Engrailed (Gln50→Lys) homeodomain–DNA complex at 1.9 Å resolution: structural basis for enhanced affinity and altered specificity

Lisa Tucker-Kellogg[1], Mark A Rould[2,3], Kristen A Chambers[2,3], Sarah E Ades[3], Robert T Sauer[3] and Carl O Pabo[2,3]*

Background: The homeodomain is one of the key DNA-binding motifs used in eukaryotic gene regulation, and homeodomain proteins play critical roles in development. The residue at position 50 of many homeodomains appears to determine the differential DNA-binding specificity, helping to distinguish among binding sites of the form TAATNN. However, the precise role(s) of residue 50 in the differential recognition of alternative sites has not been clear. None of the previously determined structures of homeodomain–DNA complexes has shown evidence for a stable hydrogen bond between residue 50 and a base, and there has been much discussion, based in part on NMR studies, about the potential importance of water-mediated contacts. This study was initiated to help clarify some of these issues.

Results: The crystal structure of a complex containing the engrailed Gln50→Lys variant (QK50) with its optimal binding site TAATCC (versus TAATTA for the wild-type protein) has been determined at 1.9 Å resolution. The overall structure of the QK50 variant is very similar to that of the wild-type complex, but the sidechain of Lys50 projects directly into the major groove and makes several hydrogen bonds to the O6 and N7 atoms of the guanines at base pairs 5 and 6. Lys50 also makes an additional water-mediated contact with the guanine at base pair 5 and has an alternative conformation that allows a hydrogen bond with the O4 of the thymine at base pair 4.

Conclusions: The structural context provided by the folding and docking of the engrailed homeodomain allows Lys50 to make remarkably favorable contacts with the guanines at base pairs 5 and 6 of the binding site. Although many different residues occur at position 50 in different homeodomains, and although numerous position 50 variants have been constructed, the most striking examples of altered specificity usually involve introducing or removing a lysine sidechain from position 50. This high-resolution structure also confirms the critical role of Asn51 in homeodomain–DNA recognition and further clarifies the roles of water molecules near residues 50 and 51.

Addresses: [1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA, [2]Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, MA 02139, USA and [3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

*Corresponding author.
E-mail: pabo@mit.edu

Key words: altered-specificity mutation, DNA binding, homeodomain, protein–DNA interactions, X-ray crystallography

Received: 12 May 1997
Revisions requested: 5 June 1997
Revisions received: 7 July 1997
Accepted: 9 July 1997

Structure 15 August 1997, 5:1047–1054
http://biomednet.com/elecref/0969212600501047

Introduction

Altered-specificity variants can provide powerful tools for studying protein–DNA recognition. The homeodomain, one of the key DNA-binding motifs used in eukaryotic gene regulation, provides a very attractive system for this type of analysis: hundreds of related homeodomain sequences are known, and there is a wealth of relevant biochemical and structural data [1]. Biochemical and genetic studies indicate that residue 50 is especially important in determining the differential specificity of homeodomain–DNA recognition [2–5], playing a role in distinguishing between binding sites of the form TAATNN. Glutamine is the most common residue at position 50, but cysteine, serine and lysine occur in other subfamilies [1]. The tightest and most specific binding occurs when lysine is present at position 50. Biochemical studies of an engrailed Gln50→Lys variant (QK50) revealed that QK50 actually binds more tightly to TAATCC than wild-type engrailed binds to TAATTA (Table 1) [6]. We have pursued structural studies of this Lys50 variant to understand how it forms such a stable complex, to elucidate the role of position 50 in homeodomain–DNA recognition, and — more generally — to explore the structural requirements for designing altered-specificity variants. We find that the Lys50 sidechain projects directly into the major groove of the DNA and makes a set of hydrogen bonds with the

Table 1

Equilibrium dissociation constants (in nM) for the complexes with the wild type engrailed homeodomain, the QA50 variant, and the QK50 variant.

Version of engrailed	DNA site*	
	TAATTA	TAATCC
Wild type	0.079	21
QA50	0.19	3.4
QK50	0.32	0.0088

*Only one strand of the DNA subsite is indicated; binding studies used 20 bp duplex DNA sites.

guanines at base pairs (bps) 5 and 6 of the optimal TAATCC binding site. The Lys50 sidechain and these new contacts are accommodated without requiring any major changes in the overall architecture of the homeodomain–DNA complex.

Results

We have crystallized the engrailed QK50 variant in complex with the duplex TAATCC binding site, and have solved this structure at 1.9 Å resolution. The DNA duplex used for cocrystallization is homologous to that used in studies of the wild-type complex and gave cocrystals that are nearly isomorphous to the wild-type cocrystals studied by Kissinger *et al.* [7]. Our refined model (using data collected at –150°C) has a free R factor of 25.1% and a conventional R factor of 20.5% for data from 6.0–1.9 Å resolution (Table 2). The overall structure of the QK50 complex is very similar to that of the wild type engrailed complex (studied at 2.8 Å resolution at room temperature): alignment of the complexes by superimposing Cα atoms of the homeodomain (residues 10–55) and P and C1′ atoms of the TAATNN subsites (i.e. superimposing 24 atoms of each DNA duplex) gives a root mean square (rms) deviation of 0.48 Å, and confirms that the folding and docking are exceedingly similar. As in the wild-type complex, the homeodomain folds as a globular domain with three α helices, and helix 3, the 'recognition' helix, fits into the major groove of the DNA. An extended N-terminal arm contacts the minor groove. Given that the wild-type and QK50 complexes are so similar, we focus our attention on Lys50 and Asn51. These residues are critical for site-specific recognition and have been the focus of much discussion when comparing NMR and crystal structures of homeodomain–DNA complexes [7–14].

Interactions between Lys50 and the DNA

Figure 1 shows the electron density for Lys50 from the solvent-flattened MIRAS (multiple isomorphous replacement with anomalous scattering) map (using data collected at 10°C) and also shows the final refined model (using data collected at –150°C). The overall placement of the lysine sidechain is exceedingly clear: Lys50 projects into the major groove towards the guanines of bps 5 and 6, and many of the key contacts involve hydrogen bonds to the O6 and N7 atoms of these guanines (Figures 1–4). The high resolution of our structure determination allows us to see and refine alternate conformations for the terminal atoms of Lys50. Conformation 1 (60% occupancy) places the terminal NH_3^+ group near the guanines of bps 5 and 6. In this conformation, the closest contacts involve the O6 of the guanine at bp 5 (2.76 Å) and the O6 and N7 of the guanine at bp 6 (3.27 and 3.17 Å, respectively). The N7 of the guanine at bp 5 is slightly farther away (3.92 Å), but there is a bridging water molecule that contacts the N7 of this guanine (2.88 Å) and the terminal NH_3^+ of the Lys50 sidechain (3.17 Å). Conformation 2 (40% occupancy) only moves the terminal Nζ of the lysine by 1.42 Å but the altered sidechain dihedral points the NH_3^+ somewhat more towards the guanine at bp 5 and the thymine at bp 4. In this conformation, the closest contacts involve the O6 of the guanine at bp 5 (2.78 Å) and the O4 of the thymine at bp 4 (3.04 Å). The N7 of the guanine at bp 5 is 4.11 Å away, but there are good contacts from the bridging water molecule (which in this conformation is 3.17 Å from the Nζ and 2.88 Å from the N7 of the guanine).

The role of Asn51

This high resolution cocrystal structure also provides important information about the role of Asn51 in homeodomain–DNA recognition. The structure of the wild type engrailed complex at 2.8 Å resolution [7] indicated that Asn51 makes a pair of hydrogen bonds with the adenine at bp 3 of the TAATTA site, and similar contacts were seen

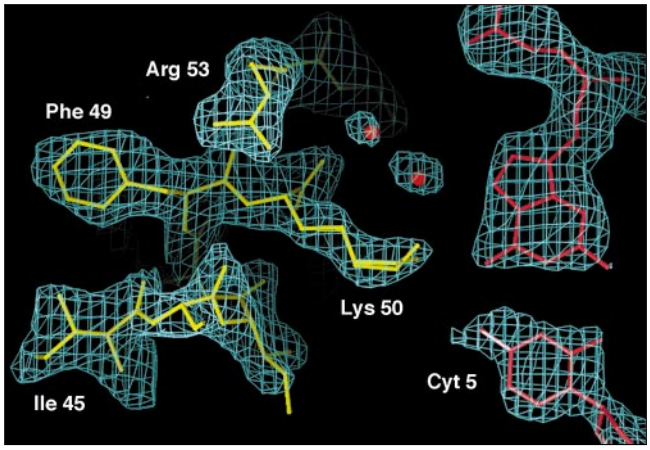
Table 2

Data collection (–150°C) and refinement statistics.

Data collection	
Resolution (Å)	1.9
Measured reflections	92,610
Unique reflections	27,136
Completeness to 1.9 Å (%)	94
Completeness in 1.97–1.90 Å shell (%)	84
R_{merge}* (%)	3.5
Refinement	
R factor† (%)	20.5
R_{free}† (%)	25.1
Rms deviation of bond lengths (Å)	0.011
Rms deviation of bond angles (°)	1.46
Number of nonhydrogen atoms	1,990
Number of water molecules	183
Rms ΔB (Å²)‡	2.32

*$R_{merge} = \Sigma | I - \langle I \rangle | / \Sigma I$, where I = observed intensity and $\langle I \rangle$ = average intensity of multiple observations of symmetry-related reflections. †The R factors exclude 2564 reflections for which $F < 2\sigma(F)$. Using all data from 6.0–1.9 Å, the R factor is 21.7% and the R_{free} is 26.5%. ‡Rms ΔB is the root mean squared difference between temperature factors of covalently bonded atoms.

Figure 1



Solvent-flattened MIRAS electron-density map contoured at 1.5σ, in the vicinity of Lys50 and bp 5 of the TAATCC subsite. The model is our final refined low-temperature structure, but a rigid-body motion has been used to adjust for differences in cell dimensions (between the 10°C map and the –150°C refined structure). The protein is shown in yellow and the DNA in red. The two conformations of Lys50 are essentially superimposed when seen from this orientation.

in the α2 homeodomain–DNA complex [8]. There has been much discussion about these contacts, since NMR studies of the Antp–DNA complex have suggested that Asn51 has multiple, rapidly-interchanging conformations and have indicated that Asn51 might make water-mediated contacts with the bases [9,10]. Our structure of the QK50 variant confirms that Asn51 forms a pair of direct hydrogen bonds with the adenine at bp 3, with distances of 3.04 Å to the N7 atom and 3.09 Å to N6. The conformation and contacts that we observe for Asn51 are in excellent agreement with those observed in studies of the Oct-1 [11], paired [12], a1/α2 [13], and even-skipped [14] homeodomain–DNA complexes. The remarkable consistency of these structures (determined independently and in different crystal forms) suggests that these crystallographic studies are giving the correct (time-averaged) conformation of Asn51. The fact that Asn51 is so strictly conserved among the hundreds of known homeodomains (see [1] for a review) and the fact that homeodomain binding sites almost invariably have adenine at bp 3, suggests that the Asn51–adenine contacts may be similar in all homeodomain–DNA complexes.

Water molecules in the binding interface

In addition to the direct hydrogen bonds made with the adenine at bp 3, Asn51 is flanked by several well-ordered water molecules at the protein–DNA interface. Perhaps the most striking interaction involves a water molecule that bridges from the Oδ1 of Asn51 to the N6 of the adenine at bp 4 (Figures 2 and 4). This water molecule has excellent hydrogen-bonding geometry, with distances of 3.01 Å to the Oδ1 and 3.11 Å to the N6 atom. This water molecule also bridges to a second water molecule which, in turn, contacts the N7 of the adenine at bp 4 (Figures 2 and 4). The distance between these two water molecules is 3.15 Å, and the distance from the second water molecule to the N7 is 3.08 Å. A third water molecule in this hydrogen-bonding network contacts the N7 of the guanine at bp 5

We note that there is a 'tilt' in the Asn51 sidechain amide that allows Asn51 to maintain good hydrogen bonds with the adenine and yet also allows it to interact well with the bridging water molecule. (This tilt involves a rotation around χ_2, and the observed χ_2 angle (–37°) in our complex

Figure 2

Stereo diagram showing base contacts made by Lys50, Asn51 and associated water molecules in the QK50–TAATCC cocrystals. Both conformations of the Lys50 sidechain are shown, and the three key water molecules are represented as black spheres; hydrogen bonds are shown as dashed lines. The numbering scheme for base pairs corresponds to that used in Figures 3 and 4. A Cα trace is shown for part of helix 3.

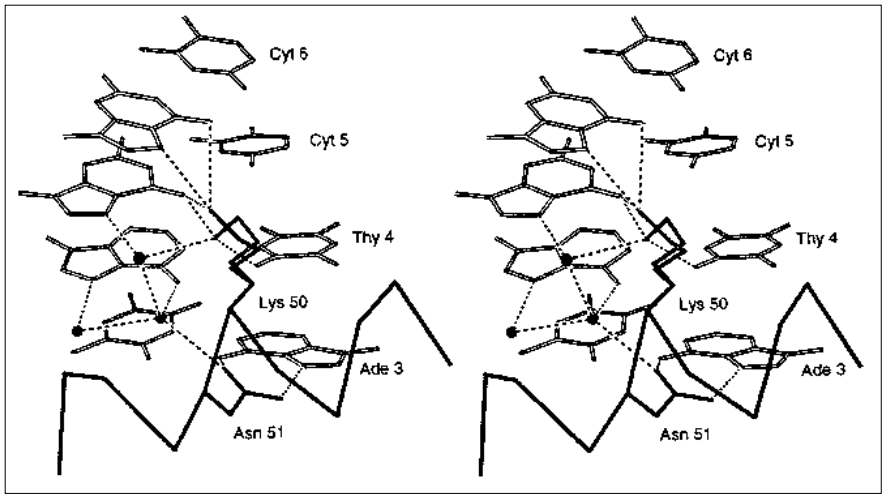
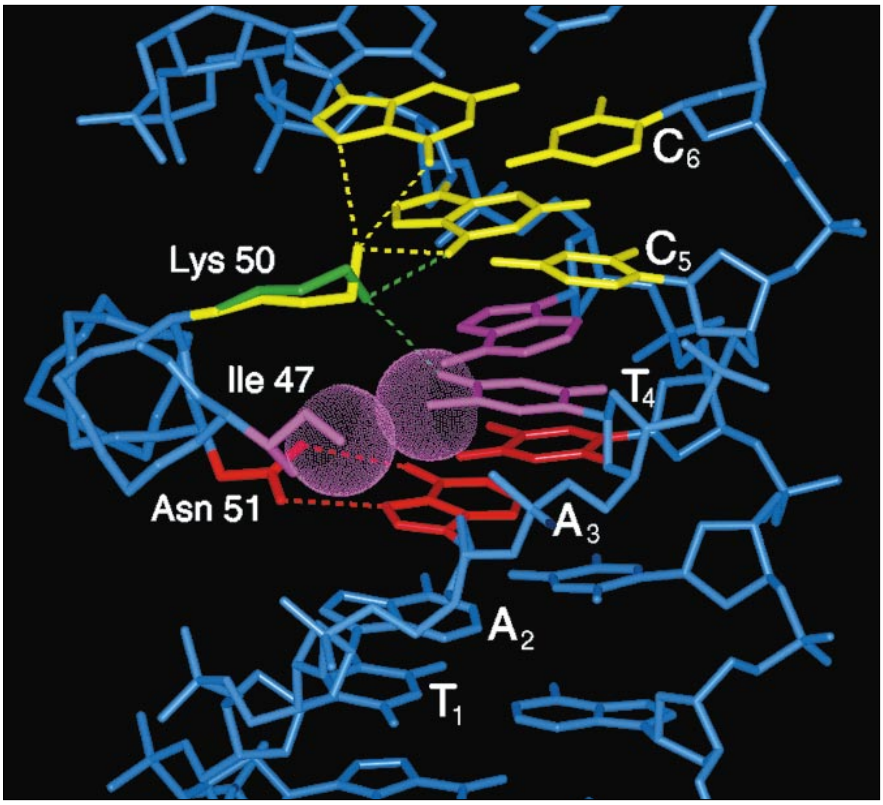


Figure 3



Major groove contacts of the QK50–TAATCC complex. Three residues make base contacts in the major groove: Asn51 makes a pair of hydrogen bonds with the adenine at bp 3 (red); Ile47 makes hydrophobic contacts with the methyl group of the thymine at bp 4 (purple); the primary conformation of Lys50 (yellow) makes hydrogen bonds with the O6 of the guanine at bp 5 and with the O6 and N7 atoms of the guanine at bp 6; the secondary conformation of Lys50 (green) makes hydrogen bonds with the O6 of the guanine at bp 5 and with the O4 of the thymine at bp 4. Hydrogen bonds are shown as dashed lines and van der Waals contacts are indicated with dotted spheres. For clarity, water molecules have been omitted in this figure.

leaves the amide plane 27° out of the plane of the adenine.) The observed tilt of the Asn51 sidechain underscores the potential importance of the bridging water molecule. It seems quite plausible — as suggested by Wilson *et al.* [15] — that the water-mediated contacts with the adenine at bp 4 may augment the sequence specificity provided by the Ile47–thymine contact (Figure 3) and thus may help explain the preference of the homeodomain for the canonical TAAT subsite. Examining other refined structures of homeodomain–DNA complexes shows similar tilt angles for the Asn51 sidechains. Remarkably, we note that a conceptually analogous bridging interaction also occurs in the α2 portion of the a1/α2 homeodomain–DNA complex: here the terminal atoms of the Arg54 sidechain, rather than a water, participate in a hydrogen-bonding network that bridges from the Oδ1 of Asn51 (2.99 Å) to the O6 of the guanine at bp 4 (3.09 Å).

Discussion

The role of position 50

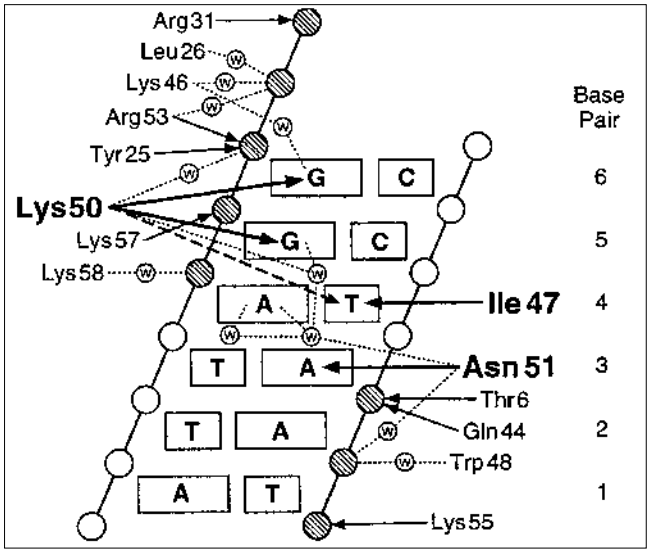
As emphasized in the early biochemical studies [2–5], residue 50 plays a key role in determining the differential specificity of the homeodomain, helping to explain how homeodomains can distinguish one TAATNN site from another. Correlating all the available data, however, highlights the fact that different residues at position 50 confer very different degrees of specificity for their respective sites. Reviewing the earlier papers [2–5], shows that the most striking cases of altered-specificity mutations usually involve introducing or removing a lysine residue from position 50: Key constructs are a Lys50→Gln variant of the bicoid homeodomain [3], a Ser50→Lys variant of the paired homeodomain [4], and a Gln50→Lys mutation in the *fushi tarazu* homeodomain [5]. In every case, the Lys50 variants of these homeodomains prefer to bind a sequence of the form TAATCC and they presumably all make contacts similar to those seen in the crystal structure reported here.

There are other cases in which the sidechains at position 50 only have a marginal role in determining DNA-binding specificity. For example, the Oct-1 homeodomain, with a cysteine at position 50, shows sequence specificity for a 4 bp subsite (typically with the sequence AAAT) but shows little specificity at positions that would correspond to bp 5 and 6 of our current numbering scheme. Changing this cysteine to glutamine has little effect on DNA-binding affinity [16,17]. Even when glutamine, which is one of the most common residues at position 50, occurs in the wild-type proteins it may have only a modest energetic contribution to binding: wild-type engrailed prefers a TAATTA site, but a variant (QA50) which has alanine at position 50 binds the

TAATTA site only 2.4-fold less strongly than the wild-type protein (Table 1) [6]. This modest energetic contribution of the Gln50 sidechain ($\Delta\Delta G$ = 0.5 kcal/mol) is fully consistent with the 2.8 Å resolution crystal structure of the wild type engrailed complex, in which the only direct contact between Gln50 and the DNA bases is a van der Waals contact with the methyl group of the thymine at bp 6.

When placed at position 50, lysine seems to have a clearer sequence specificity than other residues tested and makes a greater energetic contribution to binding. Thus, comparing the QK50 and QA50 variants of engrailed shows that changing Lys50 to alanine gives a 390-fold reduction in affinity for the TAATCC sequence (Table 1; $\Delta\Delta G$ = 3.4 kcal/mol) [6]. The numerous direct contacts made by Lys50 in our QK50 structure, and the way that these lysine–guanine contacts fit so well within the conserved structure of the complex, provide a simple explanation for the efficacy of this residue in site-specific recognition. Remarkably, these are the first direct hydrogen-bonding contacts reported for residue 50 in any homeodomain–DNA complex.

Figure 4



Sketch of major groove contacts in the QK50–TAATCC complex. Residues that make base contacts in the major groove are shown in boldface. Phosphates are represented with circles, and hatched circles mark phosphates that are contacted by the homeodomain. Arrows represent direct protein–DNA contacts. Small circles marked 'W' denote water molecules, and finely-dotted lines represent water-mediated contacts. The numbering scheme for the base pairs corresponds to that used in Figures 2 and 3; base pairs are numbered to represent a typical homeodomain binding site in the form TAATNN.

Base-specific electrostatic interactions

Our structure of the QK50–DNA complex also highlights the role that 'electrostatic readout' of the major groove may play in site-specific recognition. Recent surveys of sidechain–base interactions in the known protein–DNA complexes show that arginine–guanine and lysine–guanine interactions are remarkably common [18,19]. As hydrogen bonds involving one charged partner can be very strong [20], and because the N7 of guanine is the most electronegative region in the major groove [21], such contacts may make a major contribution to site-specific recognition. (A key lysine residue in the N-terminal arm of λ repressor also forms hydrogen bonds with a pair of guanines in the major groove [22,23].) We presume that the binding affinity of the QK50 variant reflects both the intrinsic affinity of these lysine–guanine interactions and the very favorable structural framework (provided by the rest of the homeodomain–DNA complex), which holds lysine in an ideal position for making these contacts.

Prospects for other altered-specificity variants

One of the underlying structural issues in protein–DNA recognition — and one that is especially important when thinking about altered-specificity variants — involves the complex interrelationship between the overall folding and docking arrangement of a protein and the geometric requirements for particular sidechain–base interactions (COP and L Nekludova, unpublished data). To what extent does the overall folding and docking determine which sidechain–base interactions are geometrically plausible at a given position? How often will a strictly local substitution, such as the Gln50→Lys change in engrailed, allow new DNA contacts that are as favorable or more favorable than the wild-type contacts?

In thinking about these issues, it is interesting to re-examine the role of Gln50 in the engrailed homeodomain. After seeing the important role that glutamine–adenine contacts play in other protein–DNA complexes, one might have imagined that glutamine would make a pair of hydrogen bonds with one of the adenines in the preferred TAATTA binding site. However, (as mentioned above) the only direct contact between Gln50 and the DNA is a van der Waals contact with the methyl group of the thymine at bp 6, and changing Gln50 to alanine only gives a modest (0.5 kcal/mol) reduction in affinity for the TAATTA site.

Modeling studies readily confirm the problems that would be involved in trying to make canonical glutamine–adenine contacts from position 50 of the homeodomain, and modeling thus helps explain why these contacts do not occur in the wild-type complex. Surveying known protein–DNA complexes shows that the most favorable glutamine–adenine interactions (such as those seen in the λ repressor [23,24] and the 434 repressor [25] complexes) involve a pair of hydrogen bonds between the sidechain and the base, and in these situations the terminal atoms of the sidechain (Cγ, Cδ, Nε and Oε) are roughly coplanar with the adenine base. While keeping the protein and DNA backbones fixed in the conformation of our QK50 crystal structure, we attempted to superimpose the same

'canonical' glutamine–adenine contacts seen in the phage repressors onto residue 50 and bps 5 or 6 of the engrailed complex (L Nekludova and COP, unpublished data). Regardless of what sidechain χ angles are used during modeling, there is no way that canonical glutamine–adenine contacts can fit into the structural context provided by the homeodomain. The position and orientation of the polypeptide backbone at position 50 (vis-à-vis the DNA) provides an ideal geometric arrangement for the lysine–guanine interactions, but simply does not work as well for optimizing potential glutamine–adenine interactions.

Other studies have revealed similar limitations in the design and selection of DNA-binding proteins with altered specificity. For example, biochemical and genetic studies involving systematic variation in position 51 of the Oct-1 homeodomain and in bp 3 of the AAAT binding site failed to reveal any other sidechain–base combination that would work as well as the wild-type Asn51–adenine arrangement [26]. Glutamine was particularly disruptive when placed at position 51 (causing a 1,100-fold reduction in binding to the AAAT subsite), and it appears that the structural context provided by the rest of the homeodomain–DNA complex plays a critical role in determining which sidechain–base interactions will be possible at any given position. The basic idea is very simple and yet has broad implications for our understanding of protein–DNA recognition: given the distinct sizes, shapes, and conformational preferences of the sidechains, only one or two may fit well at a given position in a complex. The overall folding and docking arrangement of the protein (and the overall structure of the DNA) will help to determine which contacts are possible.

Conclusions

These structural and biophysical studies of the QK50 variant provide an interesting perspective on current studies of protein–DNA recognition. The analysis of multiple conformations and of water-mediated contacts has some meaningful role in the understanding of homeodomain–DNA interactions, but we find that a lysine variant which can make direct hydrogen bonds with the DNA bases binds more tightly and specifically than the wild-type engrailed homeodomain. The crystal structure of this altered-specificity complex shows there is nothing mysterious about the tight binding: the homeodomain presents the Lys50 sidechain in a very favorable geometric and structural context (fixed by the conserved folding and docking arrangement of the homeodomain), and the lysine can make a set of direct, sequence-specific hydrogen bonds with the O6 and N7 groups of the guanines. (There is also a water-mediated contact and an alternative conformation of the terminal atoms that allows a hydrogen bond with the O4 of a thymine.) Our structure gives a satisfying explanation for the affinity and specificity of the lysine QK50 variant, but challenging problems remain as we try to understand the limits of altered-specificity variants. How often will such favorable substitutions be possible? How do the overall folding and docking arrangements help to determine what sidechain–base interactions will be possible at a given position?

Biological implications

Homeodomains are one of the most important eukaryotic DNA-binding motifs, and they occur in many transcription factors that control differentiation and determine cell fate. Homeodomains contain 60 amino acids, which fold to form a module with three α helices and an extended N-terminal arm. Homeodomain–DNA interactions have been studied intensively both because of the intrinsic importance of the homeodomain, and because the homeodomain has become a paradigm for the analysis of protein–DNA interactions. Previous structural studies have shown that helix 3, the 'recognition' helix, docks into the major groove and makes many of the base contacts. Biochemical and genetic studies have suggested that residue 50 of the homeodomain is especially important for differential recognition, distinguishing among sites of the form TAATNN. However, none of the previously determined structures of homeodomain–DNA complexes has provided evidence for a stable hydrogen bond between residue 50 and a base, and there has been much discussion about the potential significance of water-mediated contacts in homeodomain–DNA recognition.

Biochemical data, showing that a Gln50→Lys (QK50) variant of the engrailed homeodomain has very high affinity and specificity for a TAATCC site, motivated solving the structure of this complex, and we find a set of very favorable Lys50–guanine contacts that readily explain the biochemical data. The QK50–DNA structure also confirms the conserved docking arrangement of the homeodomain and the critical Asn51–adenine contacts seen in the crystal structures of other homeodomain–DNA complexes. The fact that there is a rigidly conserved docking arrangement may help explain why other sidechains (including the wild-type glutamine) cannot make such energetically favorable contacts from position 50. More generally, our analysis suggests limits (only certain sidechains will fit at certain positions) that may occur in the design and selection of altered-specificity DNA-binding mutants. Finally, our data suggest that direct sidechain–base interactions, when geometrically compatible with the other contacts in a complex, can provide greater affinity and specificity than water-mediated contacts.

Materials and methods

Protein expression and purification

The engrailed QK50 domain used in these studies contains 60 amino acids from the *Drosophila* engrailed protein, but the glutamine residue at position 50 of the wild-type homeodomain is replaced by lysine and an N-terminal methionine is introduced in cloning. (Thus the sequence of our peptide is the same as that shown in Figure 1 of reference [7], except that lysine is present at position 50.) The QK50 variant was

Table 3

MIRAS phasing statistics (data collected at 10°C).

	Native	Derivative 1	Derivative 2	Derivative 3	Derivative 4
Iodinated bases*	–	T2, T8', C17	T2, T8', C18	T2, C17	T2, C18
R_{merge} (%)	4.4	4.7	5.5	4.8	5.0
R_{cross} on I (%)	–	18.3	27.6	16.6	25.5
R_{cross} on F (%)	–	13.1	20.2	12.3	18.7
Completeness to 2.25 Å (%)	90	88	76	99	90
Completeness to 2.0 Å (%)	76	79	65	86	81
R_{cullis}	–	0.495	0.551	0.529	0.564
Phasing power	–	3.25	2.59	3.01	2.40

*Numbering scheme for bases corresponds to that used in Figure 1b of reference [7]. R_{cross} on I = sum($|I_N - I_D|$) / sum(I_N). R_{cross} on F = sum($|F_N - F_D|$) / sum(F_N). R_{Cullis} = sum($||F_{PH} +/- F_P| - |F_{Hcalc}||$) / sum($|F_{PH} +/- F_P|$), for centric reflections only. Phasing power = $(\text{sum}(F_{Hcalc}^2) / \text{sum}((|F_{PH}| - |F_{PHcalc}|)^2))^{0.5}$.

expressed in *Escherichia coli* strain BL21 cells containing the DE3 plasmid [18]. Cultures were induced for 2.5 h with 0.3 mM isopropyl-β-D-thiogalactopyranoside (IPTG) at 37°C. Soluble protein was purified using ion exchange and reverse phase chromatography, and the purity of the peptide was confirmed by gel electrophoresis, mass spectroscopy, amino acid analysis, and protein sequencing (William Lane, Harvard Microchemistry Facility).

DNA complex formation and crystallization

The complex was formed in 1 M ammonium acetate (to keep it soluble), with a 2:1 molar ratio of QK50 peptide to duplex DNA. The DNA used for cocrystallization,

5′-T T T T G C C A T G T A A T C C C C G G A
A A A C G G T A C A T T A G G G G C C T A-5′

contains one TAATCC subsite, and when these DNA duplexes stack in the crystal, a related subsite with the sequence AAATCC is formed by the juxtaposed DNA duplexes. Crystals of the QK50–DNA complex were grown using the hanging-drop vapor diffusion method [27]. Well buffer contained 0.73–0.80 M ammonium acetate (pH 8.0) and 1% PEG 400. The best crystals grew in three days at room temperature from a 2 μl hanging drop containing the complex at a concentration of 10 mg/ml. Note that these conditions are somewhat different from those used for crystallizing the wild-type complex, which had been studied at 2.8 Å resolution [7]. As with the wild-type crystals, the QK50–DNA complex crystals form in space group C2, but have cell parameters of a = 129.9 Å, b = 45.45 Å, c = 72.75 Å, β = 118.7° at 10°C. (The wild-type crystals have a = 131.2 Å, b = 45.5 Å, c = 72.9 Å, β=119.0° at room temperature [7].) Under cryo conditions (–150°C), the cell parameters of the QK50–DNA crystals are a = 127.7 Å, b = 45.3 Å, c = 72.5 Å, β = 119.5°. As expected from studies of the wild-type complex, the asymmetric unit of the QK50 crystals contains one DNA duplex, one homeodomain at the TAATCC site, and a second homeodomain at the AAATCC site that is formed by juxtaposed duplexes.

Phasing and refinement

Structure determination of the QK50–DNA cocrystals proceeded in two stages: MIRAS was used for initial phasing and model refinement at 2.0 Å resolution using data collected at 10°C; and final refinement to 1.9 Å used data collected under cryo conditions. For MIRAS phasing, data were collected on a Rigaku R-Axis IIC detector equipped with Yale/MSC mirrors. Two crystals of the native and of each double- or triple-iodinated DNA derivative were used and data were processed with DENZO/SCALEPACK (written by Z Otwinowski and W Minor and distributed by Molecular Structure Corp.). A constant temperature of 10°C was maintained at the crystal with an FTS AirJet crystal cooling system. Derivatives were local scaled to the native using MAXSCALE [28]. Cross-phased heavy-atom refinement (each derivative is refined separately using phases derived only from the other three derivatives) was carried out with the program PHARE [29]. Solvent flattening [30] was used to improve the phases. Heavy-atom parameters were then re-refined to convergence using the solvent-flattened phases as parent phases (without updating the phases during refinement), and new MIRAS phases were recalculated [31]. This process of refining the heavy-atom parameters using the solvent-flattened phases, recalculating MIRAS phases, and solvent flattening was repeated four times to give the final electron-density map, free of any model bias (Figure 1; Tables 3 and 4). The 2.8 Å model for the wild-type complex [7], with the appropriate changes in amino acid and nucleotide sequences, was rebuilt into this density using TOM/FRODO [32,33] and refined to 2.0 Å with XPLOR [34]. Higher resolution data, extending to 1.9 Å, were obtained

Table 4

MIRAS figure of merit versus resolution.

Resolution shell (Å)	9.4	6.2	4.6	3.6	3.0	2.6	2.2	2.0	20–2.0
Figure of merit	0.93	0.95	0.93	0.89	0.86	0.79	0.65	0.46	0.73*

*Subsequent solvent flattening increases the mean figure of merit to 0.84.

using cryocrystallographic methods. Crystals were cryoprotected by adding glycerol to the hanging drop, with a final concentration of 30% (v/v) immediately prior to flash cooling. Data were collected at –150°C and processed as before. In rebuilding to the cryo data, rigid-body refinement was used to adjust for the differences in cell parameters. Throughout refinement we made repeated use of simulated annealing omit maps and monitored the free R factor to avoid overfitting the experimental data. The same free R list was used for the 10°C data and the –150°C data. Local scaling was used to correct for absorption errors and anisotropic diffraction. The two monomers were refined independently and have almost identical DNA contacts, but discussion in this paper focuses on the homeodomain at the TAATCC site as this binding site has the same sequence as that used in the biochemical studies [6] and because this complex does not have a nick in the DNA. (As in the wild-type complex, the other homeodomain binds to a 'nicked' site formed by the juxtaposition of neighboring DNA duplexes.)

Accession numbers

Coordinates are being deposited with the Brookhaven Data Bank. While they are being processed, a set of coordinates may be obtained by sending an e-mail message to pabo@mit.edu.

Acknowledgements

This project was supported by an NIH grant (GM31471) to COP and used equipment purchased with support from the PEW Charitable Trusts. We thank Lena Nekludova for help with some of the modeling studies cited in this paper, and we thank Ernest Fraenkel and Joel Pomerantz for helpful comments on this manuscript. LT-K was supported by an NSF Predoctoral Fellowship.

References

1. Gehring, W.J., Affolter, M. & Bürglin, T. (1994). Homeodomain proteins. *Annu. Rev. Biochem.* **63**, 487–526.
2. Hanes, S.D. & Brent, R. (1989). DNA specificity of the bicoid activator protein is determined by homeodomain recognition helix residue 9. *Cell* **57**, 1275–1283.
3. Hanes, S.D. & Brent, R. (1991). A genetic model for interaction of the homeodomain recognition helix with DNA. *Science* **251**, 426–430.
4. Treisman, J., Gönczy, P., Vashishtha, M., Harris, E. & Desplan, C. (1989). A single amino acid can determine the DNA binding specificity of homeodomain proteins. *Cell* **59**, 553–562.
5. Percival-Smith, A., Müller, M., Affolter, M. & Gehring, W.J. (1990). The interaction with DNA of wild-type and mutant *fushi tarazu* homeodomains. *EMBO J.* **9**, 3967–3974.
6. Ades, S.E. & Sauer, R.T. (1994). Differential DNA-binding specificity of the engrailed homeodomain: the role of residue 50. *Biochemistry* **33**, 9187–9194.
7. Kissinger, C.R., Liu, B., Martin-Blanco, E., Kornberg, T.B. & Pabo, C.O. (1990). Crystal structure of an engrailed homeodomain–DNA complex at 2.8 Å resolution: a framework for understanding homeodomain-DNA–interactions. *Cell* **63**, 579–590.
8. Wolberger, C., Vershon, A.K., Liu, B., Johnson, A.D. & Pabo, C.O. (1991). Crystal structure of a MATα2 homeodomain–operator complex suggests a general model for homeodomain–DNA interactions. *Cell* **67**, 517–528.
9. Billeter, M., Qian, Y.Q., Otting, G., Müller, M., Gehring, W. & Wüthrich, K. (1993). Determination of the nuclear magnetic resonance solution structure of an *Antennapedia* homeodomain–DNA complex. *J. Mol. Biol.* **234**, 1084–1097.
10. Billiter, M., Guntert, P., Luginbühl, P. & Wüthrich, K. (1996). Hydration and DNA recognition by homeodomains. *Cell* **85**, 1057–1065.
11. Klemm, J.D., Rould, M.A., Aurora, R., Herr, W. & Pabo, C.O. (1994). Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA-binding modules. *Cell* **77**, 21–32.
12. Wilson, D.S., Guenther, B., Desplan, C. & Kuriyan, J. (1995). High resolution crystal structure of a paired (Pax) class cooperative homeodomain dimer on DNA. *Cell* **82**, 709–719.
13. Li, T., Stark, M.R., Johnson, A.D. & Wolberger, C. (1995). Crystal structure of the MATa1/MATα2 homeodomain heterodimer bound to DNA. *Science* **270**, 262–269.
14. Hirsch, J.A. & Aggarwal, A.K. (1995). Structure of the even-skipped homeodomain complexed to AT-rich DNA: new perspectives on homeodomain specificity. *EMBO J* **14**, 6280–6291.
15. Wilson, D.S., Sheng, G., Jun, S. & Desplan, C. (1996). Conservation and diversification in homeodomain–DNA interactions: a comparative genetic analysis. *Proc. Natl. Acad. Sci. USA* **93**, 6886–6891.
16. Verrijzer, C.P., Alkema, M.J., van Weperen, W.W., van Leeuwen, H.C., Strating, M.J.J. & van der Vliet, P.C. (1992). The DNA binding specificity of the bipartite POU domain and its subdomains. *EMBO J.* **11**, 4993–5003.
17. Ingraham, H.A., *et al*., & Rosenfeld, M.G. (1990). The POU-specific domain of Pit-1 is essential for sequence-specific, high-affinity DNA binding and DNA-dependent Pit-1–Pit-1 interactions. *Cell* **61**, 1021–1033.
18. Ades, S.E. (1995). *The Engrailed Homeodomain: Determinants of DNA-Binding Affinity and Specificity*. PhD Thesis, Massachusetts Institute of Technology, USA.
19. Mandel-Gutfreund, Y., Schueler, O. & Margalit, H. (1995). Comprehensive analysis of hydrogen bonds in regulatory protein–DNA complexes: in search of common principles. *J. Mol. Biol.* **253**, 370–382.
20. Fersht, A.R., *et al*., & Winter, G. (1985). Hydrogen bonding and biological specificity analyzed by protein engineering. *Nature* **314**, 235–238.
21. Saenger, W. (1984). *Principles of Nucleic Acid Structure*. Springer-Verlag, New York, NY, USA.
22. Clarke, N.D., Beamer, L.J., Goldberg, H.R., Berkower, C. & Pabo, C.O. (1991). The DNA binding arm of λ repressor: critical contacts from a flexible region. *Science* **254**, 267–270.
23. Beamer, L.J. & Pabo, C.O. (1992). Refined 1.8 Å crystal structure of the λ repressor–operator complex. *J. Mol. Biol.* **227**, 177–196.
24. Jordan, S.R. & Pabo, C.O. (1988). Structure of the lambda complex at 2.5 Å resolution: details of the repressor–operator interactions. *Science* **242**, 893–899.
25. Aggarwal, A., Rodgers, D., Drottar, M., Ptashne, M. & Harrison, S.C. (1988). Recognition of a DNA operator by the repressor of phage 434: a view at high resolution. *Science* **242**, 899–907.
26. Pomerantz, J.L. & Sharp, P.A. (1994). Homeodomain determinants of major groove recognition. *Biochemistry* **33**, 10851–10858.
27. Blundell, T.L. & Johnson, L.N. (1976). *Protein Crystallography*. Academic Press, San Diego, CA, USA.
28. Rould, M.A. (1997). Screening for heavy atom derivatives and obtaining accurate isomorphous differences. *Methods Enzymol*. **276**, 461–472.
29. Collaborative Computational Project Number 4. (1994). The CCP4 suite: programs for protein crystallography. *Acta Cryst. D* **50**, 760–763.
30. Wang, B.C. (1985). Resolution of phase ambiguity in macromolecular crystallography. *Methods Enzymol.* **115**, 90–112.
31. Rould, M.A., Perona, J.J. & Steitz, T.A. (1992). Improving MIR phasing by heavy atom refinement using solvent-flattened phases. *Acta Cryst. A* **48**, 751–756.
32. Jones, T.A. (1978). A graphics model building and refinement system for macromolecules. *J. Appl. Cryst.* **11**, 268–272.
33. Israel, M. & Chirino, A.J. (1991). *TOM/FRODO version 3.0.* University of Alberta, Alberta, Canada and California Institute of Technology, CA, USA.
34. Brünger, A.T. (1992). *X-PLOR Version 3.1: a System for X-ray Crystallography and NMR*. Yale University Press, New Haven, CT, USA.