Protein Science (1994), *3*:1779–1787. Cambridge University Press. Printed in the USA.

Structural studies of the engrailed homeodomain

NEIL D. CLARKE,[1] CHARLES R. KISSINGER,[2] JOHN DESJARLAIS,[3]
GARY L. GILLILAND,[4] AND CARL O. PABO[5]

[1] Department of Biophysics and Biophysical Chemistry, Johns Hopkins School of Medicine, Baltimore, Maryland 21205
[2] Agouron Pharmaceuticals, Inc., San Diego, California 92121
[3] DuPont Merck Pharmaceutical Co., Experimental Station, Wilmington, Delaware 19880
[4] Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute and National Institute of Standards and Technology, Rockville, Maryland 20950
[5] Howard Hughes Medical Institute, Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

(RECEIVED May 31, 1994; ACCEPTED July 1, 1994)

Abstract

The structure of the *Drosophila* engrailed homeodomain has been solved by molecular replacement and refined to an *R*-factor of 19.7% at a resolution of 2.1 Å. This structure offers a high-resolution view of an important family of DNA-binding proteins and allows comparison to the structure of the same protein bound to DNA. The most significant difference between the current structure and that of the 2.8-Å engrailed–DNA complex is the close packing of an extended strand against the rest of the protein in the unbound protein. Structural features of the protein not previously noted include a "herringbone" packing of 4 aromatic residues in the core of the protein and an extensive network of salt bridges that covers much of the helix 1–helix 2 surface. Other features that may play a role in stabilizing the native state include the interaction of buried carbonyl oxygen atoms with the edge of Phe 49 and a bias toward statistically preferred side-chain dihedral angles. There is substantial disorder at both ends of the 61 amino acid protein. A 51-amino acid variant of engrailed (residues 6–56) was synthesized and shown by CD and thermal denaturation studies to be structurally and thermodynamically similar to the full-length domain.

Keywords: crystallography; homeodomain; protein stability; protein structure

Homeodomains are common eukaryotic DNA-binding domains that consist of a short extended strand followed by 3 helices (Qian et al., 1989; Kissinger et al., 1990; Laughon 1991; Wolberger et al., 1991). The homeodomain itself (conventionally defined as being 60 amino acids in length on the basis of homology) contains much of the DNA-binding specificity and affinity of the much larger proteins in which it is found (Kornberg, 1993). Moderate-resolution crystal structures of 3 protein:DNA complexes that include homeodomains (engrailed, MATa2, and Oct-1) have been determined, as have the NMR structures of Antennapedia (Antp), an Antp:DNA complex, and fushi tarazu (ftz) (Qian et al., 1989, 1994; Kissinger et al., 1990; Wolberger et al., 1991; Billeter et al., 1993; Klemm et al., 1994). There is considerable interest in understanding how these domains recognize DNA.

Homeodomains are also of interest because they present an extremely simple model system for studies related to protein folding. They are small and monomeric and do not require disulfide bonds or ligands in order to fold stably. There is considerable sequence diversity among the several hundred homeodomain sequences known, but a conserved pattern of hydrophobic residues and the similarity of the known structures suggests that each of the hundreds of sequences adopts the same overall structure (Scott et al., 1989; Laughon, 1991).

We have obtained a high-resolution structure of the engrailed homeodomain in the absence of DNA (Kinemage 1). This structure is useful because (1) it allows direct comparison to the crystallographic structure of a homedomain bound to DNA and (2) it provides a more accurate coordinate set for studies of the sequence–structure relationship of this simple structural motif.

Results

Description of crystallographic structure

The engrailed homeodomain has been refined to an *R*-factor of 19.7 for data between 8.0 and 2.1 Å, with a 2σ cutoff. The *R*-factor is 18.9 for data greater than 5σ. Table 1 lists details of the crystallographic data and the geometry of the final model. Figure 1 shows the *R*-factor for various resolution shells and the cumulative *R*-factor as a function of resolution for data greater than 2σ.

Reprint requests to: Neil D. Clarke, 708 WBSB, Biophysics and Biophysical Chemistry, Johns Hopkins School of Medicine, 725 N. Wolfe Street, Baltimore, Maryland 21205; e-mail: neil.clarke@qmail.bs.jhu.edu.

Table 1. *Data collection and refinement statistics*

Space group	$P6_522$	
Unit cell dimensions (Å)		
$a = b$	44.67	
c	118.12	
Data collection statistics		
(Data set)[a]	1	2
Resolution limit	2.23	1.89
N_{unique}	3,167	6,104
N_{obs}	12,679	21,567
R_{merge}[b]	4.61	3.75
Completeness (%)	83.3	86.8
Refinement statistics		
Atoms (non-hydrogen)	466	
Missing atoms	43	
Water molecules	33	
Resolution range (Å)	8.0–2.1	
	2σ	(5σ)
R-factor[c]	19.7	(18.9)
No. reflections	4,060	(3,721)
Completeness (cumulative)[d]	92.1	(84.4)
Completeness (high res)[e]	77.7	(60.8)
RMS deviations from ideal geometry		
Bond lengths (Å)	0.013	
Bond angles (°)	1.59	
Improper torsion angles (°)	1.18	

[a] Data set 1 was used in the structure solution and in the early stages of refinement. Data set 2 was used subsequently.

[b] $R_{merge} = 100 * \sum_h \sum_i |I(h)_i - \langle I(h)\rangle| / \sum_h \sum_i I(h)_i$, where $I(h)_i$ is the intensity of the ith observation of reflection h, and $\langle I(h)\rangle$ is the mean intensity of reflection h.

[c] R factor $= 100 * \sum |F_{obs} - F_{calc}| / \sum F_{obs}$.

[d] Completeness = (observed reflections above the σ cutoff and in the resolution range)/(total possible reflections in the resolution range).

[e] High-resolution shell defined as data between 2.10 and 2.16 Å; completeness as defined in footnote d.

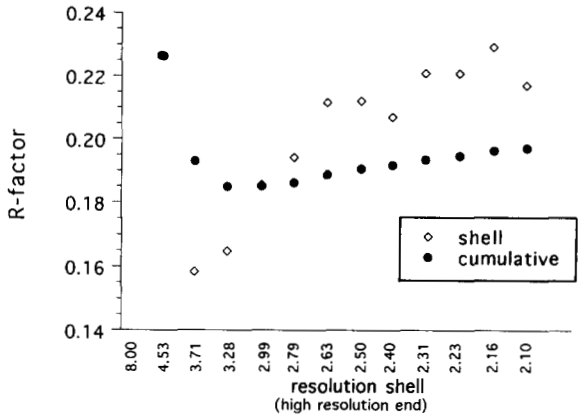

Fig. 1. R-factor as a function of resolution shell. Open diamonds show the R-factor for reflections within a particular resolution shell. Closed circles show the cumulative R-factor as data are included from low-resolution shells to high. The value on the x-axis indicates the high-resolution end of the shell; the shell consists of data between that resolution and the value of the point to the left. The resolution shells are of approximately equal volume.

A ribbon diagram of the structure is shown in Figure 2A. The sequence of engrailed with secondary-structure ranges indicated is shown in Figure 2B. The secondary-structure elements are defined here on the basis of contiguous residues having the appropriate ϕ–ψ values. All helical residues identified by this criterion also meet hydrogen bonding criteria for helices. Residues at the amino-terminus are drawn as a β-strand, although this is an isolated strand and not part of a β-sheet.

The structure shows excellent geometry (Table 1; Fig. 3). The ϕ–ψ values for helical residues are tightly clustered in the appropriate region of the Ramachandran diagram (Fig. 3). The only glycine in the protein (G39) lies near the α_L region. Most of the other loop residues and extended-strand residues also lie within a favorable region of ϕ–ψ space. Two residues (R24 and L38) have somewhat unfavorable ψ angles. For both residues, hydrogen bonds to their backbone atoms by other side chains may help stabilize these slightly unfavorable conformations. Both the carbonyl oxygen and the amide nitrogen of R24 participate in hydrogen bonds with the side chains of other residues (the oxygen with R53 and the nitrogen with N23). L38 accepts a hydrogen bond to its carbonyl oxygen from the side-chain amide nitrogen of Q12.

The backbone for residues 7–53 is well determined, as reflected in the relatively low temperature factors (Fig. 4). These residues correspond roughly to what could be considered the globular part of the homedomain. As the protein extends toward both termini (residues 3–6 and 54–56), the chain moves away from this well-packed core and the structure becomes less constrained. The current model ends at residue 56. Despite the poor density and higher temperature factors in this region, the continuity of the helix through residue 56 is clear. The structure is also poorly ordered in the first 3 residues of the model (residues 3–5), but persistent attempts to build alternative structures or to refine the structure in the absence of these residues failed to produce better agreement with the data. Side chains in the core are well determined; surface side chains differ considerably in the quality of the electron density. Side chains involved in hydrogen bonding and salt bridge networks (discussed below) are well ordered. Electron density for R18 is extremely poor but has been included in the model for the sake of including all side-chain atoms for which main-chain coordinates are reported. Q32 appears to have multiple conformations but has been built with the single best rotamer.

Comparison to an engrailed:DNA complex

The backbone of the 2.1-Å structure is very similar to the 2 homeodomains determined independently in the 2.8-Å engrailed:DNA complex: the RMS deviations (RMSDs) for Cα atoms of residues 6–56 are less than 0.5 Å in each case (Kissinger et al., 1990). The 10 most buried residues in the current structure have, as a group, RMSDs of 0.6 Å and 0.9 Å when compared to the 2 complex structures. Surface side chains are substantially different in many cases. The most significant difference between the bound and unbound homeodomains is the way the N-terminal extended segment packs against the 3-helix motif. In the current structure, the extended segment is more closely packed against the rest of the protein than is the case in the protein:DNA complex. Figure 5 and Kinemage 1 illustrate this difference by showing a superpositioning of Cα atoms for the

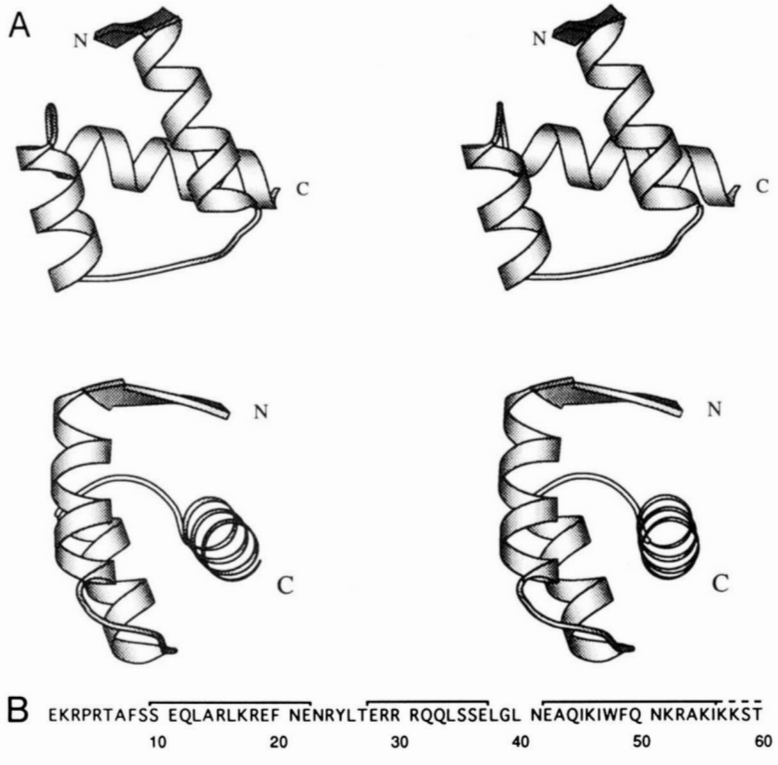

Fig. 2. A: Stereo ribbon drawing of engrailed homeodomain. **B:** Sequence of engrailed with helices indicated by bars above the sequence. The crystallographic model ends at residue 56. The dashed line after 156 indicates possible extension of a poorly ordered helix.

structure reported here and that of an engrailed:DNA complex. As previously noted, the amino-terminal arm in the complex wraps over the phosphate backbone to make contacts in the minor groove, precluding a tight packing between the arm and the rest of the structure (Kissinger et al., 1990). However, DNA binding results in a significantly more ordered structure for the arm than exists in the unbound structure reported here. This DNA-induced ordering has also been observed in NMR studies of the bound and unbound homeodomain from Antp (Qian et al., 1989; Billeter et al., 1993).

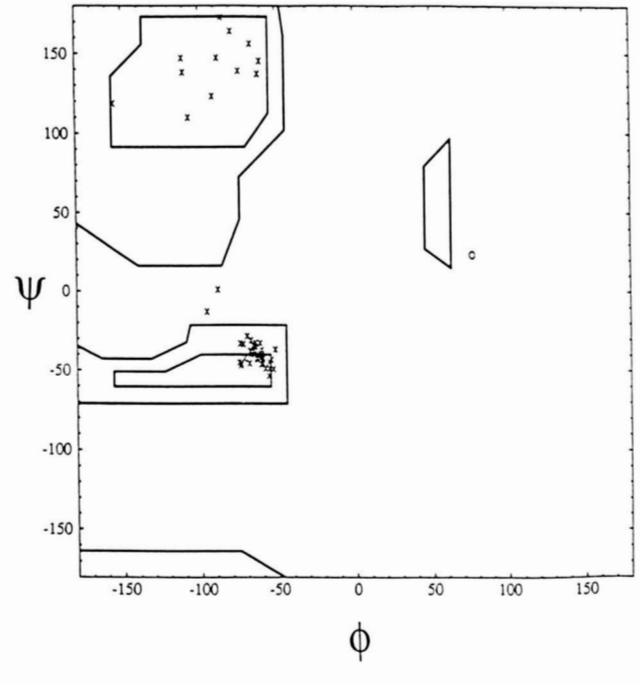

Fig. 3. Ramachandran diagram. G39 is shown as a circle; all other residues are indicated by an X.

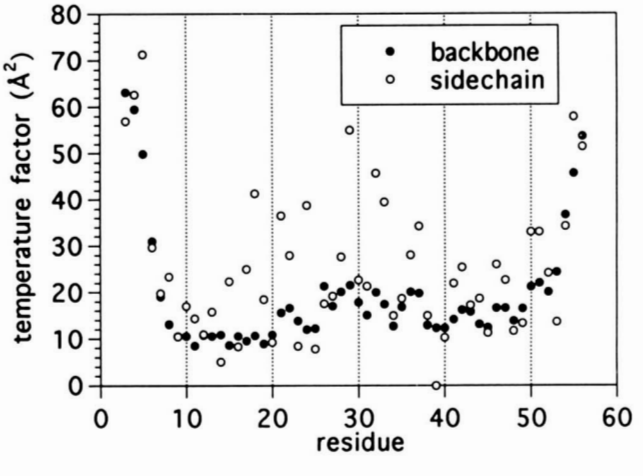

Fig. 4. Stereo drawing of the superposition of Cα atoms for the current structure on Cα atoms for one of the domains from an engrailed-DNA complex. The unbound structure is shown as a solid line. The domain as found in a DNA complex is shown as a dashed line. The extended strand at the top of the models is the N-terminal end of the domain. In the current structure, the strand is more tightly packed against the core of the protein than is the case in the protein-DNA complex. The Cα's of residues 8–54 were used to align the structures.

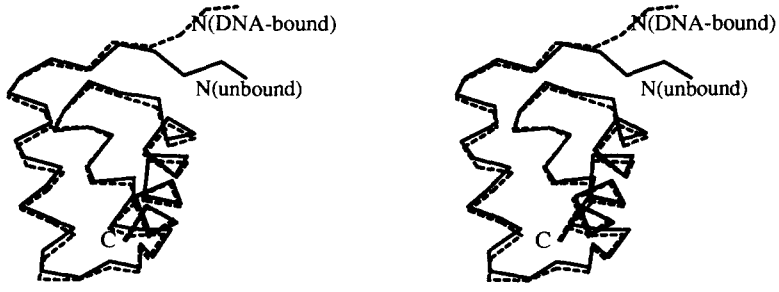

Fig. 5. Average *B*-factors for backbone atoms (solid) and side-chain atoms (open) for each residue in the structure. *B*-factors were restrained in the refinement as described in the X-PLOR manual.

A similar induction of structural order at the C-terminal end of the protein is also apparent from this analysis and from NMR studies of other homeodomains (Qian et al., 1989; Billeter et al., 1993). Disorder in residues near the C-terminal end of the domain that are known or presumed to be involved in DNA binding is especially dramatic in the case of the recently solved ftz domain (Qian et al., 1994). In addition to structural comparisons of the type presented here, analysis of thermodynamic data also suggests that DNA-binding proteins frequently recognize their cognate binding sites by an induced-fit mechanism (Spolar & Record, 1994).

Structural features of the engrailed homeodomain

In addition to their importance as DNA-binding motifs, the family of homeodomains has properties that make it an excellent system for studying sequence-structure relationships. We have identified several features of the engrailed homeodomain that may be important in stabilizing the native fold.

Cluster of aromatic side chains

Of the 10 side chains in the structure that are greater than 90% buried, 4 (F8, F20, W48, and F49) are aromatic. These 4 side chains form a cluster in which each is in contact with at least one of the others (Fig. 6; Kinemage 2). The aromatic nature of these residues is a highly conserved feature of homeodomain sequences. Among 119 unique homeodomain sequences (compiled in Laughon [1991]), residue 48 is always tryptophan, residues 20 and 49 are aromatic in all but 1 sequence each, and residue 8 is aromatic in about 80% of the sequences. A similar network of aromatic residues (also including a conserved tryptophan) has been observed in the structurally distinct 3-helix motif of the HMG box (Weir et al., 1993). The relative orientation of side chains in the engrailed cluster is similar to that observed in other proteins and is consistent with the idea that angles between the ring planes in the range of 60–90° are most favorable (Burley & Petsko, 1985).

Carbonyl oxygen–aromatic ring interactions

The edges of aromatic rings can also interact with carbonyl oxygens (Thomas et al., 1982). A particularly dramatic example of this kind of interaction can be seen in engrailed (Fig. 7; Kinemage 2). The carbonyl oxygens of 3 residues (23, 24, and 45) all lie within 3.5 Å of F49 and are very nearly in the plane of the ring. Each appears to be well oriented to form a kind of hydrogen bond with the hydrogens of the phenylalanine ring.

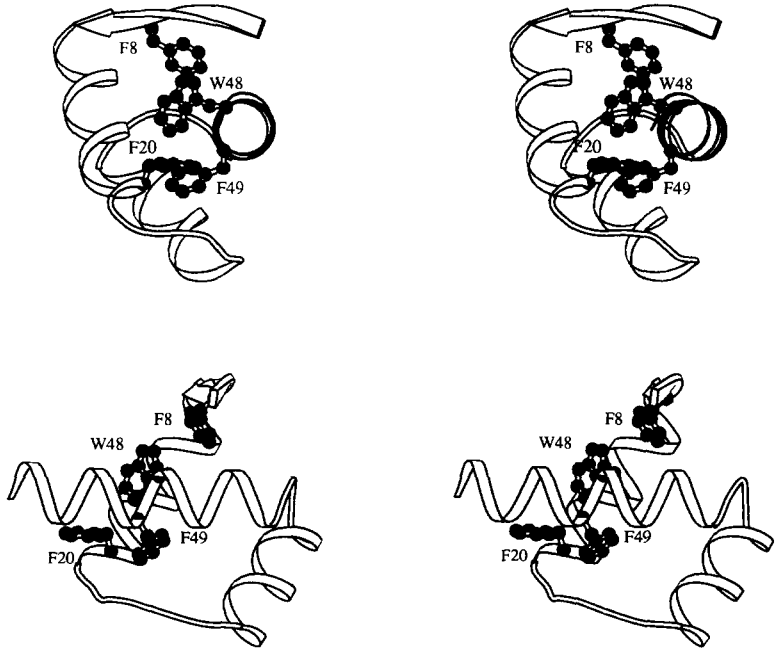

Fig. 6. Cluster of 4 aromatic residues (F8, F20, W48, W49). Two mutually orthogonal views are shown in stereo.

Fig. 7. Interactions of backbone carbonyl oxygens (residues 23, 24, and 45) with the ring of F49. Two views are shown in stereo.

Interestingly, inspection of phenylalanine residues acting as σ acceptors to other aromatic residues revealed a preference for acceptance near the ortho hydrogens, followed by the meta hydrogens, and then the para (Burley & Petsko, 1985). This order of preference is observed at F49 as well because both ortho positions and 1 meta site are involved in the interactions with the carbonyl oxygens.

Salt bridge network

An extensive network of salt bridges and hydrogen bonds covers much of the helix 1–helix 2 surface of the protein, which is the area furthest from the DNA-binding surfaces of helix 3 and the arm. Figure 8A and Kinemage 2 show the contacts made among residues R15, E19, R30, and E37. R15 and E19 on helix 1 form a salt bridge with each other and each participates in a salt bridge with a residue in helix 2. R30 makes an especially interesting set of contacts (Fig. 8B). In addition to the salt bridge to E19, R30 makes hydrogen bonds to the carbonyl oxygens of residues 23 and 25 at the beginning of the loop between helices 1 and 2 and, somewhat less ideally, to the carbonyl of residue 19 near the end of helix 1. Thus, each of the 4 possible hydrogen bonds to the terminal amines of arginine appears to be satisfied. Reexamination of omit map electron density for the engrailed–DNA complex suggests that the R30 side chain has 2 discrete conformations in that crystal structure with approximately equal occupancies. One conformation corresponds to that found in this study; the other is as originally modeled (Kissinger et al., 1990).

Side-chain–backbone hydrogen bonds

There are several other examples of side-chain hydrogen bonds to backbone atoms at or near the ends of helices. Most prominent among these is a reciprocal hydrogen bonding pattern between S9 and Q12 at the beginning of helix 1 (Fig. 9; Kinemage 2). S9, which makes an $i, i+4$ backbone hydrogen bond to start helix 1, but does not itself adopt a helical conformation, makes a side-chain hydrogen bond to the backbone amide of Q12. The side-chain carbonyl of Q12, in turn, makes a hydrogen bond to the backbone amide of S9. Reciprocal hydrogen bond partners of this type have been dubbed "capping boxes" (Harper & Rose, 1993). Q12 also hydrogen bonds to the carbonyl oxygen of L38, near the end of helix 2. Thus, the side chain of Q12 acts as both a hydrogen bond donor and acceptor to backbone atoms near the ends of helices. Other tertiary interactions of note include Q44 in a hydrogen bond with the backbone amide of T6 and R53 in a hydrogen bond with the backbone carbonyl of R24.

Use of preferred side-chain dihedral angles

The distribution of side-chain dihedral angles in proteins has been surveyed periodically. As the database of structures has grown, it has become possible to obtain statistically significant distributions for individual amino acids in different secondary structures (McGregor et al., 1987). These distributions likely reflect intrinsic differences in the energetics of particular side-chain conformations. Extensive use of preferred dihedral angles may be 1 way in which proteins achieve added stability in the native

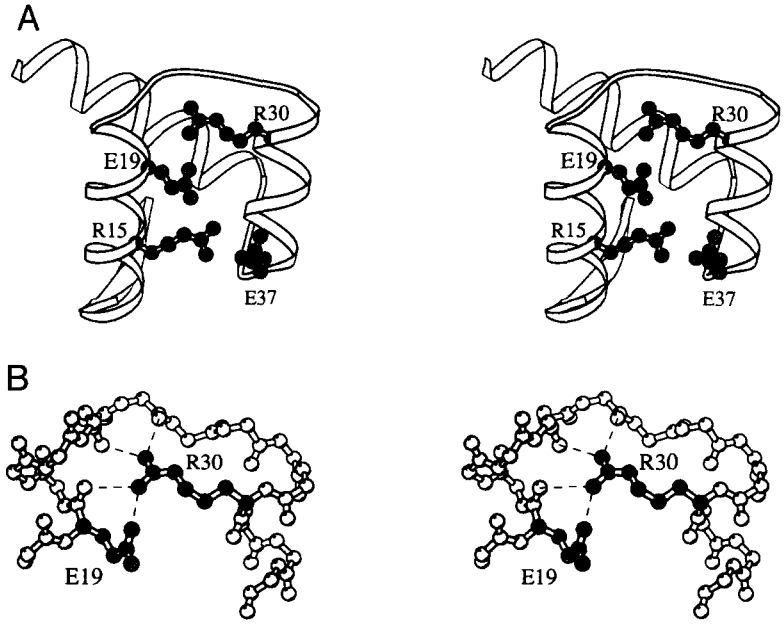

Fig. 8. A: Stereo drawing of the engrailed backbone with side chains of residues R15, E19, R30, and E37. This extensive network of salt bridges covers much of the surface of helices 1 and 2. **B:** Closeup view of R30 (dark balls), E19 (light balls), and the backbone of residues 18–33.

state. We examined the side-chain conformations of all residues in engrailed that are greater than 90% buried and compared their conformations to the distributions tabulated by McGregor et al. (1987). We chose to look at buried residues because their conformations are better determined than those of surface residues. The data in Table 2 indicate that these side chains show a greater-than-expected preference for the most favored conformations. Eight of the 10 residues adopt the preferred conformation for that residue type and secondary structure, despite the fact that the average frequency in the database for the most favored conformation is only 0.54. The dihedral angle values are also, in general, close to the relevant mean value, with 18 of 20 angles within 1 standard deviation of the mean for that residue in that secondary structure. The 1 notable exception to the bias toward statistically preferred conformations is L26 $\chi 1$. We examined the reason for this by performing empirical energy calculations at 5° increments of both $\chi 1$ and $\chi 2$ on a leucine with the backbone dihedrals of L26 ($\phi = -107°$, $\psi = 111°$) but in the absence of any other atoms. The conformational search showed a strong preference for a $\chi 1$ angle of 185–190°, similar to what we observed in engrailed. The preference for a *trans* dihedral angle was attributable to steric clashes with the backbone atoms of residue 26 and suggests that the observed side-chain conformation is strongly preferred by the local backbone conformation. As

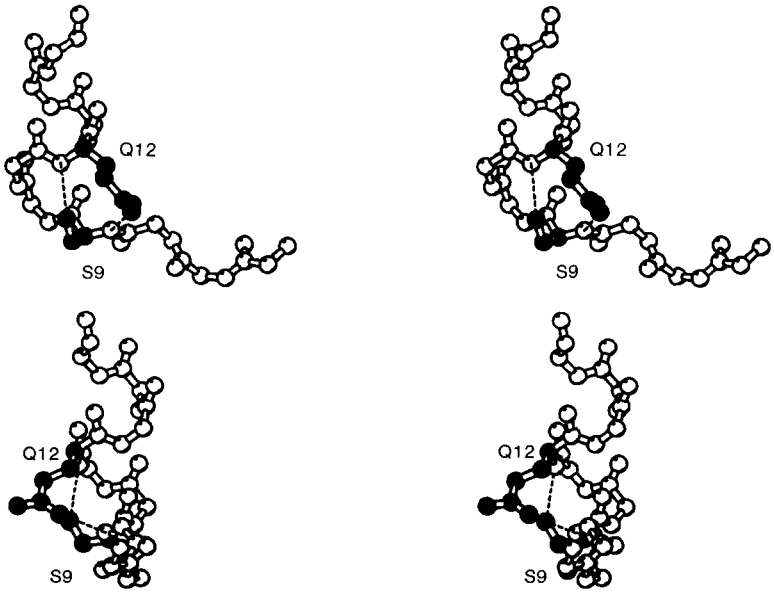

Fig. 9. N-terminal capping box of helix 1. Atoms of the S9 and Q12 side chains are shown as dark balls.

Table 2. *Side-chain dihedral angles for the 10 buried side chains in engrailed*[a]

Residue	Secondary structure	χ_1 (σ from mean of class)	χ_2 (σ from mean of class)	Rank order of conform. class	Frequency of conform. class
F8	–	304 (0.86)	90 (0.20)	1	0.58
L16	α-Helix	281 (0.46)	177 (0.11)	1	0.49
F20	α-Helix	184 (0.36)	94 (0.94)	1	0.59
L26	–	206 (0.83)	61 (0.05)	2	0.12 (0.60)
L34	α-Helix	279 (0.62)	180 (0.06)	1	0.49
L38	–	297 (0.11)	176 (0.06)	1	0.51
L40	–	280 (0.83)	162 (0.81)	1	0.51
I45	α-Helix	300 (0.78)	178 (0.40)	1	0.66
W48	α-Helix	187 (1.11)	85 (0.33)	1	0.42
F49	α-Helix	264(1.64)	79 (0.50)	2	0.40 (0.59)

[a] – in the "secondary structure" column indicates that the residue is not in a regular secondary structure. The values for the 2 side-chain dihedral angles ($\chi 1$ and $\chi 2$) for each residue are shown in columns 3 and 4. In parentheses in those columns is the degree to which the observed dihedral angle differs from the mean value for that amino acid in that secondary structure class; the difference is expressed as the number of standard deviations away from the mean. The values for these calculations are taken from Table 4 of McGregor et al. (1987). The fifth column shows whether the observed conformational class was the most frequently found (as it is for 8 of the residues) or the second most frequently found conformation in the McGregor et al. (1987) survey of high-resolution structures. The last column gives the frequency with which that conformational class was found in the survey. The number in parentheses that appears for the 2 residues that do not adopt the most common conformation is the frequency of that conformation. These values are also extracted from the data of McGregor et al. (1987).

a control, similar calculations for a leucine in a helix reproduced the preference for *gauche+* dihedral angles observed in the structure survey of McGregor et al. (1987).

Structural and thermodynamic properties of a 6–56 fragment of engrailed homeodomain

Homeodomains are short fragments of much larger proteins and are conventionally defined on the basis of homology limits as being 60 amino acids in length. This structure determination and others show that a significant number of these residues are disordered in the isolated domains in the absence of DNA (Qian et al., 1989, 1994; Kissinger et al., 1990; Wolberger et al., 1991). We synthesized a peptide corresponding to residues 6–56 of engrailed to test whether the crystallographically well-determined part of the structure could itself adopt a stable conformation. The 51-amino acid peptide was compared to the full-length homeodomain by CD and by thermal denaturation. Figure 10A shows that the CD spectrum of the shorter peptide is very similar to that of the full-length homeodomain. Figure 10B shows that the T_m and the van't Hoff ΔH for unfolding for the full-length and 6–56 peptides are likewise similar, suggesting that 51 amino acids suffice for essentially wild-type folding of the engrailed homedomain.

Discussion

The 2.1-Å crystal structure of the engrailed homedomain provides a high-resolution view of a simple structural motif that offers a number of valuable features for the study of DNA binding and protein folding. We also find that a 51-amino acid fragment of the homeodomain folds into a structure indistinguishable by CD from that of the full-length domain and has similar thermodynamic properties. The engrailed homeodomain is essentially fully folded at room temperature and has a T_m of about 45°. There are a surprisingly large number of salt bridges and hydrogen bonds among side chains and between side chains and backbone atoms on the surface of the protein. Although it is not proven that these provide net stabilization to the native state, it may be that the relatively small core/surface ratio of this small domain requires such interactions to a greater extent than larger proteins. The clustering of aromatic residues, the interactions of carbonyl oxygens with an aromatic ring, and the preponderance of preferred side-chain dihedrals may contribute to stability.

Methods

Structure solution

The crystallization of the engrailed homeodomain has been described previously (Liu et al., 1990). The crystals are hexagonal ($P6_522$; $a = b = 44.67$ Å, $c = 118.12$ Å), with 1 molecule in the asymmetric unit, and diffract to a limit of about 1.9 Å. Data were collected on a Siemens multiwire proportional area detector and processed with XENGEN (Howard et al., 1987). The structure was solved by molecular replacement using X-PLOR with PC refinement of rotation peaks (Brünger et al., 1987; Brünger, 1990). Various search models derived from the engrailed–DNA complex were attempted. The successful rotation-function model consisted of residues 7–54, with 3 residues (18, 29, and 30) replaced by alanine. The rotation-function model was used in an X-PLOR translation search. Because the handedness of the screw axis was not known, translation searches were conducted separately in both $P6_522$ and $P6_122$. A single translation-function solution in $P6_522$ was found that scored approximately 8σ higher than others by the crystallographic function and also scored well in a packing function. Rigid-body refinement of this solution lowered the R-factor to 46.5% for data to 3.2 Å. The solution was confirmed using a single isomorphous derivative (K_2PtCl_4). Difference maps phased with the molecular replacement model gave an approximately 9σ peak for a site previously identified in Patterson maps.

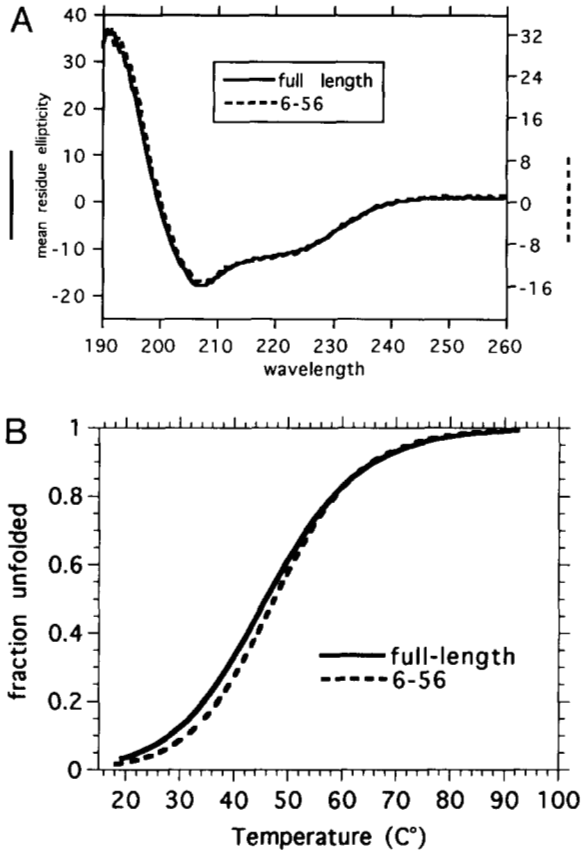

Fig. 10. A: CD spectrum of the full-length engrailed homeodomain (solid line) and the 6–56 variant (dashed line). The mean residue ellipticity is in units of degree cm^2 $dmol^{-1}$. Note that 6–56 is plotted on a slightly different *y*-axis (right). This may reflect errors in protein concentration determination. Mean residue ellipticity for 6–56 has been length normalized to that of the full-length protein. **B:** Thermal denaturation of full-length homeodomain and the 6–56 variant, monitored by CD at 222 nm. The data were fit by nonlinear regression to a 2-state denaturation model. A complete folding/unfolding transition was assumed based on more detailed studies of 6–56 (see Methods). The midpoint of the transition (T_m) is 46.0° for the full-length domain and 47.5° for 6–56. The van't Hoff enthalpy values are 23.7 kcal mol^{-1} and 26.4 kcal mol^{-1}, respectively.

Refinement

Model bias was reduced somewhat by combining molecular-replacement model phases with experimental phases from a single isomorphous derivative. Phase combination was performed using the PHASES program package written by W. Furey (W. Furey & S. Swaminathan, 1990, paper presented at the American Crystallographic Association Fortieth Anniversary Meeting, New Orleans, Louisiana). Refinement was primarily done with X-PLOR, although TNT was used for several cycles when the model was moderately well refined (Tronrud et al., 1987). Refinement was stalled for some time at an *R*-factor of about 23% and, during this period, we generally did not find normal molecular dynamics/simulated annealing or the use of simulated annealing omit maps to be helpful. However, by imposing harmonic restraints to the starting coordinates (e.g., 50 kcal mol^{-1} $Å^2$ for waters, 5 kcal mol^{-1} $Å^2$ for nonsolvent atoms), simulated annealing and manual rebuilding allowed us to break through this impasse. The requirement for harmonic restraints to reduce the structural change in any 1 annealing cycle was perhaps due to the relatively large number of long surface side chains and the large fraction of the structure that is poorly constrained at the ends. Individual temperature factors were refined with restraints on neighboring atoms; the values for these restraints were those suggested in the X-PLOR manual (Brünger, 1992). Thirty-three water molecules are present in the current model. The geometric parameters of Engh and Huber (1991) were used in the final refinement. Coordinates have been deposited in the Protein Data Bank.

Structure analysis and depiction

Solvent-accessibility calculations were done using the Lee and Richards algorithm as implemented in X-PLOR (Lee & Richards 1971; Brünger, 1992). The maximum possible buried area (to which the area buried in the structure is normalized) is here defined for each side chain as the accessible area of the side chain in the presence of the backbone atoms of the residue but in the absence of the rest of the structure. Analysis of the L26 conformational preferences employed a simple van der Waals-type potential currently being used to analyze alternative core packing arrangements (J. Desjarlais, unpubl. results). All structural figures were drawn with MOLSCRIPT (Kraulis, 1991).

CD

CD spectra were taken in 10 mM Tris, pH 7.4, with an Aviv CD60 spectropolarimeter. Several protein concentrations in the range of 40–150 μg/mL were used. Thermal denaturation experiments monitored by CD at 222 nm were conducted on a Jasco-710 spectropolarimeter with a temperature probe in contact with the protein solution. The temperature was increased from ambient temperature to about 90° over a 90-min period, with ellipticity recorded every 20 s. Peptide 6–56 was also examined from 2° to 90°, equilibrating for 2 min at each temperature using an Aviv model 62DS with a Hewlett-Packard temperature controller. Ellipticity below 20° and above 70° was found to vary linearly with temperature, being greater at low temperatures. The temperature dependence was similar at both temperature extremes. Although there is no obvious physical justification for a linear temperature dependence, baseline correction of all data between 2° and 90° was performed by subtraction of the apparent temperature-dependent ellipticity calculated by extrapolation of the data from 2 to 20°. Nonlinear regression of the corrected data produced an excellent fit to a 2-state model ($R > 0.999$), and suggested that the protein is approximately 99% folded at 20 °C and >99% unfolded at temperatures above 70 °C. Interpretation of the temperature scans between 20° and 90° therefore assumes a complete unfolding transition.

Protein and peptide purification

The purification of the protein used in crystallization has been described (Liu et al., 1990). The gene fragment used in expression of this protein encodes a 61-amino acid polypeptide corresponding to the initiator methionine (which is not encoded by engrailed), an Asp that immediately precedes the conventionally defined boundary for homedomains, and residues 1–59 of

the homeodomain. Preparation of protein for CD experiments was performed in a similar manner except that expression was from a synthetic gene under the control of a T7 promoter (gift of S. Ades and R.T. Sauer). This gene encodes the initiator methionine and residues 1-60. The molecular weight of this protein was confirmed by mass spectrometry, and it includes the N-terminal methionine. A peptide corresponding to residues 6-56 of engrailed was synthesized by standard solid-state peptide synthesis methods using a MilliGen 9050 PepSynthesizer with FMOC chemistry. The peptide was purified by reverse-phase (C4) chromatography, and the purity and molecular weight were confirmed by mass spectrometry.

Acknowledgments

We thank S. Ades and R. Sauer for the synthetic engrailed gene. Mass spectral analyses were carried out at the Middle Atlantic Mass Spectrometry Laboratory, an NSF Regional Instrumentation Facility. This work was supported by the Howard Hughes Medical Institute at Johns Hopkins (C.O.P., N.D.C., C.R.K.) and later at M.I.T. (C.O.P.), the National Institute of Standards and Technology, the Markey Center for Macromolecular Structure and Function at Johns Hopkins, and a National Research Council Research Associateship (N.D.C.). Certain commercial equipment, instruments, and materials are identified in this paper in order to specify the experimental procedure. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology nor does it imply that the material or equipment identified are necessarily the best available for the purpose.

References

Billeter M, Qian Y, Otting G, Muller 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.

Brünger A. 1990. Extension of molecular replacement: A new search strategy based on Patterson correlation refinement. *Acta Crystallogr A 46*: 46-57.

Brünger A. 1992. *X-PLOR version 3.1: A system for X-ray crystallography and NMR*. New Haven, Connecticut: Yale University Press.

Brünger A, Kuriyan J, Karplus M. 1987. Crystallographic *R* factor refinement by molecular dynamics. *Science 235*:458-460.

Burley S, Petsko G. 1985. Aromatic-aromatic interaction: A mechanism of protein structure stabilization. *Science 229*:23-28.

Engh R, Huber R. 1991. Accurate bond and angle parameters. *Acta Crystallogr A 47*:392-400.

Harper E, Rose G. 1993. Helix stop signals in proteins and peptides: The capping box. *Biochemistry 32*:7605-7609.

Howard A, Gilliland G, Finzel B, Poulos T, Ohlendorf D, Salemmed F. 1987. Use of an imaging proportional counter in macromolecular crystallography. *J Appl Crystallogr 20*:383-387.

Kissinger C, Liu B, Martin-Blanco E, Kornberg T, Pabo C. 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.

Klemm J, Rould M, Aurora R, Herr W, Pabo C. 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.

Kornberg T. 1993. Understanding the homeodomain. *J Biol Chem 268*: 26813-26816.

Kraulis P. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J Appl Crystallogr 24*:946-950.

Laughon A. 1991. DNA binding specificity of homeodomains. *Biochemistry 30*:11358-11367.

Lee B, Richards FM. 1971. The interpretation of protein structures: Estimation of static accessibility. *J Mol Biol 55*:379-400.

Liu B, Kissinger C, Pabo C. 1990. Crystallization and preliminary X-ray diffraction studies of the engrailed homeodomain and of an engrailed homeodomain/DNA complex. *Biochem Biophys Res Commun 171*: 257-259.

McGregor M, Islam S, Sternberg M. 1987. Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. *J Mol Biol 198*:295-310.

Qian Y, Billeter M, Otting G, Muller M, Gehring W, Wüthrich K. 1989. The structure of the *Antennapedia* homeodomain determined by NMR spectroscopy in solution: Comparison with prokaryotic repressors. *Cell 59*: 573-580.

Qian Y, Furukobo-Tokunaga K, Resendez-Perez D, Müller M, Gehring W, Wüthrich K. 1994. Nuclear magnetic resonance solution structure of the *fushi tarazu* homeodomain from *Drosophila* and comparison with the *Antennapedia* homeodomain. *J Mol Biol 238*:333-345.

Scott M, Tamkun J, Hartzell G. 1989. The structure and function of the homeodomain. *Biochim Biophys Acta 989*:25-48.

Spolar R, Record M. 1994. Coupling of local folding to site-specific binding of proteins to DNA. *Science 263*:777-784.

Thomas K, Smith G, Thomas T, Feldmann R. 1982. Electronic distributions within protein phenylalanine aromatic rings are reflected by the three-dimensional oxygen atom environments. *Proc Natl Acad Sci USA 79*: 4843-4847.

Tronrud D, Ten Eyck L, Matthews B. 1987. An efficient general-purpose least-squares refinement program for macromolecular structures. *Acta Crystallogr A 43*:489-501.

Weir H, Kraulis P, Hill C, Raine A, Laue E, Thomas J. 1993. Structure of the HMG box motif in the B-domain of HMG1. *EMBO J 12*:1311-1319.

Wolberger C, Vershon A, Liu B, Johnson A, Pabo C. 1991. Crystal structure of a MATα2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. *Cell 67*:517-528.