Published in final edited form as:

Mech Dev. 1993 September ; 43(1): 57–70.

A universal target sequence is bound in vitro by diverse homeodomains

Bill Kalionis^{*} and Patrick H. O'Farrell

Department of Biochemistry and Biophysics, University of California at San Francisco, San Francisco, CA 94143-0448, USA

Abstract

To determine the number of DNA binding proteins capable of binding a consensus Engrailed binding site, this consensus sequence was used to screen a library of *Drosophila* cDNA clones in a bacteriophage expression vector. We retrieved clones encoding 20 distinct DNA binding domains, 17 of which are homeodomains. Binding to a variety of oligonucleotides confirms the related sequence specificity of the retrieved binding domains. Nonetheless, the homeodomains have remarkably diverse amino acid sequences. We conclude that during the evolutionary divergence of homeodomains, the specificity of DNA binding has been much more highly conserved than the amino acid sequence.

Keywords

DNA recognition; *Drosophila*; Homeodomain

Introduction

The prevailing notion that evolution of new genes proceeds largely by gene duplication and divergence has been supported by sequence relationships among cloned genes. Sequences of DNA binding proteins provide a dramatic example. As has been emphasized in the literature, sequence similarities group these proteins into several gene families (Johnson and McKnight, 1989; Struhl, 1989). One such family is distinguished by a highly conserved DNA binding motif known as the homeodomain (McGinnis et al., 1984; Scott and Weiner, 1984; Desplan et al., 1985). This large group of proteins can be organized into a 'family tree' on the basis of the extent of amino acid sequence similarities (Scott et al., 1989).

If evolution of new homeodomain regulators has been associated with a selection for new DNA binding specificities, then different family members ought to have distinct specificities. It was suggested that specificity of DNA binding ought to diverge in parallel with divergence of amino acid sequences, particularly with divergence at positions that interact with DNA (Desplan et al., 1988). The history of experiments that might test this idea is rather confused. Early in vitro binding studies appeared to contradict expectations, in that the rather diverged homeodomains of the *engrailed*, *fushi tarazu*, and *even-skipped* genes all bound the same sequence even though they differed in residues expected to contact DNA (Desplan et al., 1988; Hoey et al., 1988). However, the initial interpretation of these experiments was flawed because it was assumed that the residues of the homeodomain contacting DNA would be the same as those of the

Correspondence to: P.H. O'Farrell, Department of Biochemistry and Biophysics, University of California at San Francisco, San Francisco, CA 94143-0448, USA.

^{*}Present address: Department of Obstetrics and Gynaecology, Flinders University, Adelaide, SA 5000, Australia.

structurally related helix-turn-helix proteins of prokaryotes (Desplan et al., 1988). The determination of the crystal structures for *engrailed* and Mat $\alpha 2$ complexes demonstrated that the major DNA contacts in the homeodomain: DNA complex involved amino acid residues more C-terminal (by one helical turn) than in the prokaryotic helix-turn-helix proteins (Kissinger et al., 1990; Wolberger et al., 1991). Substitutions at one of these positions (50) altered specificity, apparently changing recognition of a few base-pairs within the site (Hanes and Brent, 1989; Treisman et al., 1989; Hanes and Brent, 1991). This effect of a single substitution fit the initial suggestion that there might be incremental divergence of DNA binding specificity as evolutionary divergence altered individual contact residues, and hence, recognition of part of the site (Desplan et al., 1988). We sought to test whether there is an overall parallel between the divergence of DNA binding specificity and amino acid sequence. A direct parallel in divergence of homeodomain sequences and binding specificity would have the important implication that evolution of distinct regulatory functions is tightly coupled to divergence of binding specificity.

It is obvious that very related proteins that make similar contacts to DNA will have similar DNA binding specificity. However, the reverse is not true. Proteins that bind the same sequence are not necessarily highly homologous. For example, while the bacteriophage λ repressor and cro proteins recognize the same DNA sequence, they have entirely different sequences (for a review see Pabo and Sauer, 1984). Perhaps, like these prokaryotic proteins, some very diverged homeodomains also recognize the same DNA sequence.

Genes encoding sequence specific DNA binding proteins have been cloned by a direct screening method in which the protein products made by clones are immobilized on filters and probed with labeled oligonucleotides representing a characterized DNA recognition sequence (Singh et al., 1988; Vinson et al., 1988). This procedure has been used with the intent of cloning specific DNA binding proteins whose existence was first detected by binding to an upstream regulatory element. However, the method is capable of detecting DNA binding proteins having related specificity, and we have used it to explore the relationship between DNA binding specificity of homeodomain proteins and amino acid sequence similarities. Those amino acid residues involved in the discrimination of different DNA sequences ought to be conserved among the homeoproteins detected in this screen.

In a screen with an oligonucleotide that includes the consensus binding sequence for the Engrailed homeodomain, we have obtained twenty distinct *Drosophila* clones: 17 of these encode homeoproteins. Despite the similarity in the DNA binding specificity of these homeoproteins, there are few similarities in their amino acid sequences other than residues common to all homeodomains. Apparently, sequence divergence of homeodomain proteins has been associated with only marginal divergence of their DNA binding specificity.

Results

Isolation of cDNA clones

In vitro DNA binding using bacterially produced *engrailed* protein (Engrailed) identified a consensus sequence for binding (tCAATTAAat; upper case and lower case indicate highly and moderately conserved positions, respectively) (Desplan et al., 1988). These studies examined binding of three synthetic oligonucleotide sequences containing the consensus sequence: the non-palindromic sequence NP, tCAATTAAatga, which contains the full consensus (underlined), the symmetric LP sequence, tCAATT · AAatga, which contains two overlapping palindromically arranged copies of the core consensus sequence ATTA, a second symmetric sequence RP, tCATTT · AAatga, which deviates from the consensus at one of the core positions. Both NP and LP sites were bound by Engrailed, and binding was more avid when

the sites were arranged in three or more tandem copies. The sequence change from NP to RP resulted in a 25 fold reduction in binding (Desplan et al., 1988).

In addition to the Engrailed homeodomain, Ftz and Eve homeodomains bound these sequences (Desplan et al., 1988; Hoey and Levine, 1988). Furthermore, the LP and NP sequences function as transcriptional enhancers that respond to Engrailed, Ftz, Eve and Ubx in tissue culture transfection assays (Jaynes and O'Farrell, 1988), and also act as cell type specific enhancers in P-element transformed embryos (Vincent et al., 1990). The single base pair change in the RP site depresses in vitro binding, and greatly reduces function in transfection assays. We have used the LP sequence to screen expression libraries, and have used NP, RP and other oligonucleotides to examine binding specificity of the newly isolated clones.

We screened cDNAs that were produced from RNA isolated from 9 to 11 hour embryos and cloned in the λ phage expression vector λ gt11 (Zinn et al., 1988). The technical aspects of our procedures for ligand screening with oligonucleotide probes are detailed in Materials and Methods. In the initial screen, concatenated LP sequences detected 112 positive clones from 500000 plaques. These were purified through two additional rounds of screening. This resulted in the isolation of 83 purified plaques, and 72 of these were chosen for further study. Cleavage of each insert with frequently cutting restriction enzymes produced diagnostic patterns of fragments, and revealed that 20 different sequences were cloned, eight of which were represented by only one clone (data not shown). The failure to isolate more than one representative of eight of the clones, and the failure to isolate genes encoding proteins known to bind the LP sequence (e.g. *engrailed*, *ftz*, and *Ubx*) (Desplan et al., 1988) suggest that this screen is not saturating. We expect that further screening, especially if extended to additional cDNA libraries, will uncover many more clones that encode LP binding proteins.

Sequence analysis

Inserts from each of the cDNA clones were sub-cloned into a Bluescript vector. Degenerate oligonucleotides representing highly conserved parts of the homeodomain were used as sequencing primers (see Materials and Methods). These primers gave sequence data indicating that 17 out of the 20 clones contain a homeodomain. The entire insert of each of the three remaining clones (bk28, bk35, and bk60) was sequenced. While none of these included recognizable homology to a homeodomain, bk28 and bk35 encoded Zn finger domains that might be responsible for the DNA binding (data not shown). No previously characterized DNA binding motif was identified in bk60 (data not shown).

For those clones encoding a homeodomain, the initial sequence data allowed us to design complementary oligonucleotide primers that were used to complete the sequences of these homeodomains (Fig. 1a). Ten of the clones encode known homeodomain sequences; *lab*, *abd-A*, *Abd-B*, *cut*, *ems*, *pdm-1*, *Cfla*, *BarH1*, *zfh-1* and *zfh-2* (Scott et al., 1989; Dalton et al., 1989; Billin et al., 1991; Lloyd and Sakonju, 1991; Johnson and Hirsh, 1990; Kojima et al., 1991; Fortini et al., 1991). *zfh-2* had been shown to contain multiple Zn fingers and three homeodomains (Fortini et al., 1991). DNA sequence and restriction analysis revealed that our screen included two partially overlapping cDNA clones of *zfh-2*, *zfh-2 III_{bk69}*, which encoded only homeodomain III, and *zfh-2 I & II_{bk16}*, which encoded both homeodomains I and II (Fortini et al., 1991). Both clones were analyzed. However, because we have not established which homeodomain (either or both) encoded by the *zfh-2 I & II_{bk16}* clone is responsible for DNA binding, we omitted homeodomain I or II in analyses of homeodomain specificity (Fig. 1a). Their inclusion would have no marked effect on the results.

Sequence similarities have been used to relate homeodomains in a family tree. Close kinship of homeoproteins is revealed by a high degree of sequence similarity within the homeodomain, and occasionally by the presence of other regions of homology outside the homeodomain. For

example, a number of homeoproteins have homeodomains that are especially highly related to the homeodomain of the *paired* gene of *Drosophila* and many of these 'paired class' homeoproteins have been shown to have sequence similarity in a stretch of 18 residues N-terminal to the homeodomain (Bopp et al., 1986). Four of the newly identified homeodomains BK24, BK27, BK50, and BK36 are very similar to the homeodomain of *paired* (58%, 58%, 62% and 62% identity respectively: Fig. 1b), but lack the 18 residues N-terminal that are conserved among other paired class homeodomains. The clone encoding BK27 contains a second region of homology to *paired* in the paired box region (Bopp et al., 1986; Jun, Kalionis and Desplan, unpublished results). However, we did not sequence the entire insert of all of our cDNA clones, and have no data regarding the presence or absence of additional homology regions in our other clones.

The LIM class of homeodomains was defined by three genes, *lin-11* and *mec-3* (*C. elegans*), and *Isl-1* (rat), that share high similarity in their homeodomain sequence, as well as a second domain of similarity (Freyd et al., 1990; Karlsson et al., 1990). The BK64 and BK87 homeodomains have especially high sequence similarities to the homeodomains encoded by the LIM class genes (Fig. 1c). Pair-wise comparisons reveal striking similarity to particular members of the LIM class. The BK87 homeodomain is 83% identical to that of *lin-11*. The BK64 homeodomain has a similar relationship (85% identity) with the homeodomain of a new *C. elegans* member of this family, *ceh-14* (Bürglin et al., 1989; T. Bürglin, personal communication) (Fig. 1c). The BK64 homeodomain also shows a very high degree of relatedness to the vertebrate LIM genes *Xlim-3* (87% identity based on the 39 amino acid partial homeodomain sequence of *Xlim-3*; Taira et al., 1992) and *Gsh-4* (93% identity, Singh et al., 1991 and S. Potter, personal communication). The results of pair-wise comparisons of the LIM homeodomain family members are represented in a phylogenetic tree (Fig. 2), and show that BK64 and BK87 appear to belong to distinct subgroups within the LIM family.

DNA binding specificity

In addition to screening purposes, retention of radioactively labeled oligonucleotides by fusion proteins immobilized on nitrocellulose can be used to assay DNA binding. The ability of all of our fusion products to bind a particular oligonucleotide was assayed by processing a single filter imprinted with protein from all our clones. The sensitivity is excellent. Signal intensities extended to 67 000 cpm. In a test for specificity, an unrelated 32 bp sequence that is recognized by the glucocorticoid receptor was not bound by our clones (data not shown).

To further define binding specificities, we examined binding of a variety of related oligonucleotides, each similarly labeled and concatenated. Since each clone can produce a different amount of fusion protein, absolute amounts of radioactive oligonucleotide bound to the filter imprint derived from each clone could not be compared directly. However, normalizing the binding of different oligonucleotides to the level of LP binding gives a measure of relative binding that can be compared from one clone to another. Engrailed and Ftz-producing control clones gave results consistent with published results. For example, Engrailed binds NP and LP roughly equally and does not bind detectable levels of RP (at least 50-fold less than LP) or the *bicoid* binding site (Table 1).

We tested eight oligonucleotides differing from LP to different extents. The majority of our clones encoded DNA binding activities having a roughly similar spectrum of specificities (see data in Table 1 and summary in Table 2).

All oligonucleotides that include an intact consensus (LP, NP, NPCC, and NP-11) were bound effectively. The eleven base pair NP-11 was the most weakly bound of these oligonucleotides (roughly one third of LP or NP binding); it differs from NP by deletion of a terminal base pair. While this deletion did not impinge on the consensus sequence, it altered site spacing. When

NP-11 is concatenated, sites will be displayed on one side of the helix, and steric impediments to occupancy of adjacent sites are likely.

An ATTA sequence (often written as the complement TAAT) forms the core of the binding sites of a variety of homeodomain proteins (Beachy et al., 1988; Desplan et al., 1988; Hoey and Levine, 1988; Laughon et al., 1988; Driever and Nüsslein-Volhard, 1989). The RP oligonucleotide differs from LP only within this core (AaTA rather than ATTA). Compared to LP, this oligonucleotide has reduced binding to all but three of the clones (BarH1_{bk5}, pdm-1_{bk112}, and bk35, Table 1: these exceptions are discussed below). While the reduction suggests that the core sequence makes a substantial contribution to binding, the varying level of the reduction (extending up to about 50-fold) suggests that its importance varies from one clone to another.

It has been suggested that different homeodomain proteins discriminate between related binding sites based on interactions with sequences flanking the core ATTA sequence (Hanes and Brent, 1991). Oligonucleotides retaining the core ATTA, but altered in their flanking sequences (see oligonucleotides TAA, BCD, and PRD; Table 2) also had reduced binding (Table 1). The magnitude of the reduction varied from one protein to another (compare TAA binding to bk27 and Abd-B_{bk70}; Table 1) and depended on the particular change made in the flanking sequences (compare bk87 binding of TAA, PRD, and BCD; Table 1). Despite this variation, we note that all the clones that we isolated encode proteins with at least a small preference for the flanking sequences found in the screening oligonucleotide. Thus, these proteins have a preference for the flanking sequences found in LP, but it is not apparent what residues of the homeodomain would recognize these flanking sequences (see Discussion).

Three clones are exceptional in that they bind RP better than LP, despite the absence of an ATTA sequence, bk35, which shows the strongest RP preference (about 5-fold), also has other peculiarities in binding specificity. It fails to bind the TAA oligonucleotide (at least 1000-fold less than LP binding), which suggests that it has little or no ability to bind the core sequence ATTA. Comparisons between other oligonucleotides suggest that the protein encoded by this clone preferentially recognizes a sequence (ATGA · TCAT) found at all junctions created by ligation of RP and at 1/4 of the NP junctions (see Table 1), and binds more weakly to the related sequence (tTGA · TCAa) found at all LP junctions, and at 1/4 of the NPCC and NP-11 junctions. The bk35 clone does not encode a homeodomain, but does have a sequence resembling a Zn finger domain (data not shown). We suspect that it was cloned because of a coincidental match between the junction sequences and the binding specificity.

The pdm-1_{bk112} clone, which prefers RP over LP by 4-fold, includes a second domain contributing to DNA binding specificity (POU specific domain). As discussed below, this second DNA binding domain can contribute to RP binding.

The third clone with exceptionally strong binding of RP, BarH1_{bk5}, appears to have a preference for sequences having AATA in place of the core ATTA sequence. The BarH1 homeodomain has unusual residues at two conserved positions. A tyrosine is located at a key structural position (residue 49) occupied by a phenylalanine in virtually all other homeodomains (Fig. 1a). Additionally, one of the residues on the recognition face of helix 3 (position 47) is threonine rather than the isoleucine or valine that usually occupy this position. Perhaps altered DNA contacts by the threonine modify the specificity for the sequence at the core of the site (see Discussion).

The *paired* homeodomain has also been assigned a consensus binding sequence that differs from LP (Treisman et al., 1989, and Table 2). Several of our clones encode homeodomains that have sequence similarity to the *paired* homeodomain (see above). However, binding to the paired-type consensus sequence does not parallel assignments made according to

homologies of protein sequence. While two of the paired class of clones (bk27 and bk36) produce products that bind the paired oligonucleotide reasonably well, the other two (bk50 and bk24) do not (Table 1). Additionally, the two strongest binding activities are encoded by non-paired class clones, abd-A_{bk102} and Abd-B_{bk70} (Table 1). The binding of this oligonucleotide might be explained by the creation of LP like sequences at the junction of ligated paired oligonucleotides, rather than binding to the 'paired-consensus' sequence (Table 2).

Binding specificity of POU class proteins

Two POU domain containing clones (pdm-l_{bk112} and Cfla_{bk54}) were isolated in this screen. Both of these had previously been isolated in screens based on POU sequence similarity (Johnson and Hirsh, 1990; Billin et al., 1991; Lloyd and Sakonju, 1991). In addition to having closely related homeodomains (Fig. 1a) these clones share an additional region of similarity, the POU specific domain. While DNA binding studies with the mammalian POU domain proteins OCT-1, OCT-2, and Pit-1 have demonstrated specificity for the octamer motif, ATGCAAAT, the binding specificity is somewhat degenerate (reviewed in Ruvkun and Finney, 1991). POU class proteins have also been observed to interact with the En consensus (Ingraham et al., 1990). When all of our clones were tested for binding of the octamer motif, only the two POU class proteins bound significantly (Table 2). Since these bound the octamer motif better than they bound the LP sequence with which they were selected, it is obvious that clones encoding DNA binding proteins can be isolated by probing with sequences that bind considerably less well than the optimal binding sequence.

Studies with the mammalian proteins have shown that the POU specific domain contributes to DNA binding (Ingraham et al., 1990; Verrijzer et al., 1990; Aurora and Herr, 1992). However, our screen for LP binding clones identified a series of clones representing truncated versions of clone Cfla_{bk54}, some of which are deleted for the POU specific domain. Tests of the DNA binding of various oligonucleotides to this truncated series reveals that presence, or absence of the POU specific domain has little influence on binding to LP related sequences, but that the POU specific domain is absolutely required for significant binding of the octamer motif (Table 3). Additionally, the presence of the POU specific domain appears to increase RP binding relative to NP. Since the RP oligonucleotide (TCATTTAAATGA) has a six base pair match to the octamer motif, we suspect that the high level of RP binding is related to the octamer binding specificity. In any case, the truncated versions of clone Cfla_{bk54} show that, on its own, the encoded homeodomain has a binding specificity very much like that of non-POU class homeodomains.

Discussion

We undertook this work with the a priori bias that ligand screening with a consensus binding site for the Engrailed homeodomain would predominantly detect proteins having an amino acid sequence related to that of the Engrailed homeodomain. In part, this bias is supported by the retrieval of 17 homeodomain encoding clones out 20 (the three exceptions can probably be attributed to coincidental approximation of the binding specificity of other regulators; see Results). However, in contrast to our starting bias, other than the signature residues of a homeodomain, the clones that were retrieved in this screen bore no particular sequence relationship to the Engrailed homeodomain. The few commonalities in amino acid sequence among the retrieved homeodomains help identify residues contributing to the specificity of DNA binding. In addition, the results suggest that evolution of at least most members of the family of homeodomain regulators occurred without major diversification of DNA binding specificity. Presumably, selective pressure maintained similar binding specificity during evolution.

Specificity determinants in homeodomains

If only a subset of all homeodomains bind the same DNA sequence, this subset ought to share common amino acid residues at positions involved in sequence specific interactions. However, the homeodomains that we isolated were not restricted in sequence diversity at any position (Fig. 3, above). Three results argue that this is not due to a lack of specificity in the screen. First, we recovered primarily homeodomain proteins. Second, binding to control oligonucleotides showed no detectable binding. Third, binding to a variety of related oligonucleotides showed that the proteins specifically recognized a number of positions within the LP site, and that the importance of particular base pairs within the LP oligonucleotide is consistent with importance deduced from the consensus analysis of Engrailed binding to DNA (Desplan et al., 1988, and Table 2). We conclude that the similarity in the binding specificity among the selected homeodomains is representative of the majority of homeodomains.

It should be kept in mind that DNA binding proteins recognize sequences with degenerate relationships to their preferred binding site. Thus, while the LP sequence appears to be a 'universal' site recognized by all our clones, it is likely to represent a compromise among a number of closely related preferred binding sites. Indeed, slight variations in preference were observed.

There are at least a few exceptions to the surprising proposition that most homeodomains bind the LP sequence. The Bicoid, Mat $\alpha 2$, and the recently isolated corn Zmboxla homeodomains bind sequences that differ (Treisman et al., 1989; Hanes and Brent, 1989, 1991; Keleher et al., 1988; Bellmann and Werr, 1992). Here we examine whether a widespread, but not universal homeodomain specificity for LP like sequences can be rationalized in terms of the DNA contacts that have been identified in structural studies (Fig. 5).

The nearly superimposable crystal structures of Engrailed and MAT $\alpha 2$ homeodomains show helix three almost at right angles to the DNA, and nestled in the major groove (Kissinger et al., 1990; Phillips et al., 1991; Wolberger et al., 1991). Four helix three residues (**47, 50, 51** and **54**) have an opportunity to contact the base pairs in the major groove.

Position 51 is the only contact position conserved among all the homeodomains in our collection. Indeed, asparagine is found at this position in almost all homeodomains, including Bicoid, and Mat $\alpha 2$, but not Zmboxla. It makes a sequence specific contact with a T : A' at position 5 of the binding site (Fig. 5), a position conserved in the sites recognized by Bicoid, and Mat $\alpha 2$, but not the site recognized by Zmboxla. Thus, we suggest that this conserved residue contributes part of the conserved specificity. But, clearly the entire specificity is not accounted for by this one conserved residue.

Position 47 appears to make a specific hydrophobic contact with the thymidine methyl position 4' (Kissinger et al., 1990). Most of our homeodomains have valine or isoleucine at this position, both of which are compatible with the suggested hydrophobic interaction. Furthermore, binding of other homeodomains suggest that the residue is important for recognition of position 4. Bicoid has valine at position 47 and retains specificity for A : T at position 4 of the binding site. In contrast, Mat $\alpha 2$ has asparagine at position 47 and G : C at base-pair 4 of the binding site. Zmboxla has lysine at position 47 and has a highly diverged binding site.

In addition to the canonical residues, isoleucine and valine, position 47 of our sequences is occupied once by threonine (BarH1_{bk5}), and once by asparagine (cut). Threonine can mimic isoleucine or valine (Fig. 4). But, BarH1_{bk5} slightly prefers a site in which position 4 is mutated (A : T' to T : A'), suggesting that threonine influences site selectivity. If the side chain of threonine rotates so that its hydroxyl group, rather than its methyl group, is oriented toward the DNA a hydrogen bond can be formed with A 4' (Fig. 4). Thus, we suggest that threonine

at position 47 is a permissive residue that allows interaction with both A : T' and T : A' (and perhaps C : G'; see Fig. 4) at position 4 of the binding site. The asparagine at position 47 of the *cut* homeodomain might be even more permissive. Asparagine at this position of the Mat $\alpha 2$ sequence does not make notable contacts with the base-pairs.

Position 50 has been presented as a key determinant of homeodomain specificity based on findings that exchange of glutamine for lysine at this position alters recognition of base pairs at positions 2 and 3 of the site (Treisman et al., 1989; Hanes and Brent, 1989; Hanes and Brent, 1991). However, neither the crystal structure of Engrailed, nor that of Mat $\alpha 2$ reveal a specific contact between the side chains at position 50 (glutamine and serine, respectively) and the bases. Furthermore, our results show that glutamine, cysteine, serine, and histidine are interchangeable with no marked effect on discrimination among a number of sequences (see also Ingraham et al., 1990). We suggest that lysine at position 50 severely restricts the sequences bound, but that a more permissive residue (see discussion of Florence, Handrow and Laughon, 1991) relaxes recognition of base pair 2 and 3.

Position 54 does not appear to have a conserved role. In the MAT $\alpha 2$ structure, the long side chain of arginine 54 reaches back to make contact with the base pairs (Phillips et al., 1991), but in Engrailed this position is occupied by alanine whose short side chain does not participate in these interactions. Additionally, this position is not conserved among our sequences.

In addition to contacts in the major groove, the homeodomain has contacts in the minor groove and extensive contacts to the back bone (Otting et al., 1990; Kissinger et al., 1990; Wolberger et al., 1991). Although such contacts have less opportunity to impart sequence specificity, even subtle contributions are significant when summed over several contacts. Since several of these contacts are conserved, they could contribute to a conserved sequence specificity.

Comparison of binding of LP and TAA oligonucleotides (Tables 1 and 2) shows that similarity in binding specificity includes positions beyond the ATTA core sequence found in the consensus binding sites of several homeodomains (Beachy et al., 1988; Desplan et al., 1988; Hoey and Levine 1988; Laughon et al., 1988; Driever and Nüsslein-Volhard, 1989).

Evolutionary divergence of DNA binding specificity

If evolution selects for diversification of DNA binding specificities of homeodomains, distinct specificities ought to prevail. While our data do not preclude small differences in binding specificity, homeodomains, even those with highly diverged amino acid sequences, have retained a capacity to recognize closely related sites (Fig. 1). Thus, diversification of homeodomain regulators has been associated with only marginal divergence of the DNA binding specificity of homeodomains (note that we refer, here, to the autonomous specificity of the homeodomain, not the physiological specificity of the intact protein in the context of other possible accessory factors).

Related sequence specificities might be unavoidable if the structure of the homeodomain is uniquely suited to binding a DNA sequence having a particular conformation. However, there is nothing notable about the conformation of the DNA in homeodomain: DNA complexes (Otting et al., 1990; Wolberger et al., 1991; Kissinger et al., 1990). Furthermore, yeast Mat $\alpha 2$ binds a substantially different sequence even though its structure can be almost superimposed on that of Engrailed (Wolberger et al., 1991).

Selective pressures could dictate whether binding specificity diverges or is maintained. For example, because lambdoid phage compete for hosts, distinct immunity has a selective advantage that might have driven the evolution of different binding specificities that characterizes different lambdoid repressors. In contrast to the differences in binding specificity

among the various lambdoid repressors, each lambdoid repressor shares sequence specificity with its cognate cro protein. This similarity in repressor and cro specificities is presumably maintained by selection, since the function of the immunity system requires that cro and repressor compete for binding to the sites of the immunity region (Ptashne, 1987).

Metazoan homeodomains function in a coordinated network of control (O'Farrell et al., 1985; Scott and O'Farrell, 1986; Scott and Carroll, 1987). Overlapping target specificity would tie different regulators in a common regulatory network. Thus, participation in such a network might limit evolutionary divergence of DNA binding specificities, much as the divergence of λ repressor and cro binding specificities is constrained by their function in a much smaller network of control. One might anticipate that, like λ repressor and cro, slight variation in binding specificity, and differences in dimerization or protein:protein interaction might be important to the unique regulatory roles of the different proteins.

Since plant pattern formation is not based on the same network of homeotic genes as animal pattern formation, plant homeodomains are not likely to have faced the same evolutionary constraints. Perhaps the exceptional divergence of the DNA binding specificity of some of the plant homeodomains is a reflection of this difference (Bellmann and Werr, 1992).

Extensive conservation of homeodomain sequences between species

Individual members of the homeodomain family are conserved over long evolutionary times. For example, homeodomains of *Drosophila* homeotic genes share about 90% sequence identity with their vertebrate counterparts (for examples see Scott et al., 1989). Sequence conservation is associated with extraordinary parallels in gene organization, expression and function in diverse species (McGinnis and Krumlauf, 1992; Malicki et al., 1990; Malicki et al., 1992; McGinnis et al., 1990). Genes having less sequence similarity have less evident similarities in expression and function.

Two of our newly identified genes are highly conserved (Fig. 1c and Fig. 2). The BK64 homeodomain is 93% identical to the homeodomain of the mouse *Gsh-4* gene, and 85% identical to the homeodomain of the *C. elegans ceh-14* gene (Singh et al., 1991, and S. Potter, personal communication; Bürglin et al., 1989, and T. Bürglin, personal communication), while BK87 shares 83% identity with the *C. elegans lin-11* homeodomain (Freyd et al., 1990). While little is known about these genes, their sequence similarity suggests related biological roles. They appear to be especially conserved members of the LIM class of genes. This class was defined when the *C. elegans* genes *lin-11* and *mec-3*, and the rat *Isl-1* gene were found to share a conserved sequence rich in histidine and cysteine, as well as having related homeodomains (Freyd et al., 1990). *lin-11* is involved in the control of differentiation of some terminal cell types, and it has been suggested that this gene has a special role in lineage-directed cell fate decisions (Ferguson and Horvitz, 1985; Ferguson et al., 1987). The *bk64* and *bk87* genes are expressed in subsets of cells in the nervous system consistent with a role in directing cell fate decisions (B.K., unpublished). Perhaps the analysis of their action will uncover some universal controls involved in lineage-directed cell fate choice.

The extraordinary conservation of homeodomain sequences between distantly related species implies that divergence of most amino acid residues is restricted. As noted above, conservation of binding specificity does not place a strong constraint on sequence divergence. Consequently, the conservation of amino acid sequence and the striking parallels in function of homeodomains in different organisms probably reflects constraints imposed by other conserved interactions. Indeed, the conservation of virtually all residues of the homeodomain between distantly related species suggests that there are numerous evolutionary constraints, presumably due to homeodomain involvement in many conserved interactions.

Protein: protein interactions and combinatorial specification of DNA binding

The conservation of homeodomain binding specificity that we and others (Cho et al., 1988; Desplan et al., 1988; Hoey and Levine, 1988; Jaynes and O'Farrell, 1988; Treisman et al., 1989) have observed is surprising given that even closely related homeodomains have been shown to harbor crucial functional distinctions (Kuziora and McGinnis, 1989; Mann and Hogness, 1990). Perhaps the crucial distinctions are not in the DNA contacts, but are differences in the interactions of different homeodomains with other actors, in particular other protein domains and other proteins that modify DNA interactions. The POU class proteins have a second conserved domain, the POU specific domain that serves such a modifying role (Herr et al., 1988). Although it has no detectable DNA binding of its own, this POU specific domain plays a role in DNA binding (Ingraham et al., 1990). However, there has been some controversy whether it is an essential or a modifying role. In agreement with Strum and Herr (1988), truncations of our Cfla_{BK54} clone show that the POU specific domain is essential for binding to the octamer motif. However, this domain is not required for binding to sequences related to LP. Indeed, when on its own, the homeodomain of this protein has a DNA binding specificity just like that of other homeodomains that we have examined. Thus, the POU domain modifies the more or less standard binding specificity of the POU homeodomain (see also Ingraham et al., 1990, and Verrijzer et al., 1990).

Other homeodomain proteins may have different domains that add specificity. For example, the homeodomain of the *paired* protein is associated with a second DNA binding domain encoded by the paired box (Bopp et al., 1986). Of the several clones that we isolated with homeodomains related to *paired*, at least one, BK27, contains a paired box (Jun, Kalionis and Desplan, unpublished results). Additionally, the *zfh* genes have Zn finger domains as well as homeodomains (Fortini et al., 1991).

In addition to modifications of specificity due to other domains within the same protein, interaction with a second protein can add specificity. In the yeast, *S. cerevisiae*, three homeodomain proteins, Mat α 2, Mat α 1 and Bas 2, function with partner proteins, and different partner proteins can direct binding to different sites (Keleher et al., 1988; Tice et al., 1989; Dranginis, 1990). Recently, a human homeodomain protein, Phox1, has been shown to interact with and promote the DNA binding of the serum response factor, SRF (Grueneberg et al., 1992). Since Phox1 can also interact with the yeast MCM1 gene product, an SRF homolog, it appears that this is a highly conserved interaction. In another well characterized example, the viral regulatory protein VP-16 associates with the homeodomain of Oct-1 and alters DNA binding specificity and drives transcriptional activation (Tanaka et al., 1988; Stern et al., 1989). Based on these precedents, and our finding that many metazoan homeodomain proteins have highly related binding specificity, we suggest that the distinctive regulatory roles of metazoan homeoproteins are largely attributable to association of the homeodomains with other domains or other proteins that modify binding specificity or function.

Materials and Methods

Procedures for screening cDNA expression libraries with DNA binding site probes have generally met with low levels of success. In our experience, several major factors such as the choice of the library and the level of protein expression upon induction of fusion protein synthesis, greatly influence the outcome. The use of duplicate filters which employ non-denaturing and denaturing/renaturing treatments to minimize false positives as well as false negatives is recommended. Additionally, screening at low plating densities, use of probes of a defined size, and addition of detergent to the wash buffer appear to have contributed to the success of this screen.

Probe preparation

The DNA binding site used for screening was the self-complementary DNA binding site LP 5'(TCAAT-TAATTGA)3' (Desplan et al., 1988). 15 pmol of the oligonucleotide was kinased in a final volume of 20 μ l containing 50 mM Tris-HCl (pH 7.8), 10 mM MgCl₂, 10 mM DTT, 10 U T4 polynucleotide kinase (Biolabs) and 0.5 to 1.0 ml of ³²P- γ -dATP (7000 Ci/mmol) at 37°C for 60 min after which 400 U of T4 DNA ligase (Biolabs) and ATP to 1 mM was added and the volume adjusted to 25 μ l. The reaction proceeded for approximately 30 min at 12–14°C for blunt-ended oligonucleotide binding sites. Unincorporated nucleotides were removed by passing the sample through two consecutive Sephadex (Pharmacia) G-25 spin columns. An aliquot of each reaction was fractionated on a 10% polyacrylamide/8 M urea gel and autoradiographed. Ladders containing predominantly monomers to 20-mers were used for screening.

Library screen

Screening of a 9 to 12 hour amplified λ gt11 embryonic cDNA expression library (a gift from C. Goodman) (Zinn et al., 1988) was carried out in duplicate. Plating density was kept low (10–20000 plaques per plate) to allow maximal expression of the fusion protein. Plates were incubated for 3 h at 42°C and then a sheet of Optibind nitrocellulose (Schleicher and Schuell) soaked in 10 mM IPTG, and air dried, was placed onto the overlay. The plates were incubated for a further 3 h at 37°C and then the filters removed. A second filter treated with IPTG, as above, was placed onto the plates and incubated for a further 6 h. All filters were air dried for at least 20 min. Filters were incubated in containing 70 mM NaCl, 10 mM KPO₄, 0.5% Triton X-100, 5% fat free milk powder and 100 μ g/ml sonicated, denatured salmon sperm DNA for 60 min at room temperature. DNA binding reactions were carried out in a buffer containing 70 mM NaCl, 10 mM Tris-HCl (pH 7.6), 1 mM EDTA, 0.5 mM EGTA, 1 mM DTT, 10% glycerol, 0.25% fat-free milk powder and 100 μ g/ml sonicated, denatured salmon sperm DNA. The DNA binding reaction contained approximately 25 ng/ml of concatenated DNA binding site (2.5×10^6 cpm/ml). After incubating at 4°C overnight, the filters were batch washed three times in a buffer containing 70 mM NaCl, 10 mM Tris-HCl (pH 7.6), 1 mM EDTA, 0.5 mM EGTA, 1 mM DTT, 0.25% fat-free milk powder and 0.1% Triton X-100 for 10 min at 4°C. The duplicate set of filters was denatured and subsequently renatured using the protocol essentially as described by Vinson et al. (1988) except that 0.1% Triton X-100 was added to the wash buffer. The plaques were purified through two more rounds of purification using the denaturation/renaturation protocol. In general, the signals generated using the denaturation/renaturation treatment were stronger by up to several fold, however in several cases signals were stronger with the non-denatured filters (data not shown).

Clone characterization and DNA sequencing

The insert DNA from each λ clone was isolated and restriction analysis was used to identify 20 different classes. Insert DNA corresponding to at least one member of each class was subcloned into the *Eco* RI site of the Bluescript plasmid (Stratagene). The homeobox sequence was identified by sequencing with degenerate oligonucleotide primers made to the highly conserved regions of the homeobox. Oligonucleotides used were HB-1, PRD-1, PRD-2, POU-1, POU-2 (Bürglin et al., 1989), EN-5, PRD-2, HOM-2, HOM-2', HOM-3 (Kamb et al., 1989), P1 and P2 (Bodmer et al., 1990). After obtaining the DNA sequence with a degenerate oligonucleotide primer, complementary oligonucleotide were synthesized and used to generate the sequence across the entire homeodomain in both directions. Those clones which gave no DNA sequence with the degenerate oligonucleotide primers were sequenced across the entire insert.

DNA binding assay

Agar plates containing an overlay of *E. coli* strain Y1090 were prepared. The titers of the λ gt11 phage clones were adjusted so that 5 μ l of a dilution stock would give approximately 20 plaques when spotted (1-cm diameter spots in a grid formation). Plates were incubated for 3 h at 42°C and then a sheet of Optibind nitrocellulose (Schleicher and Schuell) soaked in 10 mM IPTG was placed onto the overlay. The plates were incubated for a further 6 h at 37°C. The filters were removed and subjected to the denaturation/renaturation protocol of Vinson et al. (1988). Prebinding conditions, binding of the probe and washing conditions were as described above. The DNA binding site probes were prepared as described above, however the degree of ligation of the sites varied and conditions used to generate an optimal ladder (monomers to 20-mers) were chosen after varying both the time of the reaction and the temperature from that indicated above (data not shown). The sites used here were LP-TCAATTAATTGA, NP-TCAATTAAATGA, NPCC-TCAATTAAATCC, RP-TCATTTAAATGA, TAA-TAATAATAATAATAA, PRD-GATTTGACGTAA, BCD-TCTAATCCCC, OCT-ATGCAAAT, GRE-GTCGACTGTACAGGATGTTCTAGCTACTCGAG. The DNA binding reaction contained approximately 10 ng/ml of concatenated DNA binding site (4×10^5 cpm/ml). After the filters were washed and air dried, the grid on the filter was used as a template to cut each filter into squares corresponding to each of the spotted phage clones and each square was placed into a scintillation vial. Two ml of scintillation fluid was added and the radioactivity in each square was determined. The λ gt11 *engrailed* homeodomain-*lacZ* fusion construct used as a control was derived from the pUR290 homeodomain construct A (Desplan et al., 1988), which encodes amino acids 410 to 512 of Engrailed. Briefly, the entire insert was isolated after EcoRI digestion and subcloned into *Eco* RI digested λ gt11. A clone which produced *engrailed* homeodomain fusion protein was obtained by screening with an antibody to detect Engrailed (gift of S. DiNardo) essentially as described by Huynh et al. (1985). The λ gt11 *ftz* homeodomain-*lacZ* fusion construct was described previously (Carroll and Scott, 1985). The fusion protein contains 399 amino acids of the *ftz* gene product but lacks five N-terminal and nine C-terminal residues.

Acknowledgments

We thank Thomas Brglin and Rolf Bodmer for supplying the degenerate oligonucleotides, Thomas Brglin and Steve Potter for providing unpublished sequence information, and Claude Desplan for supplying many of the oligonucleotides. We thank our colleagues at UCSF and in Adelaide for their comments on the manuscript. B.K. was supported by a NH & MRC C.J. Martin Fellowship, and the work was funded by a NIH grant to P. OF.

References

- Aurora R, Herr W. Mol Cell Biol 1992;12:455–467. [PubMed: 1732727]
- Beachy PA, Krasnow MA, Gavis ER, Hogness DS. Cell 1988;55:1069–1081. [PubMed: 2904838]
- Bellmann R, Werr W. EMBO J 1992;11:2267–2274.
- Billin AN, Cockerill KA, Poole SJ. Mech Dev 1991;34:75–84. [PubMed: 1680380]
- Bodmer R, Jan LY, Jan YN. Development 1990;110:661–669. [PubMed: 1982429]
- Bopp D, Burri M, Baumgartner S, Frigerio G, Noll M. Cell 1986;47:1033–1049. [PubMed: 2877747]
- Brglin TR, Finney M, Coulson A, Ruvkun G. Nature 1989;341:239–243. [PubMed: 2571091]
- Carroll SB, Scott MP. Cell 1985;43:47–57. [PubMed: 3000605]
- Cho KW, Goetz J, Wright CV, Fritz A, Hardwicke J, DeRobertis E. EMBO J 1988;7:2139–2149. [PubMed: 2901347]
- Cohen B, McGuffin ME, Pfeifle C, Segal D, Cohen SM. Genes Dev 1992;7:715–729. [PubMed: 1349545]
- Dalton D, Chadwick R, McGinnis W. Genes Dev 1989;3:1940–1956. [PubMed: 2576012]
- Desplan C, Theis J, O'Farrell PH. Nature 1985;318:630–635. [PubMed: 4079979]
- Desplan C, Theis J, O'Farrell PH. Cell 1988;54:1081–1090. [PubMed: 3046753]

- Dranginis AM. *Nature* 1990;347:682–685. [PubMed: 1977088]
- Driever W, Nüsslein-Volhard C. *Nature* 1989;337:138–143. [PubMed: 2911348]
- Ferguson EL, Horvitz HR. *Genetics* 1985;110:17–72. [PubMed: 3996896]
- Ferguson EL, Sternberg PW, Horvitz HR. *Nature* 1987;326:259–267. [PubMed: 2881214]
- Fortini ME, Lai ZC, Rubin GM. *Mech Dev* 1991;34:113–122. [PubMed: 1680376]
- Freyd G, Kim SK, Horvitz HR. *Nature* 1990;344:876–879. [PubMed: 1970421]
- Grueneberg DA, Natesan S, Alexandre C, Oilman MZ. *Science* 1992;257:1089–1095. [PubMed: 1509260]
- Hanes SD, Brent R. *Cell* 1989;57:1275–1283. [PubMed: 2500253]
- Hanes SD, Brent R. *Science* 1991;251:426–430. [PubMed: 1671176]
- Herr W, Sturm RA, Clerc RG, Corcoran LM, Baltimore D, Sharp PA, Ingraham HA, Rosenfeld MG, Finney M, Ruvkun G, Horvitz HR. *Genes Dev* 1988;2:1513–1516. [PubMed: 3215510]
- Hoey T, Levine M. *Nature* 1988;332:858–861. [PubMed: 2895896]
- Hoey T, Warrior R, Manak J, Levine M. *Mol Cell Biol* 1988;8:4598–4607. [PubMed: 2905420]
- Huynh, TV.; Young, RA.; Davis, RW. *DNA Cloning: A Practical Approach*. Glover, DM., editor. Vol. I. IRL Press; Oxford: 1985. p. 49–78.
- Ingraham HA, Flynn SE, Voss JW, Albert VR, Kapiloff MS, Wilson L, Rosenfeld MG. *Cell* 1990;61:1021–1033. [PubMed: 2350782]
- Jaynes JB, O'Farrell PH. *Nature* 1988;336:744–749. [PubMed: 2905023]
- Johnson PF, McKnight SL. *Annu Rev Biochem* 1989;58:799–839. [PubMed: 2673023]
- Johnson WA, Hirsh J. *Nature* 1990;343:467–470. [PubMed: 1967821]
- Kamb A, Weir M, Rudy B, Varmus H, Kenyon C. *Proc Natl Acad Sci USA* 1989;86:4372–4376. [PubMed: 2734290]
- Karlsson O, Thor S, Norberg T, Ohlsson H, Edlund T. *Nature* 1990;344:879–882. [PubMed: 1691825]
- Keleher CA, Goutte C, Johnson AD. *Cell* 1988;53:927–936. [PubMed: 3289753]
- Kissinger CR, Liu B, Martin-Blanco E, Kornberg TB, Pabo CO. *Cell* 1990;63:579–590. [PubMed: 1977522]
- Kojima T, Ishimaru S, Higashijima S, Takayama E, Akimaru H, Sone M, Emori Y, Saigo K. *Proc Natl Acad Sci USA* 1991;88:4343–4347. [PubMed: 1674606]
- Laughon A, Howell W, Scott MP. *Development* 1988;104(Suppl):85–93. [PubMed: 3151689]
- Lloyd A, Sakonju S. *Mech Dev* 1991;36:87–102. [PubMed: 1685891]
- Malicki J, Cianetti LC, Peschle C, McGinnis W. *Nature* 1992;358:345–347. [PubMed: 1353609]
- Malicki J, Schughart K, McGinnis W. *Cell* 1990;63:961–967. [PubMed: 1979525]
- McGinnis N, Kuziora MA, McGinnis W. *Cell* 1990;63:969–976. [PubMed: 1979526]
- McGinnis W, Hart CP, Gehring WJ, Ruddle FH. *Cell* 1984;38:675–680. [PubMed: 6091896]
- McGinnis W, Krumlauf R. *Cell* 1992;68:283–302. [PubMed: 1346368]
- O'Farrell PH, Desplan C, DiNardo S, Kassis JA, Kuner JM, Sher E, Theis J, Wright D. *Cold Spring Harbor Symp Quant Biol* 1985;50:235–42. [PubMed: 3006988]
- Otting G, Qian YQ, Billeter M, Muller M, Affolter M, Gehring WJ, Wuthrich K. *EMBO J* 1990;9:3085–3092. [PubMed: 1976507]
- Pabo C, Sauer RT. *Annu Rev Biochem* 1984;53:293–321. [PubMed: 6236744]
- Phillips CL, Vershon AK, Johnson AD, Dahlquist FW. *Genes Dev* 1991;5:764–772. [PubMed: 1673952]
- Ptashne, M. *A Genetic Switch*. Cell Press and Blackwell Scientific Publications; Cambridge: 1987. p. 33–48.
- Ruvkun G, Finney M. *Cell* 1991;64:475–478. [PubMed: 1991317]
- Scott MP, Carroll SB. *Cell* 1987;51:689–698. [PubMed: 2890437]
- Scott MP, O'Farrell PH. *Annu Rev Cell Biol* 1986;2:49–80. [PubMed: 2881561]
- Scott MP, Tamkun JW, Hartzell GW III. *Biochim Biophys Acta* 1989;989:25–48. [PubMed: 2568852]
- Scott MP, Weiner A. *Proc Natl Acad Sci USA* 1984;81:4115–4119. [PubMed: 6330741]

- Singh G, Kaur S, Slock JL, Jenkins NA, Gilbert DJ, Copeland NG, Potter SS. *Proc Natl Acad Sci USA* 1991;88:10706–10. [PubMed: 1683707]
- Singh H, LeBowitz JH, Baldwin AS, Sharp PA. *Cell* 1988;52:415–423. [PubMed: 2964277]
- Stern SA, Tanaka M, Herr W. *Nature* 1989;341:624–630. [PubMed: 2571937]
- Struhl K. *Trends Biochem Sci* 1989;14:137–140. [PubMed: 2499084]
- Sturm RA, Herr W. *Nature* 1988;336:601–614. [PubMed: 2904656]
- Taira M, Jamrich M, Good PJ, Dawid IB. *Genes Dev* 1992;6:356–366. [PubMed: 1347750]
- Tanaka M, Grossniklaus U, Herr W, Hernandez N. *Genes Dev* 1988;2:1764–1778. [PubMed: 2853685]
- Tice BK, Fink GR, Arndt KT. *Science* 1989;246:931–935. [PubMed: 2683089]
- Treisman J, Gönczy P, Vashishtha M, Harris E, Desplan C. *Cell* 1989;59:553–562. [PubMed: 2572327]
- Verrijzer CP, Kal AJ, van der Vliet PC. *Genes Dev* 1990;4:1964–1974. [PubMed: 1980478]
- Vincent JP, Kassis JA, O'Farrell PH. *EMBO J* 1990;9:2573–2578. [PubMed: 1973384]
- Vinson CR, LaMarco KL, Johnson PF, Landschulz WH, McKnight SL. *Genes Dev* 1988;2:801–806. [PubMed: 3061875]
- Wolberger C, Vershon AK, Liu B, Johnson AD, Pabo CO. *Cell* 1991;67:517–528. [PubMed: 1682054]
- Zinn K, McAllister L, Goodman CS. *Cell* 1988;53:577–587. [PubMed: 3370670]

(a)

Conserved:	** R	* * Q L *	Y	* ** * L	* ** * W F * N * R	* **	% identity
	10	20	30	40	50	60	
<i>en</i>	EKRPRTA	FSS EQLARLKREF	NENRYLTERR	RQQLSSELGL	NEAQIKIWFQ	NKRAKIKKST	
<i>bk50</i>	HR-N--T-TT	Y--HE-E-A-	EKSH-PDVYS	-EE-AMKVN-	P-VRVQV---	-R---WRARR	35
<i>bk36</i>	-H--T-TQ	----E-EAA-	AKSH-PDIYC	-EE-ARTTK-	---R-QV---	-R---YR-QE	43
<i>bk24</i>	Q--H--R-TP	A--NE-E-C-	SKTH-PDIFM	-EEIAMRI--	T-SRVQV---	-R---W--RK	40
<i>bk27</i>	FR-N--T--P	---EE-EK--	DKSH-PCVST	-ER---RTS-	S--RVQV--S	-R---WRRHQ	42
<i>bk87</i>	KR-V--TIKA	K--EV--TA-	-QTPKP-RHI	-E--AK-T--	PMRV-QV---	---S-ERRMK	40
<i>bk64</i>	N-----TITA	K--ET--TAY	-NSPKPARHV	-E---QDT--	DMRVVQV---	-R---E-RLK	40
<hr/>							
<i>pdm-1 bk112</i>	RRKK--SIET	TIRGA-EKA-	LA-QKP-SEE	IT--ADR-SM	EKEVVVRV--C	-R-Q-E-RIN	25
<i>Cfla bk54</i>	KRKK--SIEV	SVKGA-EQH-	HKQPKPSAQE	ITS-ADS-Q-	EKEVVVRV--C	-R-Q-E-RM-	23
<i>zfh-1 bk96</i>	KV-V---INE	--QQ--QHY	SL-ARPSRDE	FRMIAAR-Q-	DPRVVQV---	-N-SRER-MQ	28
<i>zfh-2 III bk69</i>	N--L--TILP	---NF-YECY	QSESNPSRKM	LEEI-KKVN-	KKRVVQV---	-S---D---R	33
<i>BarH1 bk5</i>	QRKA----TD	H--QT-EKS-	ERQK--SVQE	-E-AHK-D-	SDC-V-T-Y-	-R-T-W-RQ-	40
<i>ems bk95</i>	P--I-----P	S--LK-EHA-	ES-Q-VVGAE	-KA-AQN-N-	S-T-V-V---	-R-T-H-RMQ	45
<i>labial bk79</i>	NNSG--N-TN	K--TE-EK--	HF-----RA-	-IEIANT-Q-	--T-V-----	-R-M-Q--RV	50
<i>abd-A bk102</i>	RR-G-QTYTR	F-TLE-EK--	HF-H---R--	-IEIAHA-C-	T-R-----	-R-M-L--EL	47
<i>Abd-B bk70</i>	VRKK-KPY-K	F-TLE-EK--	LF-A-VSKQK	-WE-ARN-Q-	T-R-V-----	-R-M-N--NS	40
<i>cut bk88</i>	S-KQ-VL--E	--KEA-RLA-	ALDP-PNVGT	IEF-AN----	ATRRT-TN--H	-H-MRL-QQV	33

(b)

<i>prd</i>	QRRCRTTFS	A SQLDELERAF	ERTQYPDIYT	REELAQRNTL	TEARIQVWFS	NRRARLRKQH	% identity
<i>bsh4</i>	---S---T-	E--EA-----	S-----V--	-----T-A-	-----	-----HS	80
<i>bsh9</i>	---S---N	D-I-A---I-	A-----V	-----S-G-	-----	-----L	80
<i>bk50</i>	H N T T Y H	KSH V S	MKV I V V Q	---KW ARR			62
<i>bk36</i>	-H---TQ	E--A---A-	AKSH-----C	-----RT-K-	N-----Q	---KYKQEK	62
<i>bk27</i>	F--N---P	E--E---KE-	DKSH--CVS-	--R-SS--S-	S---V-----	---KW-RHQ	58
<i>bk24</i>	-K-H--R-TP	A--N---C-	SK-H---FM	---I-M-IG-	--S-V---Q	---KWK-RK	58
<i>Antp</i>	RK-G-Q-YTR	Y-TL---KE-	HFNR-LTRRR	-I-I-HALC-	--RQ-KI--Q	---MKWK-EN	37
<i>en</i>	EK-P--A--S	E--AR-K-E-	NENR-LTERR	-QQ-SSELG-	N--Q-KI--Q	-K--KIK-ST	38

(c)

<i>lin-11</i>	RRGPRTTIKA	KQLETLKNAF	AATPKPTRHI	REQLAAETGL	NMRVIQVWFQ	NRRSKERRMK	% identity
<i>mec-3</i>	-----Q	N--DV-NEM-	SN---SK-A	-AK--L---	S-----	-----L-	72
<i>Isl-1</i>	TTRV--VLNE	---H--RTCY	--N-R-DALM	K--VEM---	SP--R---	-K-C-DKKRS	45
<i>Xlim-1</i>	-----A-	-----	-----Q-	-----	-----	-----	97
<i>ap</i>	TKRM--SF-H	H--R-M-SY-	-INHN-DAKD	LK--SQK--	PKA-L----	-A-A-W--M	45
<i>ceh-14</i>	NKR-----S-	S-----Q-Y	QTSS--A-V	-----S---	D--V-----	A--K--L	68
<i>Xlim-3</i>	-----T	-----Y	DNS--A-V	-----SS-	D--V-----	A--K--L	72
<i>bk64</i>	NKR-----T-	-----T-Y	NNS--A-V	-----SQD-	D--V-----	A--K--L	68
<i>bk87</i>	K-RV-----	-----V--T-	NQ-----	-----K---	P-----	-K-----	83
<i>Antp</i>	-KRG-Q-YTR	Y-TLE-EKE-	HFNRYL--RR	-IEI-HALC-	TE-Q-KI---	---M-WKKEN	33
<i>en</i>	EKR--AFSS	E--AR--RE-	NENRYL-ERR	-Q--SS-L--	-EAQ-KI---	-K-A-IKKST	38

Fig. 1.

Amino acid sequences of the homeodomains. (a) Comparison of homeodomain sequences obtained in this screen to the *engrailed* homeodomain. Clones below the line contained homeodomain sequences which were identical at both the protein and DNA level to previously published sequences. The name originally assigned to the clones in this screen is given in subscript. The regions shown to correspond to α -helical regions of the homeodomain are indicated, as are the highly conserved (upper case) and moderately conserved (asterisk) residues. Amino acids that are identical to the corresponding amino acid in the *engrailed* homeodomain are indicated by a dash. The percentage of identical residues to that of the *engrailed* homeodomain is shown on the right. (b) Comparison of homeodomains encoded by

bk50 bk24 bk27 bk36 with the paired class of homeodomain proteins. Amino acids that are identical to the *paired* homeodomain are indicated by a dash. The percentage of identical residues to that of the *paired* homeodomain is shown on the right. Homeodomain sequences from the engrailed and Antennapedia class (*Antp*) are also presented for comparison. (c) Comparison of the homeodomains encoded by *bk87* and *bk64* with the LIM homeodomains. Amino acids that are identical to the *lin-11* homeodomain are indicated by a dash. The percentage of identical residues to that of the *lin-11* homeodomain is shown on the right. Homeodomain sequences from the engrailed and Antennapedia class are also presented for comparison. Sources of homeodomain sequences used for comparisons were described in the text. The *apterous* (*ap*) homeodomain sequence has been placed into the LIM class of homeodomains (Cohen et al., 1992).

Fig. 2.

Phylogenetic tree based on pair-wise comparisons. Homeodomain sequences from the LIM, engrailed and Antennapedia (Antp) classes are presented. The percentage sequence identity between homeodomain sequences at the branch points are shown on the scale to the left. In some instances there are several different pair-wise comparisons that could be used to establish the level of sequence divergence at the branch point. There was a high level of concordance among these comparisons, but in a few cases the branch point that is diagrammed here represents a compromise among comparisons that differ by several percent. The thicker lines interconnect individual homeodomains shown to bind the Engrailed consensus sequence in the present study. The figure emphasizes that homeodomain sequences have been remarkably well conserved between species, despite the fact that such conservation is not required to maintain related binding specificity. See Fig. 1 and text for sources of homeodomain sequences used for the comparisons.

Fig. 3.

Comparison of the amino acid residues at each position in the homeodomain and their frequency of occurrence. A compilation of all amino acids that appear at each of 60 positions of the homeodomain (above) for the homeodomain sequences isolated in this screen (from Fig. 1). and (below) for a control set of an equal number of homeodomain sequences considered to be very diverged (Antp, bed, cad, cut, H2.0, k8, mab5, JML1001, mec3, ro, zen1, zen2, AbdB, eve, gsbBSH4, and hox1.5), taken from the compilation of Scott et al. (1989). The frequency of occurrence of each amino acid is presented as a subscript. The numbering of amino acids along the homeodomain is according to Wolberger et al. (1991). The conservation of particular residues in our set of selected homeodomains could be imposed by selection for binding specificity, or it could reflect conservation common to all homeodomains. In this latter situation, these positions would also be conserved in homeodomains that had not been selected for binding to LP. The pattern of conservation and divergence in the selected and unselected sets is comparable.

Fig. 4.

Model for threonine 47 recognition of a T : A base pair. A model for the proposed permissive interactions of threonine with either A : T or T : A base-pairs. We suggest that threonine 47 can make a hydrophobic contact with the methyl group of thymidine (left side of the figure) similar to that made by isoleucine 47 in the Engrailed: DNA crystal structure (Kissinger et al., 1990). However, unlike isoleucine, threonine at position 47 also promotes interaction with a T : A base-pair (see text). This could be achieved as shown on the right side of the figure. Note that the relative positions of the DNA (represented by the attachment point of the bases to the backbone; the X's) and the position of the α carbon of the threonine side chain are the same in the left and right parts of this figure. When the hydroxyl group of threonine is directed toward the DNA, a hydrogen bond can form between the hydroxyl of threonine and the ring nitrogen (N9) of adenosine.

	p	X ₀	p	T ₁	p	C ₂	p	A ₃	p	A ₄	p	T ₅	p	T ₆	p	A ₇	p	A ₈
	p	X ₀ '	p	A ₁ '	p	G ₂ '	p	T ₃ '	p	T ₄ '	p	A ₅ '	p	A ₆ '	p	T ₇ '	p	T ₈
PO ₄	R ₃₁		R ₅₃		K ₅₇ Y ₂₅							T ₆ F ₈		W ₄₈		K ₅₅		
M					Q ₅₀				I ₄₇		N ₅₁							
m														R ₃		R ₅		
<u>Common Specificity</u>																		
BK50	-	-	-	-	R ₅₇ -	-	-	-	V ₄₇	-	-	-	-	-	-	-	-	-
BK36	-	-	-	-	R ₅₇ *	-	-	-	V ₄₇	-	-	-	-	-	-	-	-	-
BK24	-	-	-	-	-	-	-	-	V ₄₇	-	-	-	-	-	-	-	-	-
BK27	-	-	-	-	R ₅₇ -	S₅₀	-	-	V ₄₇	-	-	-	-	-	-	-	-	-
BK87	-	-	-	-	R ₅₇ K ₂₅ -	-	-	-	V ₄₇	-	-	I ₉	-	-	-	-	-	-
BK64	-	-	-	-	-	K ₂₅ -	-	-	V ₄₇	-	-	I ₈	-	-	-	-	-	-
pdm-1	I ₃₁	-	-	-	-	K ₂₅ C₅₀	-	-	V ₄₇	-	-	I ₈	K ₃	-	-	-	-	-
Cf1	I ₃₁	-	-	-	-	K ₂₅ C₅₀	-	-	V ₄₇	-	-	I ₈	K ₃	-	-	-	-	-
zfh1	F ₂₁	-	-	-	R ₅₇ R ₂₅ -	-	-	-	V ₄₇	-	-	I ₈	-	-	-	-	R ₅₅	-
zfh2	L ₃₁	-	-	-	-	N ₂₅ -	-	-	V ₄₇	-	-	I ₈	-	-	-	-	-	-
Bar	-	-	-	-	-	-	-	-	T₄₇	-	-	-	K ₃	-	-	-	-	-
lab	-	-	-	-	-	-	-	-	V ₄₇	-	-	-	-	-	-	-	-	-
ems	-	-	-	-	-	-	-	-	-	-	-	-	S ₃	-	-	-	-	-
abdA	-	-	-	-	-	-	-	-	-	-	-	Q ₆ Y ₈	-	-	-	-	-	-
AbdB	-	-	-	-	-	-	-	-	-	-	-	K ₆ Y ₈	K ₃	-	-	-	-	-
cut	I ₃₁	-	-	-	-	H₅₀	-	-	N₄₇	-	-	V ₆ -	K ₃	-	-	-	R ₅₅	-
<u>Altered Specificity</u>																		
Bicoid	L ₃₁	-	-	-	-	K₅₀	-	-	V ₄₇	-	-	-	-	-	-	-	R ₅₅	-
Mat a2	L ₃₁	-	-	-	-	S₅₀	-	-	N₄₇	-	-	H ₆	-	P ₃	-	Q₅	-	-
Zmbox 1a	K ₃₁	-	-	-	R ₅₇ -	E₅₀	-	-	K₄₇	-	T₅₁	G ₆ -	T ₃	-	K₅	S ₅₅	-	-

Fig. 5.

Alignment of the binding site with the contact residues of Engrailed. Based on the structural data of Kissinger et al. (1990), we have indicated the presumed alignment of the contact residues of the Engrailed homeodomain with the Engrailed binding site. The base pairs are numbered according to Wolberger et al. (1991). The upper line is in the 5' to 3' direction and the bases of the complementary (lower) strand are designated as prime ('). The third through fifth lines, represent the amino acids that contact the DNA via back bone (PO₄), and the bases in the major groove (M) and in the minor groove (m). Below this, the conservation of residues among the homeodomains isolated in our ligand screen are indicated with a dash (-) and residues that differ from Engrailed are indicated. The last three lines show the residues of homeodomains that have a diverged sequence specificity. The changes that might be expected to have a major influence on binding specificity are indicated in bold (see text).

Table I
Relative binding of oligonucleotides to protein products produced by clones

Clone	LP ^a	NP	NPCC	NP11	RP	TAA	PRD	BCD	OCT	Type ^b
bk28	1	1.1	0.57	0.43	0.02	0.71	0.06	0.09	–	Zn finger
bk50	1	1.8	0.94	0.41	0.03	0.02	–	–	–	HD paired
bk36	1	0.89	0.32	0.18	0.07	0.10	0.14	0.02	0.009	HD paired
zfh-1 _{bk96}	1	1.4	0.45	0.17	0.08	0.07	–	–	–	HD
bk24	1	0.78	0.71	0.19	0.10	0.05	0.03	0.004	–	HD paired
zfh-2 I & II _{bk96}	1	1.0	1.2	0.37	0.12	0.23	0.01	–	–	HD
bk64	1	1.1	0.68	0.17	0.12	0.3	0.10	0.01	–	HD LIM
bk27	1	1.6	0.75	0.26	0.16	0.002	0.32	–	–	HD paired
bk87	1	1.2	0.66	0.41	0.19	0.27	0.06	–	–	HD LIM
labial _{bk79}	1	1.1	0.65	0.15	0.19	0.06	0.06	–	–	HD labial
bk60	1	2.3	1.2	0.15	0.36	0.07	0.04	0.07	–	unknown
Abd-B _{bk70}	1	2.2	0.96	0.34	0.39	0.46	0.33	–	–	HD Abd B
cut _{bk13}	1	2.1	0.80	0.12	0.42	0.02	0.05	–	–	HD
abd-A _{bk102c}	1	1.1	1.1	–	0.50	0.2	0.7	–	–	HD Antp
ems _{bk95}	1	1.5	0.41	0.32	0.54	0.05	0.2	–	–	HD
Ctfa _{bk54}	1	0.68	0.76	0.55	0.62	0.13	0.02	–	1.9	HD POU
zfh-2 III _{bk69}	1	3.2	2.4	0.37	0.87	0.05	–	–	0.006	HD
BarH1 _{bk5}	1	1.9	0.68	0.34	2.6	0.01	0.09	–	–	HD
Pdm-1 _{bk112}	1	1.4	2.0	0.27	3.9	0.29	–	0.005	11.8	HDPOU
bk35	1	3.6	0.13	0.03	5.3	–	–	–	–	Zn finger
ftz	1	1.4	0.89	0.21	0.44	0.24	0.09	–	–	–
en	1	0.79	0.33	0.08	–	0.04	0.12	–	–	–

All values must be × 100.

^aThe binding to all sites was normalized to that of LP binding, which was set at 100. A dash in a column indicates that the radioactivity bound to the filter was not above background.

^bType indicates the type of DNA binding domain or the class of homeodomain (HD) found in each clone.

^cThe absolute number of counts bound to the abd/Abk102 clone were significantly lower than for the other clones.

Table II

Summary of the sequence relationships of oligonucleotides used in binding experiments, and the binding activity exhibited by the majority class of clones

Sequence	Binding		Matches to	
			ATTA	Flanking
Consensus ^a		- c/tAATTAA/g		
LP ^b	++++	<u>TCAATTAAT</u> TGA ^e	4	3
NP	++++	<u>TCAATTAA</u> aTGA	4	3
NPCC	++++	<u>TCAATTAA</u> aTcc	4	3
NP11	+++	<u>TCAATTAA</u> aTG -	4	3
RP ^b	++	<u>TCA</u> tTTAAaTGA	3	3
TAA ^b	++	- ttATTAtta	4	1
PRD ^c	+	gattTgAcgTaA * gtAATTAcgTGA	4	2
BCD	+ / -	gggATTAgA	4	1
OCT ^d	-	atgcaaAtgCAAaTAta	3	2

^aThe bases common to the binding sites of many different homeodomains are indicated in bold. The flanking positions are highly conserved among Engrailed binding sites. Somewhat less conserved positions bracket those indicated here, and these may have a small influence on binding (Desplan et al., 1988).

^bSome of the oligonucleotides contain more than one copy of the indicated match to the consensus. LP and RP are symmetric sequences having two palindromic copies of the homology. The TAA oligonucleotide is a simple repeat sequence that has permuted copies of the region of homology. The repetition of the sequence is likely to influence the binding.

^cThe PRD sequence has essentially no homology to the consensus, but self ligation of the oligonucleotide produces a junction sequence (second line) having substantial homology (product of self ligation at the end marked with an asterisk, 1/4 of all junctions).

^dWhen aligned at end of the oligonucleotide, the OCT sequence has no homology to the consensus, but all junction fragments can be aligned with modest homology (second line).

^eThe underlining shows matches to the consensus, and upper case shows matches with the selecting oligonucleotide (LP).

Table III
Influence of the POU specific domain (PSD) on DNA binding specificity

Clone ^a	LP ^b	NP	NPCC	NP11	RP	TAA	PRD	BCD	OCT	Extent of clone ^c	
										PSD	HD
Cfla _{pk} 54	100	68	76	55	62	13	2	–	190		
Cfla _{pk} 110	100	83	140	34	88	62	10	–	258		
Cfla _{pk} 5	100	94	99	51	130	63	4	<1	310		
Cfla _{pk} 99	100	69	26	17	15	29	3	–	–		
Cfla _{pk} 76	100	51	24	11	10	21	2	–	–		
Cfla _{pk} 74	100	78	69	15	19	28	9	–	–		
ftz	100	140	89	21	44	24	9	–	–		
en	100	79	33	8	–	4	12	–	–		

^a Cfla λgt11 phage clones isolated in this screen.

^b The LP site was used as the site to which the binding of different oligonucleotides were normalized. A dash in a column indicates that counts bound to the filter by the test site were not detectable above background.

^c Diagrammatic representation of the phage clones. The POU specific domain (PSD) is indicated by the hatched box and POU (HD) homeodomain by the open box. The position of the fusion between the *lacZ* amino acid sequence and the *Drosophila* amino acid sequence for clones Cfla_{pk}54 and Cfla_{pk}110 is not known and is indicated by a break. Not drawn to scale.