Differential DNA-Binding Specificity of the Engrailed Homeodomain: The Role of Residue 50†

Sarah E. Ades and Robert T. Sauer*

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

Received April 1, 1994®

ABSTRACT: To assess the importance of residue 50 in determining the binding specificity of the homeodomain from the engrailed transcription factor of *Drosophila*, the DNA-binding properties of isolated homeodomains containing glutamine (wild type), alanine, and lysine at this position have been studied. In binding site selection experiments using the wild-type engrailed homeodomain, TAATTA was identified as a high-affinity, consensus binding site. When the glutamine at position 50 was replaced by a lysine (QK50), the binding site preference changed to TAATCC. The half-life and affinity of the complex between the QK50 protein and a DNA site containing TAATCC were increased significantly compared to the half-life and affinity of the complex between the wild-type protein and a TAATTA site. This suggests that Lys50 forms a more favorable interaction with the TAATCC DNA than Gln50 does with the TAATTA site. In fact, the wild-type Gln50 side chain (which forms a hydrophobic interaction with the last A:T base pair of the TAATTA site in the cocrystal structure [Kissinger, C. R., Liu, B., Martin-Blanco, E., Kornberg, T. B., & Pabo, C. O. (1990) *Cell 63*, 579–590]) appears to play only a small role in determining binding affinity and specificity for the TAATTA site, as the QA50 mutant has only a 2-fold reduced affinity for the TAATTA site and discriminates between the TAATTA and TAATCC sites as well as the wild-type protein. As a result, determinants in addition to Gln50 must be involved in establishing the differential binding specificity of the engrailed homeodomain.

Understanding the determinants of binding specificity is one of the central challenges in the study of protein–DNA interactions. The homeodomain, a 60-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 α-helix (Figure 1). Not surprisingly, the identity or general chemical character of these four amino acids is 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 α-helix 3) and these bases appear to play a key role in determining differential DNA-binding specificity (Hanes & Brent, 1989; 1991; Treisman et al., 1989; Percival-Smith et al., 1990). Thus, when the lysine at position 50 of the *Drosophila* bicoid homeodomain is replaced with a glutamine as found in the 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 cocrystallized 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 60-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*

† Supported by NIH Grant A1-16892.
® Abstract published in *Advance ACS Abstracts,* July 1, 1994.

0006-2960/94/0433-9187$04.50/0

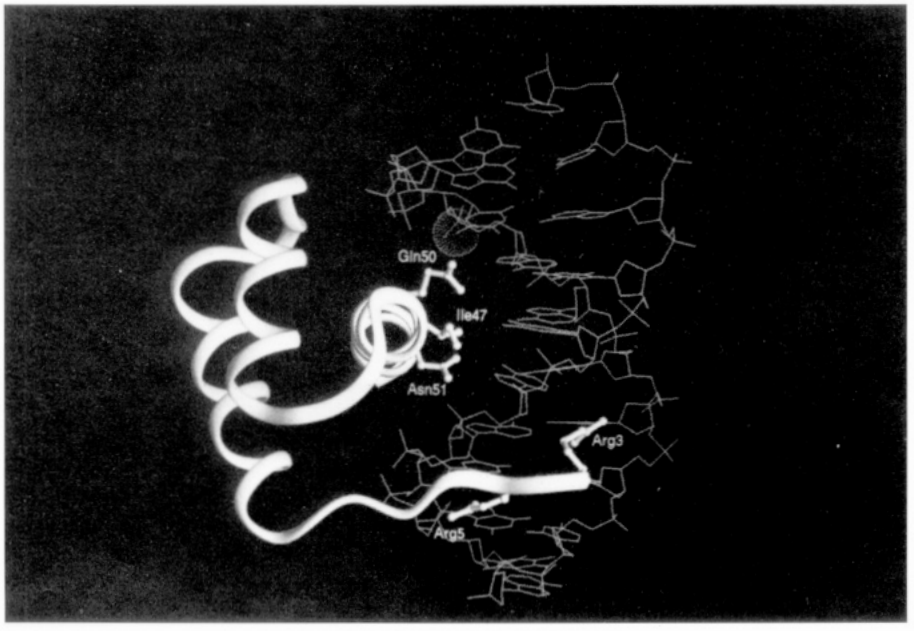


FIGURE 1: Molecular graphics representation of the engrailed homeodomain bound to DNA (Kissinger et al., 1990). The protein backbone is shown as a ribbon trace, and the five side chains that make base contacts are shown in ball-and-stick representation. The thymine methyl group which interacts with Gln50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

coli. The gene was cloned between the *Nde*I and *Cla*I sites of the T7 expression phagemid pAED4 to create the plasmid pSEA100. pAED4 (a gift from Don Doering) contains the pUC19 backbone and the 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 *Bgl*II and *Bss*HII 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/pSEA 100 and X90(DE3)/pSEA 100, respectively. Cells were grown with aeration at 37 °C in 1 L of LB broth plus 150 μg/mL ampicillin to an OD_{600} of 0.7–1.0, and transcription from the T7 promoter was induced by the addition of IPTG to 0.4 mM. After 3 h, 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 $CaCl_2$, 10 mM $MgCl_2$, 2 mM NaN_3, 1.4 mM β-mercaptoethanol, and 50% glycerol]; 10 μL of phenylmethanesulfonyl fluoride (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 supernatent 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_{18} 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 M^{-1} 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 to 300 nm were collected at 20 °C in 1 nm steps with an averaging time of 1 s and averaged over five repeats. Samples contained 25 μg/mL (wild type and QK50) or 18 μg/mL (QA50) protein in 50 mM potassium phosphate (pH 7.0) and 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-labeled with [γ^{32}P]ATP and T4 polynucleotide kinase using standard protocols (Sambrook et al., 1989). After one strand had been end-labeled, the complementary oligonucleotide was added, the mixture was heated to 90 °C, and

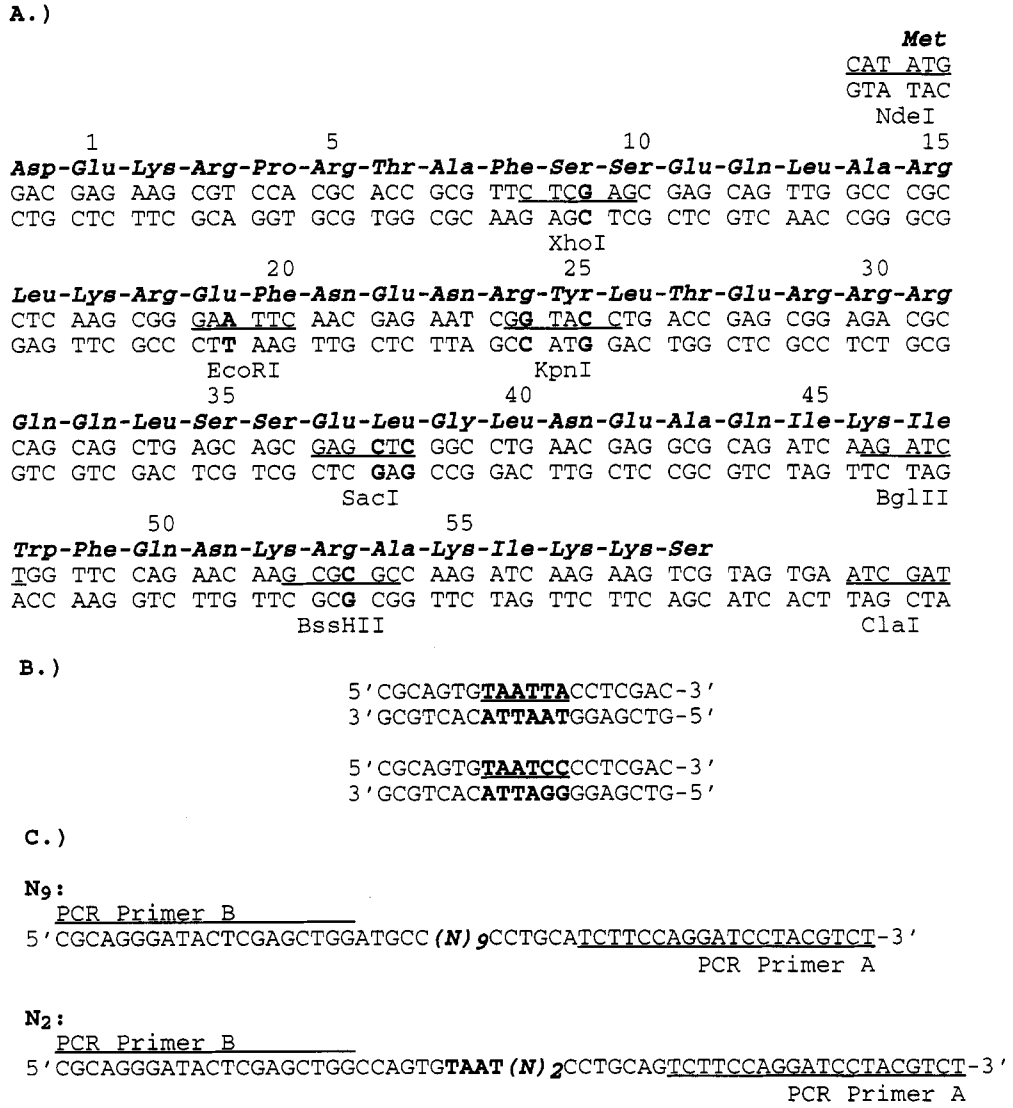


FIGURE 2: (A) Sequence of the gene constructed to encode the engrailed homeodomain. Unique restriction enzyme sites are indicated. The amino acid numbering is according to Qian et al. (1989) to maintain consistency with other homeodomains. (B) Sequences of DNA fragments used for binding assays. (C) Sequences 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.

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, radiolabeled 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 was loaded onto a 10% 0.5X TBE polyacrylamide gel running at 300 V, and after the samples had entered the gel, the voltage was reduced to 155 V. It was necessary to load running gels to obtain consistent results. Prior to loading, gels were prerun for a minimum of 45 min at 300 V. Tracking dyes were not added to the samples but 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 θ, 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 radiolabeled TAATTA site to give roughly 80–90% binding, and aliquots were dispensed into tubes with 0.5 nM–1.0 μM concentrations of unlabeled 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 TAAT-

CC site was determined as a function of θ (the fraction of bound radiolabeled DNA calculated as described above), the equilibrium dissociation constant (K_d) for the radiolabeled binding site, the total concentration of the unlabeled competitor DNA (I_T), and the total concentration of protein (P_T) by fitting data to the following equation:

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

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 $P_T \simeq [P] + [PI]$ (i.e., the concentration of radiolabeled DNA is small compared to P_T and I_T).

To measure dissociation rates, sufficient protein was equilibrated with radiolabeled DNA to give roughly 80–90% binding, unlabeled competitor DNA was added to a final concentration of 0.1 μM, and at different times, 20 μL aliquots were loaded directly onto a 10% 0.5X TBE polyacrylamide gel running at 300 V. 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 θ represents the fraction of DNA bound at time t and θ_0 represents the fraction bound at time zero.

Binding Site Selection. The DNA oligonucleotide N_9 (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_9 and extended for 1 h at 37 °C with sequenase v2.0 (USB) in the presence of unlabeled nucleotides and a small amount of [α-^{32}P]dATP. Unincorporated nucleotides were removed using a G25 Sephadex quick spin column (Boehringer Mannheim), and the labeled duplex DNA was purified on a 10% TBE polyacrylamide gel.

High-affinity binding sites for the engrailed homeodomain were selected using a gel retardation assay. Roughly 0.1 nM of labeled randomized DNA (N_9, Figure 2C) was equilibrated in 50 μL of binding buffer with 0.1 nM, 1 nM, 10 nM, 100 nM, or 1 μM of the engrailed homeodomain for at least 2 h. At this time, 30 μL from each reaction mixture was loaded on a 10% 0.5X TBE 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 $MgCl_2$, 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 of 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 mixture contained 5 mM $MgCl_2$, PCR reaction buffer (Perkin-Elmer Cetus GeneAmp kit), 20 pmol of end-labeled primer A, 20 pmol of primer B, 1 mM dNTP's, and 1 U Amplitaq (Perkin-Elmer Cetus). The reaction mixture was layered with 60 μL of mineral oil and amplified by 20 cycles of 94 °C × 30 s, 55 °C × 30 s, and 72 °C × 40 s followed by a final extension at 72 °C × 10 min using a Perkin-Elmer Cetus Model 480 thermacycler. Amplified DNA was purified on 10% TBE acrylamide 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 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 unlabeled 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_2 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 homeogeneity 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 3 bottom). The CD spectrum at 20 °C (Figure 3 top), where the protein is fully folded, is basically that expected for an α-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% α-helix (assuming a value of −33 000 for 100% helix). We were initially concerned that the aberrant CD spectrum 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 α-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; Chakrabartty et al., 1993).

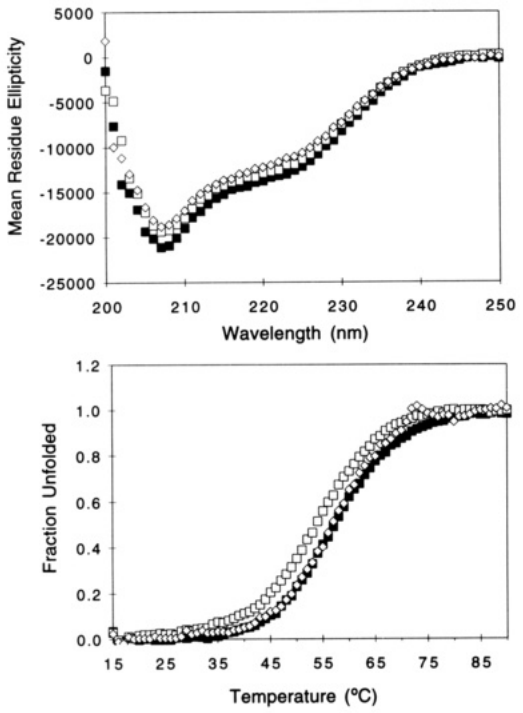


FIGURE 3: (top) CD spectra of the proteins: wild type (3.3 μM, filled boxes), QA50 (2.4 μM, open diamonds), and QK50 (3.3 μM, open boxes). (bottom) Thermal denaturation of the wild-type, QA50, and QK50 proteins (symbols and concentrations as in top panel). Fitting of the denaturation curves using nonlinear least-squares methods yields the following values: wild type, t_m = 55.5 °C, ΔH = 35.9 kcal/mol; QA50, t_m = 56.8 °C, ΔH = 39.0 kcal/mol; and QK50, t_m = 53.8 °C, ΔH = 34.3 kcal/mol.

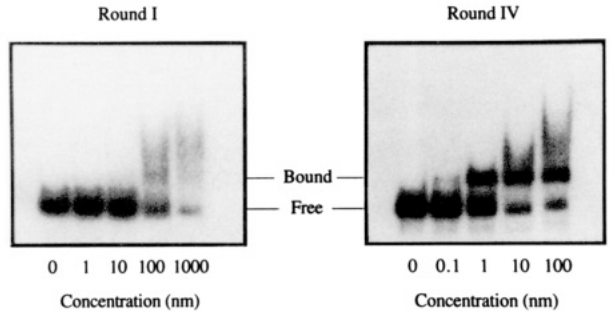
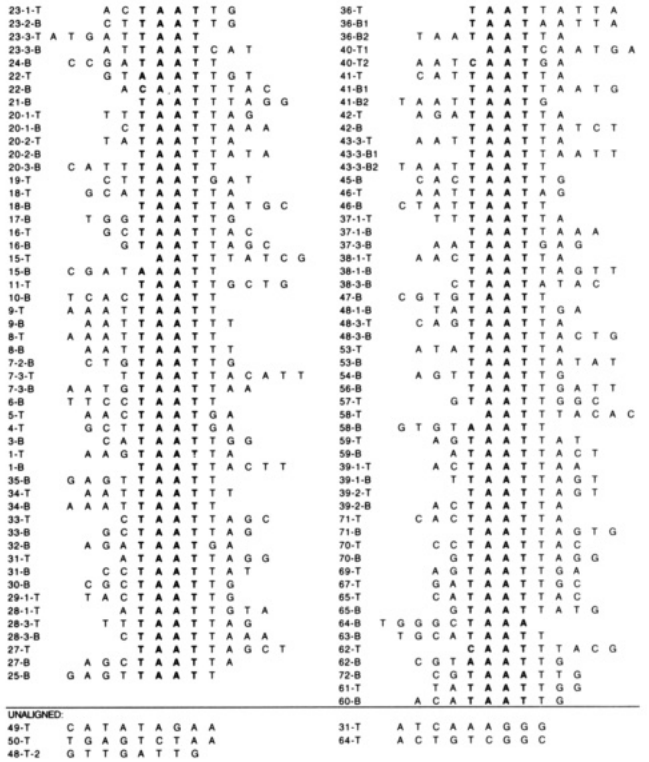


FIGURE 4: Gel mobility shift assays from the first and final rounds of binding site selection for the wild-type engrailed homeodomain. The concentration of the engrailed homeodomain in each lane is indicated. The left-most lane of each gel is a no protein control.

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_9 which contains nine 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 5, top). 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 5, bottom). 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_2, 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.



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

FIGURE 5: (top) 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. (bottom) Tabulation of the aligned data.

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

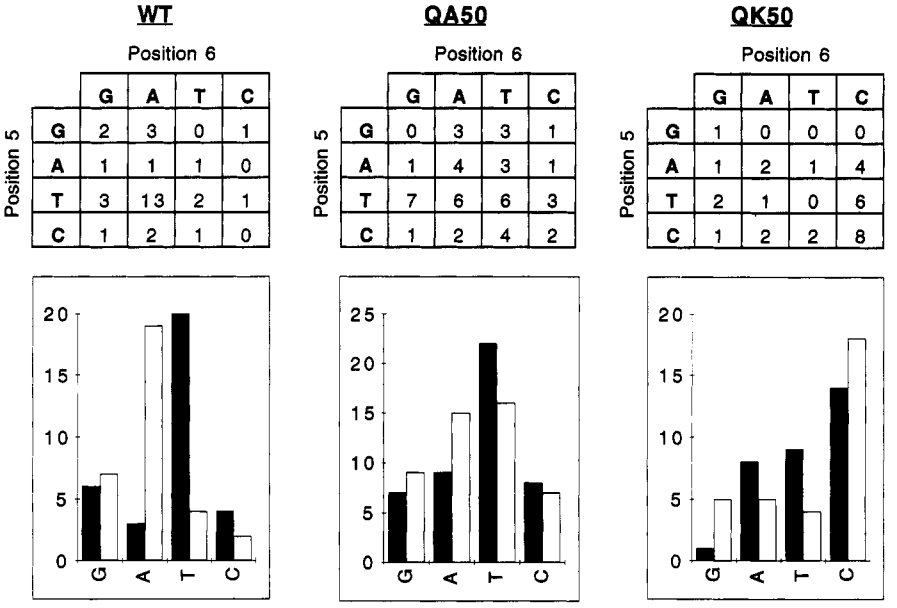


FIGURE 6: Binding site preference at positions 5 and 6 for the wild-type, QA50, and QK50 engrailed homeodomains following *in vitro* selections using TAATNN (N_2). Individual sequences are listed in the tables. The charts present the data as the number of occurrences of each base at position 5 (filled bars) and position 6 (open bars).

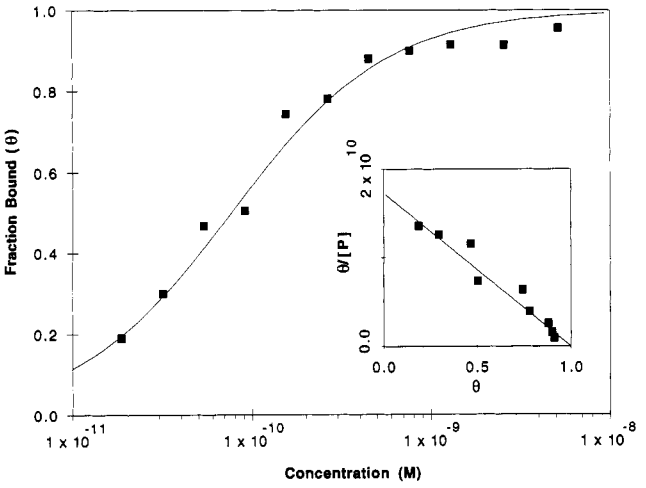


FIGURE 7: Equilibrium binding curve for the wild-type engrailed homeodomain binding to a DNA fragment containing the TAATTA site. The solid line represents a theoretical curve with $K_d = 7 \times 10^{-11}$ M. In the inset, the same data are plotted in Scatchard form.

Table 1: Equilibrium and Kinetic DNA-Binding Constants

protein	binding site	K_d (M)	half-life (s)
WT	TAATTA	$7.9 (\pm 2.3) \times 10^{-11}$	≤2.5
QA50	TAATTA	$1.9 (\pm 0.5) \times 10^{-10}$	<2.5
QK50	TAATTA	$3.2 (\pm 1.6) \times 10^{-10}$	<2.5
WT	TAATCC	$2.1 (\pm 0.8) \times 10^{-9}$	n.d.[a]
QA50	TAATCC	$3.4 (\pm 2.6) \times 10^{-9}$	n.d.[a]
QK50	TAATCC	$8.8 (\pm 4.7) \times 10^{-12}$	289

[a] 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.

50. The CD spectra and thermal denaturation profiles of these mutant proteins were very similar to those of the wild-type engrailed homeodomain (Figure 3), 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 ($k_d \simeq 0.28$ 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 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 7.3×10^{-11} M (Figure 7). This suggests that the association reaction must be close to the diffusion limit for bimolecular reactions (calculated $k_a \simeq 3 \times 10^9$ M^{-1} s^{-1}). The equilibrium binding of the wild-type engrailed homeodomain to the TAATCC fragment ($K_d \simeq 2.1 \times 10^{-9}$ 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 the wild type and, like the wild type, shows significantly reduced binding to the TAATCC site.

The QK50 protein binds to the TAATTA site only 4-fold less well than the 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

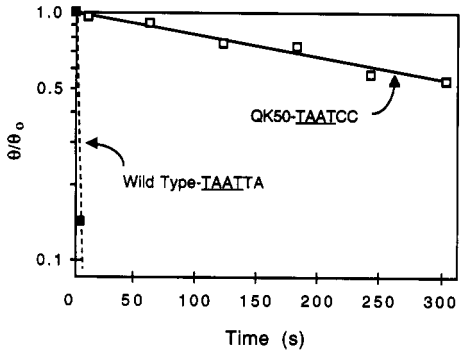


FIGURE 8: Dissociation kinetics of complexes of the wild-type engrailed homeodomain and the TAATTA site (filled boxes) and the QK50 homeodomain and the TAATCC site (open boxes). Time is measured from the addition of cold competitor DNA. The half-life of the wild-type complex is ~2.5 s, and the half-life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

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/mol. 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 Cβ 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/mol to this discrimination: both the wild-type and QA50 homeodomains bind 17–25-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.

ACKNOWLEDGMENT

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*, 2120–2134.

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