun, B. I., Young, K. E., & Beachy, P. A. (1994) *EMBO J.* 13, 3551-3560.
- Ekker, S. C., Young, K. E., von Kessler, D. P., & Beachy, P. A. (1991) *Embo J.* 10, 1179-86.
- Hanes, S. D., & Brent, R. (1989) *Cell* 57, 1275-83.
- Hanes, S. D., & Brent, R. (1991) *Science* 251, 426-30.
- Hanes, S. D., Riddihough, G., Ish-Horowicz, D., & Brent, R. (1994) *Mol. Cell. Biol.* 14, 3364-3375.
- Jaynes, J. B., & O'Farrell, P. H. (1988) *Nature* 336, 744-749.
- Kissinger, C. R., Liu, B., Martin-Blanco, E., Kornberg, T. B., & Pabo, C. O. (1990) *Cell* 63, 579-590.
- Klemm, J. D., Rould, M. A., Aurora, R., Herr, W., & Pabo, C. O. (1994) *Cell* 77, 21-32.
- Kornberg, T. (1981) *Dev. Biol.* 86, 363-381.
- Laughon, A. (1991) *Biochemistry* 30, 11357-11367.
- Lawrence, P., & Morata, G., (1976) *Dev. Biol.* 50, 321-337.
- Lin, L., & McGinnis, W. (1992) *Genes Dev.* 6, 1071-1081.
- Ohkuma, Y., Horikoshi, M., Roeder, R. G., & Desplan, C. (1990) *Proc. Natl. Acad. U. S. A.* 87, 2289-2293.
- Percival-Smith, A., Müller, M., Affolter, M., & Gehring, W. J. (1990) *Embo J.* 9, 3967-3974.

Poole, S. J., Kauvar, L. M., Drees, B., & Kornberg, T. (1985) *Cell* 40, 37-43.

Scott, M. P., Tamkun, J., & Hartzell, G. W. (1989) *Biochim. Biophys. Acta* 989, 25-48.

Treisman, J., Gönczy, P., Vashishtha, M., Harris, E., & Desplan, C. (1989) *Cell* 59, 553-562.

Treisman, J., Harris, E., Wilson, D., & Desplan, C. (1992) *BioEssays* 14,

Wolberger, C., Vershon, A. K., Liu, B., Johnson, A. D., & Pabo, C. O. (1991) *Cell* 67, 517-528.

Zeng, W., Andrew, D. J., Mathies, L. D., Horner, M. A., & Scott, M. P. (1993) *Development* 118, 339-352.