

Differential DNA sequence recognition is a determinant of specificity in homeotic gene action

Stephen C. Ekker, Doris P. von Kessler and Philip A. Beachy

Howard Hughes Medical Institute, Department of Molecular Biology and Genetics, The Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

Communicated by M. Bienz

The homeotic genes of *Drosophila* encode transcriptional regulatory proteins that specify distinct segment identities. Previous studies have implicated the homeodomain as a major determinant of biological specificity within these proteins, but have not established the physical basis of this specificity. We show here that the homeodomains encoded by the *Ultrabithorax* and *Deformed* homeotic genes bind optimally to distinct DNA sequences and have mapped the determinants responsible for differential recognition. We further show that relative transactivation by these two proteins in a simple *in vivo* system can differ by nearly two orders of magnitude. Such differences in DNA sequence recognition and target activation provide a biochemical basis for at least part of the biological specificity of homeotic gene action.

Key words: *Deformed* gene/development/DNA recognition/homeodomain/*Ultrabithorax*

Introduction

The homeotic genes of *Drosophila* are clustered on the third chromosome in a linear order corresponding to that of the segments whose identities they specify. Transcriptional regulation by homeotic gene products depends upon sequence-specific DNA recognition by the homeodomain, a conserved 61 amino acid sequence present within the proteins encoded by each gene (reviewed by Hayashi and Scott, 1990). Homeodomain genes have also been found in a broad spectrum of other animal groups, from simple invertebrates to mammals (Graham *et al.*, 1989; Kenyon and Wang, 1991; reviewed in McGinnis and Krumlauf, 1992). The degree of sequence conservation across species is striking: the homeodomains encoded by the *Drosophila* homeotic gene *Deformed* (*Dfd*) and its human counterpart *Hox-4.2*, for example, are identical at 55 of 61 amino acid residues despite 600 million years of evolutionary divergence (Regulski *et al.*, 1987; Graham *et al.*, 1989).

As in *Drosophila*, some homeodomain genes in other species are also organized in chromosomal clusters. Individual homeodomains within these clusters typically display greater similarity between species than do adjacent homeodomains within a species. The functional relationship underlying this pattern of evolutionary conservation has been demonstrated in *Drosophila* embryos, where the re-programming of segment identity by ectopic expression of *Drosophila* homeotic genes can be closely mimicked by

ectopic expression of the mammalian homeodomain gene counterparts (Malicki *et al.*, 1990; McGinnis *et al.*, 1990). Other experiments with chimeric *Drosophila* proteins have shown that target specificity is determined primarily by the identity of the homeodomain (Kuziora and McGinnis, 1989, 1991; Gibson *et al.*, 1990; Mann and Hogness, 1990).

Biochemical studies have established clear differences in DNA sequence recognition for divergent homeodomains such as those encoded by *bicoid*, *caudal* and *TTF-1* (Dearolf *et al.*, 1989; Driever and Nüsslein-Volhard, 1989; Treisman *et al.*, 1989; Guazzi *et al.*, 1990). In contrast, studies of homeotic and other closely related homeodomain proteins have focused qualitatively upon the ability of these proteins to recognize each other's binding sites promiscuously, to the extent that no DNA sequence could be assigned as a unique binding site for a particular protein (Desplan *et al.*, 1988; Hoey and Levine, 1988; reviewed in Hayashi and Scott, 1990). Most of the sites studied contained a TAAT motif, and the binding results probably reflect the importance of this motif for DNA sequence recognition by a particular class of homeotic and related homeodomain proteins (see, for example, Ekker *et al.*, 1991 and Florence *et al.*, 1991). These studies, while suggesting that homeodomains encoded by homeotic genes share related DNA sequence specificities, do not present the type of systematic and quantitative comparison of DNA sequence preferences needed to determine the role of differential DNA sequence recognition in the biological specificity of homeotic gene products.

We present here such a comparison of the DNA sequence recognition properties of proteins encoded by the *Drosophila* homeotic genes *Ultrabithorax* (*Ubx*) and *Dfd*. Using purified homeodomain peptides in oligonucleotide selection experiments, we have identified distinct 9 bp consensus DNA binding sites for *Ubx* (5'-T-T-A-A-T-G>T-G>A-C-C-3') and *Dfd* (5'-T/C-T-A-A-T-G>T-A>G-A-C-3'); these sites share a central core of sequence 5'-T-A-A-T-G>T-3', but differ to either side. Base preference indices derived from the selection experiments correctly predicted the order of base preference at these positions when sequence variants were tested for binding *in vitro*; these sequence preferences also correlated well with the levels of transactivation measured by β -galactosidase reporter gene expression assays in yeast.

The relative activation of reporter gene targets by intact *Ubx* and *Dfd* proteins in this system can differ by nearly two orders of magnitude. The major determinants of this specificity are the differential DNA sequence preferences of the *Ubx* and *Dfd* homeodomains, as suggested by the following results: first, the full-length *Ubx* protein showed an *in vitro* sequence preference very similar to that of the *Ubx* homeodomain peptide and secondly, a chimeric *Dfd* protein modified to contain a *Ubx* homeodomain showed a *Ubx*-like specificity on defined target sites in yeast. Using sequence selection experiments with chimeric homeodomains, we have mapped the determinants responsible for

differential sequence recognition within the *Ubx* and *Dfd* homeodomains to two distinct DNA-contacting regions: one of these encompasses the amino-terminal arm and the other comprises the carboxy-terminus, including a part of the third or 'recognition' helix. These results corroborate ectopic expression studies of chimeric *Dfd/Ubx* proteins in embryos (Lin and McGinnis, 1992), which identify functional roles for the amino-terminal and carboxy-terminal regions of these homeodomains in determining target specificity. All these results suggest that differential DNA sequence preferences of individual homeodomains are a major determinant of the biological specificity of homeotic gene products. Preservation of these functional differences in DNA sequence recognition may account for some of the remarkable conservation of individual homeodomain sequences among vertebrate and invertebrate animal groups.

Results

Identification of a TAAT core sequence element recognized by the *Dfd* homeodomain

In order to identify optimal DNA binding sites, we used purified *Ubx* and *Dfd* homeodomain peptides (UbxHD and DfdHD) to select high-affinity sites from a population of oligonucleotides containing stretches of random sequence. To map determinants responsible for differences in sequence preference, we also carried out selections with chimeric homeodomain peptides (Figure 1A). The preparation of UbxHD has been described previously (Ekker *et al.*, 1991). Like UbxHD, DfdHD includes 10 amino acid residues beyond the carboxy-terminus of the 61 residue canonical homeodomain. This extension was designed to include residues carboxy-terminal to the homeodomain that are conserved in the *Dfd* homologues of several other species (Regulski *et al.*, 1987); a similar extension, although of different sequence, is conserved among *Ubx* homologues (Wysocka-Diller *et al.*, 1989). DfdHD and the four chimeric homeodomains were expressed in and purified from *Escherichia coli* in a similar fashion to UbxHD (see Materials and methods); samples of the purified peptides are shown in Figure 1B. Initial selection experiments with DfdHD were essentially identical to earlier experiments with UbxHD, which utilized an immobilized homeodomain peptide matrix and an oligonucleotide with a 12 bp stretch of random sequence. From the oligonucleotides selected with the DfdHD matrix, 21 of the 31 analyzed (68%) contained a 5'-TAAT-3' core sequence (data not shown). This proportion is comparable to that observed in earlier experiments with *Ubx* (57 of 88 or 65% contained a 5'-TAAT-3' core, and no other sequence elements were present at frequencies this high; Ekker *et al.*, 1991). These experiments are in good agreement with the work of Regulski *et al.* (1991), in which all genomic binding sites identified for intact *Dfd* protein contained a TAAT sequence element.

Differences in DNA sequence preference between the *Ubx* and *Dfd* homeodomains

The TAAT element preferred by both homeodomains was used to facilitate the characterization of differences in DNA sequence recognition. The approach we took, which resembles that of Blackwell and Weintraub (1990), used a 64 base oligonucleotide containing seven bases of random sequence to each side of a 5'-T-A-A-T-3' sequence core (see Materials and methods for details). This oligonucleotide also

contained flanking end sequences usable for priming in the polymerase chain reaction (PCR), sequencing and generating double-stranded DNA. The presence of the TAAT sequence fixed the positions of binding sites within individual oligonucleotides and permitted simultaneous sequence determination of the entire pool of selected oligonucleotides. Specific binding during each round of selection was favored by allowing complexes between homeodomain peptide and the ³²P-labelled 64 bp oligonucleotide to decay in the presence of excess specific, unlabelled competitor DNA. Specific protein-DNA complexes were then separated from unbound DNA by polyacrylamide gel electrophoresis, and the DNA in these complexes was amplified after each round of enrichment. Since the specific competitor in these selections lacked homology to the primer, the competitor did not interfere with subsequent amplification or sequence determination steps.

Three rounds of selection were performed with the TAAT core oligonucleotide for UbxHD, DfdHD and each of the four chimeras. Increasing enrichment after each round (data not shown) was evident from increases in complex stability and from dideoxy nucleotide chain termination sequence analysis using a ³²P-labelled primer (see Materials and methods). Sequence analysis results after three rounds of enrichment are shown in Figure 2A, using a primer of the polarity that displays the ATTA complement of the core (similar results were obtained using a primer of the opposite polarity). With the TAAT core defined as positions 2-5, specific base preferences are apparent at positions 1, 6, 7, 8 and 9 for UbxHD (second set from left in Figure 2A) as well as DfdHD (right-most set in Figure 2A). Quantitative analysis allowed the ordering of base preferences at positions flanking the core. For this analysis we used a storage phosphor screen (Molecular Dynamics) to acquire a digital image of the gel shown in Figure 2A. Peaks corresponding to bases at positions flanking the TAAT core were integrated and peak values were normalized to remove variation between lanes due to differences in the amount of labelled DNA loaded. Normalized peak values within each of the four lanes for a particular protein were then compared with the values in the corresponding lane for the unselected oligonucleotide. The ratios of selected to unselected peak values were then used to construct a preference index at each position for each protein, these indices are shown in Figure 2B. Each index is scaled so that the sum of the index for all four bases at a particular position is always equal to four; the value for a particular base is >1 if its presence is favorable for selection and <1 if its presence is unfavorable.

The base preference indices from Figure 2B are represented graphically in Figure 3A. To facilitate comparison with the chimeric homeodomains (see below), Figure 3A is arranged with results for DfdHD at the upper left and results for UbxHD at the lower right. In accordance with our previous convention (Ekker *et al.*, 1991), the information from Figure 2 was converted to its complement so that the core appears as 5'-TAAT-3' in Figure 3A. The index values were transformed by subtracting 1 so that bars extending above and below zero denote bases selected for and against, respectively. Clear differences between DfdHD and UbxHD are evident at positions 1, 7 and 8. Note how DfdHD (upper left of Figure 3A) shows a nearly equal preference for a T or a C at position 1, a clear preference for an A at position

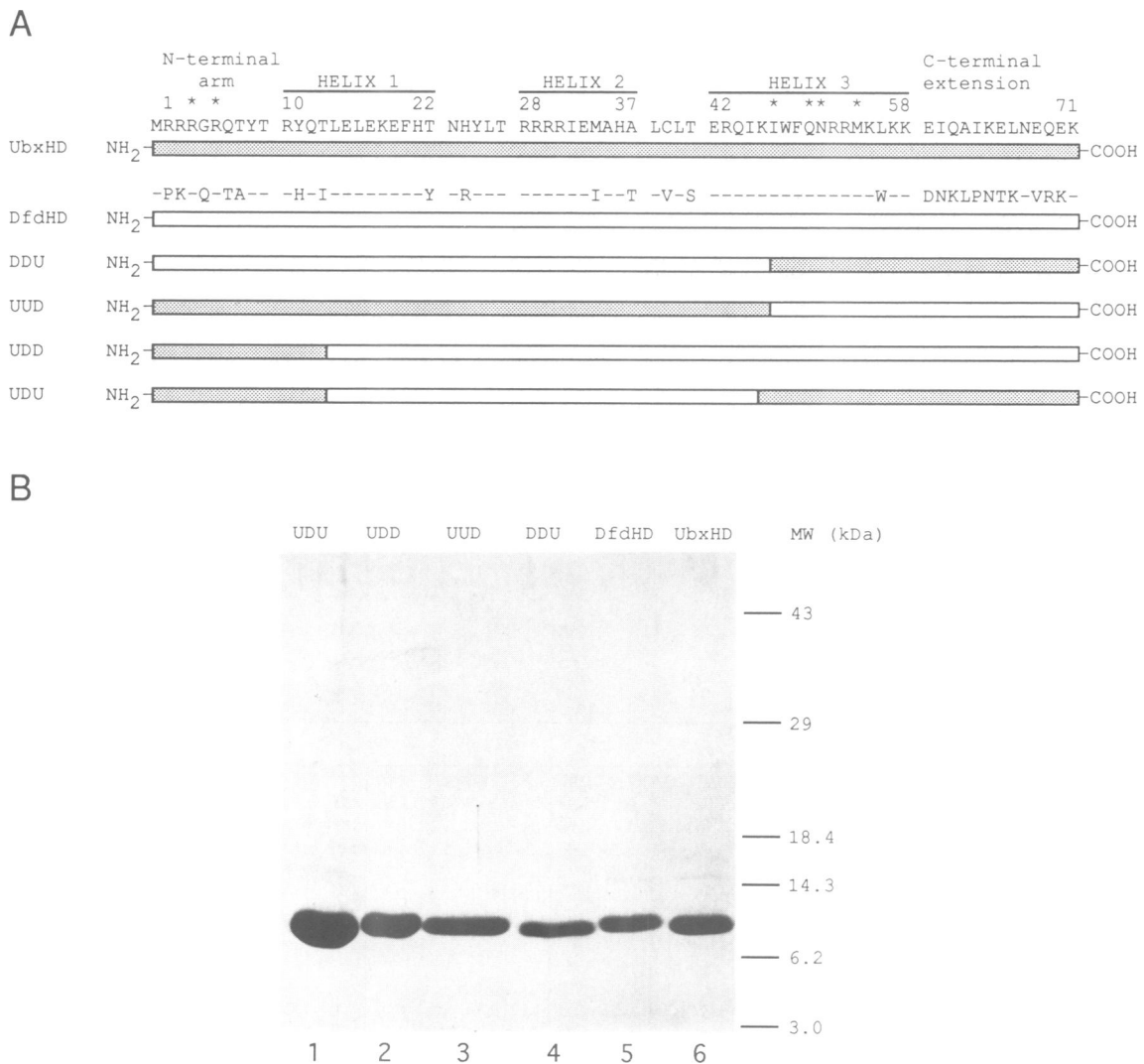


Fig. 1. Structure and purification of homeodomain peptides. **(A)** Structure of Ub_xHD, DfdHD and four chimeric homeodomains. The amino acid sequences of the Ub_xHD and DfdHD peptides are shown above the shaded and open bars, respectively, with identity indicated by a hyphen; the origin of the amino acid residues for each of the four chimeras is shown schematically. The numbering scheme and position of α -helices corresponds to that of the *engrailed* homeodomain (Kissinger *et al.*, 1990). An asterisk denotes a predicted sequence-specific contact residue. **(B)** Purified homeodomain protein samples. Coomassie-stained samples of each homeodomain protein are displayed after PAGE (15% polyacrylamide; 22, 17, 3.6, 5.7, 3 and 3 μ g protein, respectively, for lanes 1–6). The relative molecular masses and mobilities of markers are indicated at the right.

7, and a very weak A at position 8; Ub_xHD, in contrast, specifically prefers a T at position 1, a G or an A at position 7 and a C at position 8 (lower right of Figure 3A). From such analyses the consensus sequence preferences are 5'-T/C-T-A-A-T-G>T-A>G-A-C-3' for DfdHD and 5'-T-T-A-A-T-G>T-G>A-C-C-3' for Ub_xHD. These results confirm the previously reported sequence preference for Ub_xHD (5'-T-T-A-A-T-G>T-G>A-3'; Ekker *et al.*, 1991); in addition, the current method represents an improvement in sensitivity since the Ub_xHD consensus sequence is extended by 2 bp at the 3' end.

Binding of Ub_xHD and DfdHD in vitro conforms to predictions of selection experiments

Dissociation rate constants were measured for Ub_xHD and DfdHD on a variety of single binding site DNA oligonucleotides to quantify the differences observed in the base preferences of these proteins. We measured dissociation rate

constants because under appropriately selected conditions these measurements are insensitive to variations in concentration of protein, or DNA, or in protein activity. Furthermore, we had previously shown for Ub_xHD that dissociation rates of complexes with a given set of DNA sites parallel the magnitudes of corresponding equilibrium binding coefficients (Ekker *et al.*, 1991), indicating that differential sequence specificity is determined primarily by differences in stability of homeodomain–DNA complexes.

Dissociation rate measurements for complexes of Ub_xHD and DfdHD with a series of binding sites are shown in Table I. Individual sites are grouped according to sequence relationships with other binding sites; position(s) that vary within a group are underlined. The longest measured half-life for a Ub_xHD complex was with the sequence 5'-T-T-A-A-T-G-G-C-C-3' (sequence a), in agreement with predictions of the selection experiments; consequently this appears to be the optimal 9 bp binding site for *Ubx*. These results further show that the base preference indices are good

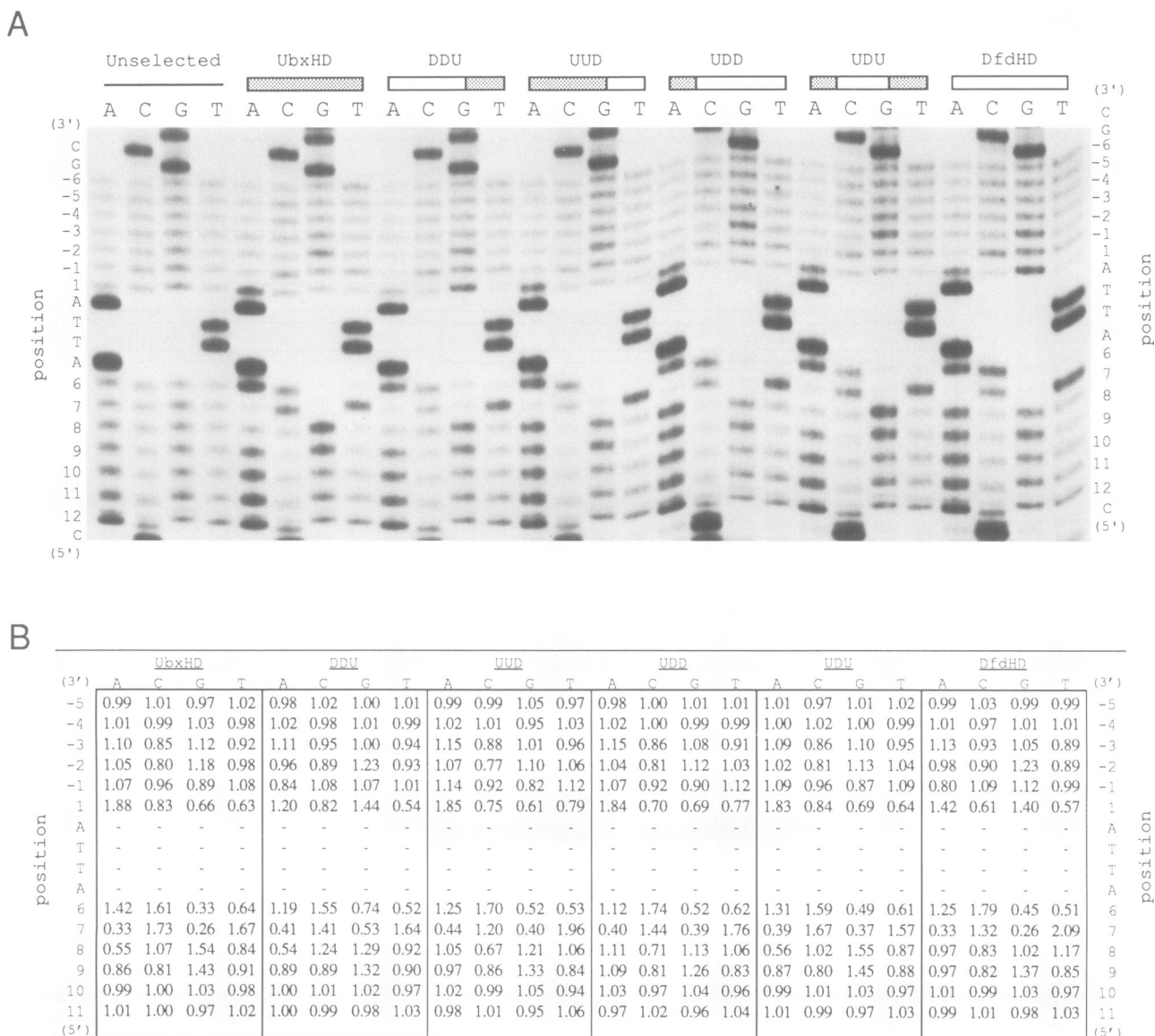


Fig. 2. Sequence analysis and preference indices for six homeodomain peptides. (A) Sequence analysis after three rounds of enrichment for each protein and for unselected DNA. The position of the 5'-ATTA-3' core is indicated, with numbers identifying each position within the random sequence portions of the oligonucleotide. (B) Base preference indices. Indices are shown for each protein at each position (see text and Materials and methods for details). An index value of 1.0 indicates no preference relative to the unselected control.

predictors of the effect of single base substitutions upon binding. For example, relative complex stabilities of UbxHD with sequences in group II are ordered as predicted by the sequence preference indices at position 6 (G > T > A > C); position 7 preference indices for UbxHD are similarly confirmed by sequences in group III.

Dissociation rate constant measurements with DfdHD were similarly consistent with base preference indices. For example, sequences in group III match the order predicted by the preference indices and confirm the differences between DfdHD (A > G) and UbxHD (G > A) at position 7. At position 6 (group IV), the bases with a positive preference index (G and T) yield more stable complexes with DfdHD than those with negative indices (C and A); the detailed order of these complex stabilities is not as expected, however, for reasons we do not yet understand (see below for a discussion of context effects at positions 6 and 7). Group V sequences support the predicted preferences by

DfdHD for A and C at positions 8 and 9, and group VI sequences show that base changes at the extreme ends of the 9 bp site produce the predicted effects for both DfdHD and UbxHD.

One particularly interesting class of exceptions to the selection experiment predictions shows that identity of a base at one position can influence the order of base preference at a neighboring position. This is most clearly seen for UbxHD: when position 7 is a G, the order of preference at position 6 is G > T > A > C (group II), as predicted from the preference indices. When position 7 is the non-optimal base A, however, the order of preference at 6 shifts to G = T > C > A (group IV). Non-independence of binding effects by substitutions at positions 3' of the TAAT core has been noted previously (Percival-Smith *et al.*, 1990) and is probably indicative of multiple modes of interaction with DNA by the amino acid residue at position 50 of the homeodomain (see Discussion).

Table I. Dissociation rates of UbxHD and DfdHD complexes with various DNA sequences

		Position									UbxHD			DfdHD		
		1	2	3	4	5	6	7	8	9	$k_d \times 100$ (min^{-1})	$t_{1/2}$ (min)	$t_{1/2}$ (rel)	$k_d \times 100$ (min^{-1})	$t_{1/2}$ (min)	$t_{1/2}$ (rel)
I	a	T	T	A	A	T	G	G	C	C	-0.89 ± 0.02	78	1.0	-2.4 ± 0.5	29	0.4
II	b	T	T	A	A	T	<u>G</u>	G	C	T	-1.0 ± 0.2	69	0.9	-1.29 ± 0.04	54	0.7
	c	T	T	A	A	T	<u>T</u>	G	C	T	-1.6 ± 0.1	43	0.6	-1.9 ± 0.1	36	0.5
	d	T	T	A	A	T	<u>A</u>	G	C	T	-2.6 ± 0.2	27	0.34	-3.3 ± 0.2	21	0.28
	e	T	T	A	A	T	<u>C</u>	G	C	T	-5.4 ± 0.2	13	0.16	-5.0 ± 0.4	14	0.18
III	b	T	T	A	A	T	G	<u>G</u>	C	T	-1.0 ± 0.2	69	0.9	-1.29 ± 0.04	54	0.7
	f	T	T	A	A	T	G	<u>A</u>	C	T	-1.9 ± 0.3	36	0.5	-0.94 ± 0.01	74	1.0
	g	T	T	A	A	T	G	<u>C</u>	C	T	-7.7 ± 0.1	9.0	0.12	-13.9 ± 0.9	5.0	0.07
IV	f	T	T	A	A	T	<u>G</u>	A	C	T	-1.9 ± 0.3	36	0.5	-0.94 ± 0.01	74	1.0
	h	T	T	A	A	T	<u>T</u>	A	C	T	-2.0 ± 0.8	35	0.5	-0.92 ± 0.07	75	1.0
	i	T	T	A	A	T	<u>C</u>	A	C	T	-4.8 ± 0.3	14	0.18	-3.3 ± 0.3	21	0.28
	j	T	T	A	A	T	<u>A</u>	A	C	T	-7.8 ± 1.0	8.9	0.11	-5.2 ± 0.3	13	0.18
V	e	T	T	A	A	T	C	G	<u>C</u>	<u>T</u>	-5.4 ± 0.2	13	0.16	-5.0 ± 0.4	14	0.18
	k	T	T	A	A	T	C	G	<u>A</u>	<u>C</u>	-6.5 ± 1.5	11	0.14	-1.8 ± 0.1	39	0.5
VI	d	<u>T</u>	T	A	A	T	A	G	C	<u>T</u>	-2.6 ± 0.2	27	0.34	-3.3 ± 0.2	21	0.28
	l	<u>A</u>	T	A	A	T	A	G	C	<u>T</u>	-3.0 ± 0.3	23	0.30	-4.4 ± 0.1	16	0.21
	m	T	T	A	A	T	A	G	C	<u>G</u>	-3.3 ± 0.6	21	0.27	-4.7 ± 0.3	15	0.20

Dissociation rate constants (k_d) and complex half-lives ($t_{1/2}$) were determined as described in Materials and methods. The k_d values are given as an average of at least two independent determinations \pm the standard error. All sequences except m (see Materials and methods) contain identical flanking bases.

Mapping of regions of the Ubx and Dfd homeodomains responsible for differential DNA sequence recognition

There are 17 amino acid differences between the homeodomains of UbxHD and DfdHD, and eight additional differences in the carboxy-terminal extensions of these proteins (Figure 1A). To identify the residues responsible for differences in DNA sequence recognition, we have purified four chimeric homeodomain proteins. The structures of these proteins are shown schematically in Figure 1A, with a three letter designation for each indicating the source of the amino-terminus, middle portion and carboxy-terminus (U for *Ubx* and D for *Dfd*). Purification was essentially as described for DfdHD (see Materials and methods), and samples of each are shown in Figure 1B. Three rounds of sequence selection with these chimeras were performed in parallel with UbxHD and DfdHD. Figure 2 panels A and B show the results of sequence analysis and quantification; the preference indices are presented graphically in Figure 3A. The arrangement of Figure 3A is such that proteins within a column share the same carboxy-terminus and proteins within a row share the same amino-terminus. A detailed inspection of these data serves to establish the rule that proteins with a common amino-terminus share base preferences at positions 5' to the TAAT core while proteins with a common carboxy-terminus share base preferences to the 3' side of the TAAT core. The bottom four proteins in Figure 3A (UDD, UDU, UUD and UbxHD) for example, all contain a UbxHD amino-terminus and share a clear preference for a T at position 1; in contrast, proteins DfdHD and DDU in the top row share the mixed preference for a C and T that is characteristic for DfdHD at position 1. The

greatest deviation from this rule is illustrated by the 3' preferences of protein DDU, which contains a UbxHD carboxy-terminus; even in this worst case, however, the 3' preferences were more like those of UbxHD than like those of DfdHD. The relative importance of amino- and carboxy-terminal sequences in determining the specificity of sequence recognition is best illustrated by protein UDU, in which just the middle portion of UbxHD is replaced by that of DfdHD. This protein gives sequence preferences nearly superimposable upon those of UbxHD, indicating that differences in the middle portion have little or no effect upon DNA sequence recognition.

The full-length UBX Ib protein exhibits the same fundamental sequence recognition properties as UbxHD

An assumption implicit in much previous biochemical work with homeodomain proteins is that the DNA sequence recognition properties of homeodomains accurately reflect those of the intact proteins from which they derive. To test this assumption explicitly, we performed sequence selection experiments in parallel with UbxHD and with purified, full-length UBX Ib (Beachy *et al.*, 1988). The UBX Ib protein contains each of the three internal sequences (9, 17 and 17 amino acid residues) that are variably present in *Ubx* protein isoforms due to differential splicing at a position upstream but adjacent to the homeodomain (Beachy *et al.*, 1985; O'Connor *et al.*, 1988; Kornfeld *et al.*, 1989). The biological role of these differences is not fully understood, but they are expressed in distinct embryonic tissues (Lopez and Hogness, 1991) and appear to have tissue-specific biological functions (Mann and Hogness, 1990).

After three rounds of enrichment using the TAAT core random oligonucleotide (see Materials and methods for details), Figure 3B shows that both UBX Ib and UbxHD give the consensus sequence of 5'-T-T-A-A-T-G>T-G>A-C-C-3'. Due to minor technical improvements in the selection procedures used to obtain the data in Figure 3A, less overall enrichment was observed for the enrichments shown in Figure 3B; nevertheless, the order of base preferences at each position is nearly identical for the UBX Ib determination and both UbxHD determinations. We conclude that the fundamental DNA sequence recognition properties of the *Ubx* family of proteins are determined by sequences present in the *Ubx* homeodomain peptide we have studied. This conclusion can probably be generalized to include proteins encoded by other homeotic genes, though perhaps not to proteins containing highly divergent homeodomains and/or accessory binding domains such as those of the POU (reviewed by Rosenfeld, 1991) and *paired* (Treisman *et al.*, 1991) homeodomain groups.

A second conclusion from these studies is that the biological specificity of *Ubx* isoforms probably does not derive from fundamental changes in DNA sequence recognition properties, since these variable sequences occur at positions amino-terminal to the homeodomain. More probably, the biological properties of these proteins results from protein-protein interactions, either homomeric or with other factors. We recognize that a secondary consequence of these interactions might be to modify the sequence preferences of *Ubx* protein, by analogy to the effects of MCM1 and *a1* upon binding of $\alpha 2$ homeodomain protein in yeast (Smith and Johnson, 1992); these altered preferences would depend on the presence of the interacting factors and thus would not have been detected in our experiments.

Target site activation in yeast correlates with strength of homeodomain binding *in vitro*

We tested the significance of differences in DNA sequence recognition *in vivo* using β -galactosidase as a reporter gene to monitor transactivation in a simple eukaryote, the yeast *Saccharomyces cerevisiae*. The yeast system was chosen for rapidity in testing multiple sequences and to eliminate the potential for interfering activities from homeodomain proteins present in homologous systems such as *Drosophila* embryos or cultured cells. In addition, previous work had demonstrated the feasibility of using yeast for measurement of transactivation of various target sites with several *Drosophila* homeodomain proteins (Fitzpatrick and Ingles, 1989; Samson *et al.*, 1989; Hanes and Brent, 1991). In contrast to these previous systems, however, our use of centromere plasmids and a GAL1 promoter in the regulator plasmid allowed us to control the copy number of the target and regulator plasmids and permitted inducible expression of the regulatory proteins (see Figure 4 for details).

Each of 12 target sequences was tested for activation by three proteins: UBX Ib, DFD and a DFD/UBXHD chimera in which the *Dfd* homeodomain plus five carboxy-terminal amino acids were replaced by the corresponding sequences from *Ubx* (Kuziora and McGinnis, 1989); in addition, the unmodified regulator expression plasmid was used as a control (Table II). Targets a-m contained four tandem copies of an individual site from Table I spaced according to the pattern of four individual sites present in the naturally occurring *Ubx* binding site U-A (Beachy *et al.*, 1988). Each

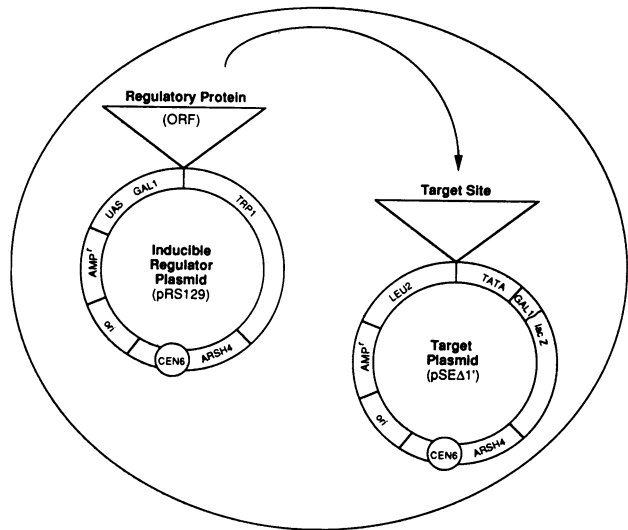


Fig. 4. Yeast transcriptional activation assay system. The plasmid pRS129 (Sikorski and Hieter, 1989) contains the *GAL1* promoter for inducible expression of homeodomain regulatory proteins. Efficiency of transcriptional activation is measured using pSE Δ 1', in which target sequences inserted near a basal promoter can activate expression of a *GAL1-lacZ* fusion protein.

of the three regulatory proteins functions as a sequence-specific transcriptional activator in *S. cerevisiae*, as shown by activation of specific targets a-m, while the control LexA-operator (target n) is activated only at extremely low levels. This LexA-operator target can be activated by a *Ubx* fusion protein containing the DNA binding domain of LexA (data not shown), as also reported by Samson *et al.* (1989). Activation by these proteins ranged across approximately two orders of magnitude, with *Dfd*-based proteins in general displaying a higher degree of activation. This may be due to differences in activation domains: *Dfd* has identifiable acidic and Gln-rich regions (Regulski *et al.*, 1987; see Ptashne, 1988 for review of transcriptional activation domains) while *Ubx*, although clearly capable of transcriptional activation (Samson *et al.*, 1989; Gavis and Hogness, 1991), does not contain these types of activation sequences.

We expected the pattern of target activation by a particular protein to reflect the *in vitro* binding properties of its homeodomain. To determine whether there was indeed such a correlation, we applied a non-parametric statistical test; as described below, this test verified the existence of an excellent correlation between strength of yeast target activation and *in vitro* binding of the cognate homeodomain. Certain individual sequences nevertheless displayed anomalous behaviour, for which we believe endogenous yeast factors are at least partially responsible. For example, target f is specifically activated in the absence of the three regulatory proteins, indicating the existence of an endogenous yeast factor capable of activating this sequence; sequence-specific endogenous repressing activities would not have been identified in our experiments but could be interfering with full induction of other targets. All of the results have nevertheless been included in our statistical evaluation of correlation.

To determine the statistical significance of correlation between *in vitro* binding and *in vivo* activation, we calculated the Spearman rank-order correlation coefficient (r_s). This requires, first, assignment of rank order (i.e. from 1 to 11)

Table II. Sequence specific target activation (expressed in units of β -galactosidase activity)

Target site	Nucleotide at position									Regulatory protein			
	1	2	3	4	5	6	7	8	9	UBX Ib	DFD	DFD/UBXHD	Control
a	T	T	A	A	T	G	G	C	C	55 ± 6	99 ± 22	115 ± 2	0.0 ± 0.0
b	T	T	A	A	T	G	G	C	T	71 ± 6	> 41 ± 4	< 107 ± 19	0.1 ± 0.1
c	T	T	A	A	T	T	G	C	T	93 ± 9	77 ± 9	156 ± 7	0.2 ± 0.1
d	T	T	A	A	T	A	G	C	T	18 ± 3	10 ± 2	15 ± 1	0.1 ± 0.0
e	T	T	A	A	T	C	G	C	T	6.6 ± 1.5	9.8 ± 1.4	3.3 ± 1.0	0.4 ± 0.1
f	T	T	A	A	T	G	A	C	T	76 ± 12	157 ± 25	117 ± 48	22 ± 4
g	T	T	A	A	T	G	C	C	T	2.0 ± 0.3	1.0 ± 0.2	0.9 ± 0.3	0.2 ± 0.1
i	T	T	A	A	T	C	A	C	T	19 ± 1	23 ± 3	11 ± 3	0.3 ± 0.1
k	T	T	A	A	T	C	G	A	C	1.7 ± 0.2	< 15 ± 2	> 1.1 ± 0.2	0.2 ± 0.1
l	A	T	A	A	T	A	G	C	T	2.0 ± 0.3	0.8 ± 0.2	1.1 ± 0.3	0.1 ± 0.1
m	T	T	A	A	T	A	G	C	G	2.1 ± 0.4	4.3 ± 1	1.6 ± 0.3	0.2 ± 0.0
n	LexA-operator									0.6 ± 0.2	0.5 ± 0.2	0.3 ± 0.1	0.1 ± 0.1

Units of β -galactosidase activity were measured after 4 h of induction (see Materials and methods). The values are given as an average of at least three independent determinations \pm the standard error. The target sites a–m each contained four tandem copies of the corresponding binding site sequence from Table I: details of spacing and the LexA-operator sequence are given in Materials and methods. Targets b and k show the greatest relative difference in activation by *Ubx* and *Dfd* proteins.

for the series of sequences tested in each type of assay. The coefficient is then computed directly from the square of differences between the rankings for each sequence in the two assays being compared: r_s thus will be small in the case of a strong correlation between the two assays (the rankings would be similar and the differences between rankings therefore small) and large in the case of no correlation between rankings. The value of r_s can then be used to estimate the statistical significance of the correlation in the form of a probability (see Materials and methods). We selected this non-parametric statistical test because it allowed us to search for correlation without assuming linearity or any other explicit relationship between the data from two different assays.

Table III presents for each pairwise combination of target activation and *in vitro* binding the probability of finding such a strong correlation on a random basis. The significance of correlation between activation by each of the three proteins is at least 15-fold better with *in vitro* binding by the cognate than that by the non-cognate homeodomain. A reasonably high significance is also observed between yeast activation and *in vitro* binding by the non-cognate homeodomains; it is important to note, however, that these correlations are no stronger than the correlation between binding for the two homeodomain peptides ($P \leq 0.025$; bottom line of Table III), which must be considered the background for comparison. We thus can conclude that differences in target activation *in vivo* correlate well with differences in DNA sequence recognition that we have measured *in vitro*. The activation pattern of the DFD/UBXHD chimera, which correlates best with binding by the *Ubx* homeodomain, further shows that the homeodomain is responsible for specifying these differences.

Beyond the observation that target activation correlates well with *in vitro* binding, we note that the relative difference in activation of particular targets by *Ubx* and *Dfd* can be quite large. For example, target k was activated by DFD at a level nearly 10-fold higher than UBX Ib, while UBX Ib activated target b nearly 2-fold better than DFD, yielding a 17-fold relative difference in activation for these two sites. The DFD/UBXHD chimera may be more appropriate for

Table III. Correlation between strength of binding and transcriptional activation

	Order of binding by UbxHD	Order of binding by DfdHD	Ratio of probabilities
Order of activation by UBX Ib	$P \leq 0.0012$	$P \leq 0.024$	1:20
Order of activation by DFD	$P \leq 0.015$	$P \leq 0.0010$	15:1
Order of activation by DFD/UBXHD	$P \leq 0.0011$	$P \leq 0.018$	1:16
Order of binding by UbxHD	–	$P \leq 0.025$	

Correlation between rank order of transcriptional activation and rank order of binding was determined using the Spearman rank-order correlation coefficient (r_s ; Siegel and Castellan, 1988) to estimate the probability (P) that the correlation occurred by chance. Experimental values used for ranking derive from Table I (*in vitro* binding) and Table II (transcriptional activation, background subtracted). Note the inherent correlation in binding between UbxHD and DfdHD (bottom line; see text).

comparison with DFD because it has the specificity of UBX Ib combined with the activation potential of DFD; the relative difference in activation of targets b and k by these two proteins is 43-fold.

Integration of single site differences through cooperative binding to multiple sites

In establishing the yeast system for assay of various targets we found that multiple sites were required to produce an easily measurable response, even with the strongest binding sites (data not shown). We also noted that the differences in yeast target activation for a set of sites usually exceeded many-fold the binding differences for the same set of sites *in vitro*. For example, the average activation by *Ubx* of the

best four targets in yeast [targets a, b, c and f (Table II) average = 73] was 38-fold better than the average activation of the four poorest targets [targets g, k, l and m (Table II); average = 1.8]; in contrast, the *in vitro* binding differences between these two groups of four sites for UbxHD were only 3.5-fold (see Table I; average $t_{1/2}$ = 56.5 for sites a, b, c and f, compared with $t_{1/2}$ = 16 for sites g, k, l and m). These observations suggest the existence of a mechanism for integration of binding to the four tandemly repeated sites present in these yeast targets.

To test whether this mechanism involved cooperative interactions of *Ubx* proteins with multiple sites, we characterized the binding of full-length *Ubx* proteins to the yeast multiple site targets b and k in Table II. These targets show 47- and 119-fold activation differences for *Ubx* and DFD/UBXHD, but only a 7-fold single site binding difference *in vitro* with UbxHD (sequences b and k; Table I). Figure 5 shows a footprint challenge assay that measures complex stability for purified UBx Ib protein (Beachy *et al.*, 1988) bound to DNA fragments from the yeast target plasmids b and k. In this assay, pre-formed complexes were challenged with excess unlabelled specific competitor, and aliquots of the mixture were subjected to DNase I treatment at the indicated times following addition of competitor. Figure 5A shows that the footprinted region containing the four site cluster (the largest footprint) decays more rapidly for yeast target b than for k. To quantify these differences, the intensity of DNase I cleavages within region II were integrated and used as an indicator of binding in order to determine dissociation rates (Figure 5B). The early time points provide an estimate of dissociation rates for the fully bound complexes; as shown in Figure 5C, these rates differed by 43-fold, very similar to the 47-fold difference measured in yeast transactivation. In addition, we noted the presence of a binding site within target plasmid sequences whose stability also improved in target b (region I). The increased stability of complexes formed with a four site cluster of stronger individual sites suggests that cooperative binding by full-length protein indeed may play a role in integration of multiple site differences *in vivo*. We have further shown that cooperativity in binding can extend to distant sites, and that this cooperativity requires the presence of amino acid sequences lying outside the *Ubx* homeodomain (Beachy, P.A., Varkey, J., Young, K.E., von Kessler, D.P. and Ekker, S.C., in preparation).

Discussion

Using biochemical techniques and yeast transcriptional activation assays we have shown that the homeodomain proteins encoded by *Ubx* and *Dfd* bind optimally to distinct DNA sequences. We have established that the DNA sequence preferences of the full-length *Ubx* protein are determined by the homeodomain plus several carboxy-terminal residues. Also consistent with the sufficiency of the homeodomain for specification of DNA sequence preference, we have demonstrated that transcriptional activation of 11 distinct target sequences in yeast reflects the identity of the homeodomain present in the regulatory protein tested, regardless of context. Finally, we have mapped differential DNA sequence recognition functions to the amino-terminus of the homeodomain for bases 5' to the common central region of the binding site, and to the carboxy-terminus (plus extension) for bases 3'.

Determinants of differential sequence recognition correspond to specificity determinants in the embryo

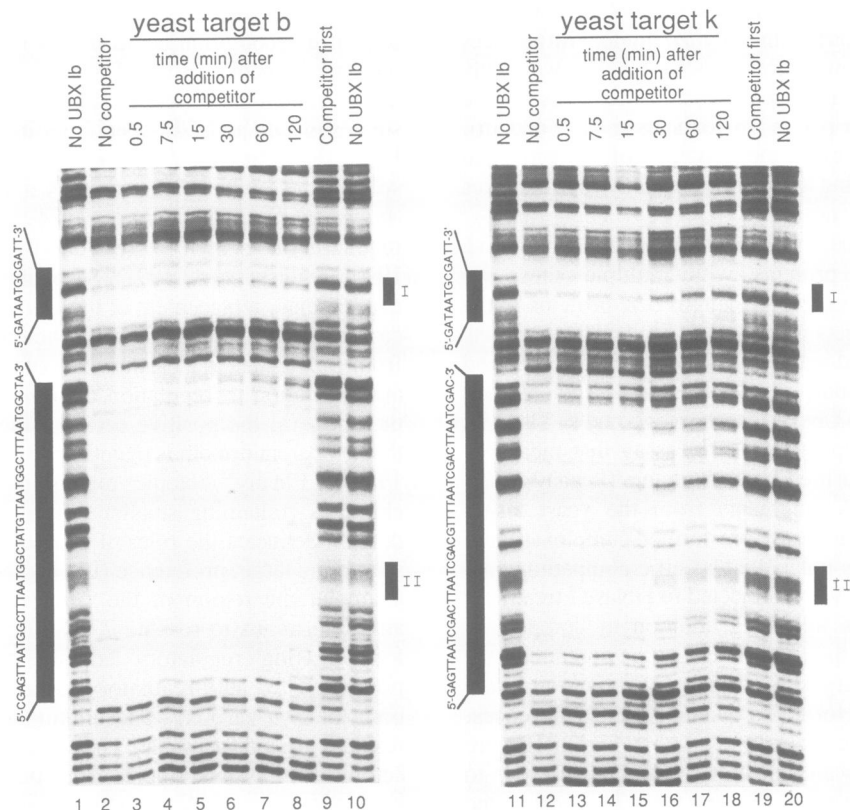
Our *in vitro* and *in vivo* results suggest that differential DNA sequence recognition could play a role in determining the biological specificity of *Dfd* and *Ubx* function. As an important complement to these results, we consider previous studies of *Dfd* and *Ubx* specificity in *Drosophila* embryos. First, the *Ubx* and *Dfd* proteins exhibit different target specificities, with *Ubx* acting as a negative regulator of *Antennapedia* (Hafen *et al.*, 1984) and *Dfd* as a positive regulator of its own expression (Bergson and McGinnis, 1990; Regulski *et al.*, 1991). Secondly, in ectopic expression experiments, replacement of the *Dfd* homeodomain and five carboxy-terminal residues with the corresponding residues from *Ubx* results in a clean switch of target specificity, from autoregulation to regulation of *Antennapedia* [although in the chimera, the positive sense of the regulatory effect in the *Dfd* parent protein is maintained (Kuziora and McGinnis, 1989)]. Finally, ectopic expression studies with other chimeras containing smaller portions of the *Ubx* homeodomain delineate the roles of homeodomain sub-regions in specifying target preference (Lin and McGinnis, 1992). For example, the region of the *Ubx* homeodomain from the amino-terminus to residue 7 is sufficient in a *Dfd* context for targeting regulatory activity to the *Antennapedia* promoter; for auto-regulatory targeting of *Dfd*, however, both the *Dfd* carboxy- and amino-terminal homeodomain residues are indispensable. The close parallels between these ectopic expression studies and our characterization of homeodomain DNA sequence preferences suggest that differential DNA sequence recognition provides a mechanistic basis for the biological specificity of homeotic gene action.

Elements common among TAAT-preferring homeodomain proteins

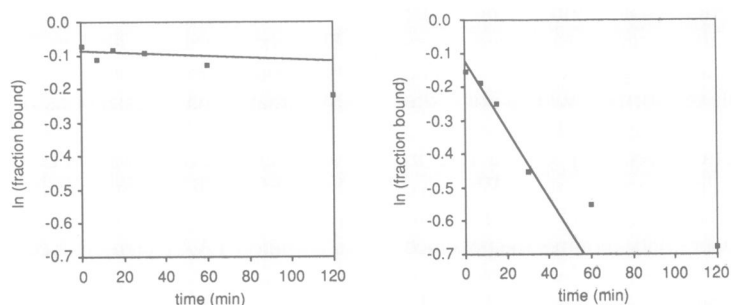
Systematic characterizations of DNA sequence preference are now available for *Dfd* and *Ubx* (Ekker *et al.*, 1991; this work) and for the *Drosophila* homeodomain protein encoded by *fushi tarazu* (Florence *et al.*, 1991). These closely related proteins bind preferentially to DNA sequences containing a common TAAT element (see Figure 6); it should therefore be possible to identify common homeodomain features that may be involved in TAAT core recognition. Because three-dimensional structure determinations have not been completed for these proteins, we rely upon analogies drawn from the structures of DNA-protein complexes that have been determined for the *Drosophila* homeodomains encoded by *engrailed* (Kissinger *et al.*, 1990) and *Antennapedia* (Otting *et al.*, 1990), and for the homeodomain encoded by the yeast $\alpha 2$ gene (Wolberger *et al.*, 1991). Such structural analogies are justified by the observation that widely diverged homeodomains can be oriented and aligned to their DNA binding site sequences by reference to an invariant interaction between residue N51 and the adenine partner of an A:T base pair (Wolberger *et al.*, 1991). A large body of evidence, some of it discussed below, indicates that the third base of the TAAT core (position 4 in Figure 6A) corresponds to this adenine.

Figure 6A presents a schematized model for DNA sequence recognition by the TAAT-preferring class of homeodomains. Base-specific recognition of the TAAT element probably involves residues within the amino- and

A



B



C

	target b	target k	Ratio
$k_d \times 10^4$	2.4	103	
$t_{1/2}$ (hours)	48	1.1	43x
β -gal (units)	71	1.5	47x

Fig. 5. Stability of UBX Ib protein complexes correlates with transcriptional activation in yeast. Preformed protein–DNA complexes were challenged with unlabelled, excess specific DNA competitor and treated with DNase I at the indicated times. (A) Stability of DNase I footprints on yeast targets b and k. Lanes 2 and 12 show the extent of protection by UBX Ib in the absence of added competitor; lanes 9 and 19 show the extent of protection when competitor is added to labelled DNA before the addition of protein. The large footprint covers a region including four tandem repeats of the sites b and k from Table II; the smaller footprint covers sequences present in the vector pSE Δ 1' (see Figure 4). Note the greater stability of both footprints in target b. (B) Dissociation rate determination. The area corresponding to region II in each lane of part (A) was quantified and used to estimate the fraction of DNA molecules remaining in complex. The natural log of these values were plotted as a function of time; to determine the initial dissociation rate and half-life for each of the two complexes, only values from the first 30 min were used for analysis (see Materials and methods). (C) Complex dissociation rates, half-lives and yeast transcriptional activation for targets b and k. The values for transcriptional activity with targets b and k (Table II) are presented with the background (control) activity subtracted.

carboxy-terminal regions of the homeodomain. By analogy to the *engrailed* structure, residues R3 and R5 in the amino-terminal arm could make minor groove contacts with bases at positions 3 and 2, respectively. Residues I47 and N51 within the third helix near the carboxy-terminus could provide major groove contacts with bases at positions 5 and 4, by analogy to *engrailed* (and $\alpha 2$ for residue 51). The role of residue 54 in the TAAT-preferring homeodomains is somewhat unclear since in $\alpha 2$ it is an arginine and contacts position 5 while, in *Antennapedia*, it is a methionine and may interact with position 6. Residues R3, R5, I47, N51 and M54 are conserved within all three homeodomains that have been rigorously shown to prefer TAAT. In addition to the base contacts made by these five residues, important contacts with the sugar-phosphate backbone of the DNA are made by a number of additional residues. With one exception (see below), however, these will not be discussed since they are conserved widely and appear to offer no further discriminatory criteria for identifying TAAT-preferring homeodomains. Other *Drosophila* homeodomains containing the five proposed TAAT sequence contact residues are listed in Figure 6C. Among the scores of other likely TAAT-preferring homeodomains not listed are those identified outside of *Drosophila*, but we expect that the same criteria will obtain.

The TAAT core sequence is not found in the binding sites of all homeodomain proteins. For example, in the well-studied binding site for $\alpha 2$ (Sauer *et al.*, 1988; Smith and Johnson, 1992), the sequence corresponding to the TAAT core is TTAC (Wolberger *et al.*, 1991). The $\alpha 2$ homeo-

domain is highly diverged; of the five proposed TAAT contact residues, only N51 is conserved. In addition, neither *caudal* (Dearolf *et al.*, 1989) nor *TTF1* (Guazzi *et al.*, 1990) appear to recognize TAAT core sequence elements. These proteins contain changes from R3 to K3, and from M54 to A54 (*caudal*) and to Y54 (*TTF1*). Substitutions like these are common (Scott *et al.*, 1989) and we expect that homeodomains with such substitutions contact residues might display preferences for non-TAAT core elements. Some homeotic selector genes fall into this category, including *labial* (R3 to S3; Diederich *et al.*, 1989), *proboscipedia* (I47 to V47; Cribbs *et al.*, 1992), and *Abdominal-B* (R3 to K3; Regulski *et al.*, 1985).

We wish to emphasize that homeodomains that prefer a TAAT core sequence may also bind to DNA sequences not containing a TAAT core; the strength of such interactions, however, is likely to be considerably lower, as has been shown for *Ubx* and *fushi tarazu* (Ekker *et al.*, 1991; Florence *et al.*, 1991). We also emphasize that the list of homeodomains in Figure 6C is not exclusive, since it is possible that homeodomains not containing all of these residues may nonetheless bind preferentially to TAAT-containing sequences. For example, the *engrailed* and *bicoid* homeodomains are capable of binding TAAT-containing sequences (Desplan *et al.*, 1988; Driever and Nüsslein-Volhard, 1989), and the crystal structure of the *engrailed* homeodomain has been solved in complex with TAAT and AAAT core sequences (Kissinger *et al.*, 1990). We are uncertain whether their inclusion in this group is appropriate, however, for lack of systematic biochemical data demonstrating preference for TAAT and because of divergences at residue 54 (A54 for *engrailed* and R54 for *bicoid*). One additional clarification in assigning homeodomains to the TAAT-preferring class can be illustrated by the TAAT sequence present in the $\alpha 2$ homeodomain binding site: this sequence is present on the opposite strand in a distinct alignment, and recognition of the TAAT involves a different set of contacts than those discussed above. Using the N51 alignment rule, the core sequence for $\alpha 2$ is TTAC, and we therefore do not include it in the TAAT core-preferring group.

Sequence specificity through differential recognition of bases flanking the TAAT core

The identity of bases flanking the TAAT core is a critical determinant for homeodomain recognition (Percival-Smith *et al.*, 1990; Ekker *et al.*, 1991; Florence *et al.*, 1991; Hanes and Brent, 1991). Biochemical and genetic evidence indicates that residue 50 plays a major role in determining sequence preference at positions one or two base pairs 3' of the TAAT core (Hanes and Brent, 1991; Percival-Smith *et al.*, 1990). Single and double base substitutions at these two positions showed greater than additive effects upon binding, suggesting that residue 50 may be capable of base-specific contacts at either or both positions (Percival-Smith *et al.*, 1990). Florence *et al.* (1991) proposed simultaneous interactions of Q50 with bases at positions 6 and 7, with recognition dependent upon the specific dinucleotide occupying these two positions. Consistent with this idea, we have shown that the order of base preference at position 6 for both *Ubx* and *Dfd* homeodomains is dependent upon the identity of the base at position 7 (see Table I). Residue 54 may also be involved in base recognition outside the TAAT core since M54 in the *Antennapedia* structure appears to interact with a base at position 6 (Otting *et al.*, 1990).

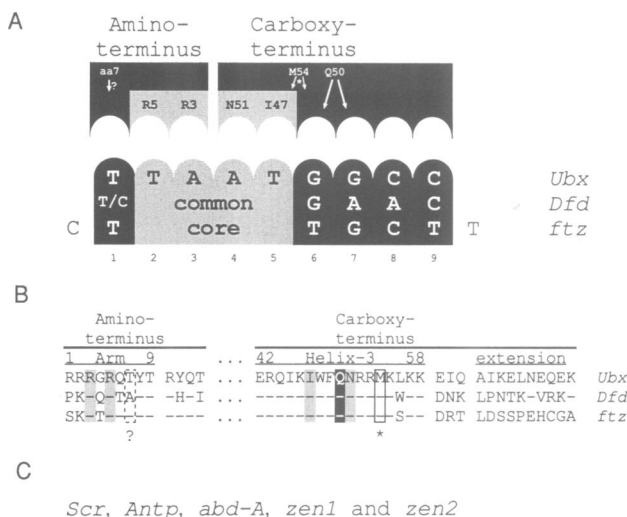


Fig. 6. The TAAT-preferring class of homeodomains. (A) Schematic summary of binding sites and corresponding contact residues for *Ubx*, *Dfd* and *fushi tarazu* homeodomains. Preferred binding sites for the three homeodomains are aligned with the common TAAT core shaded in gray and flanking bases in black. Contact residues for the amino- and carboxy-termini of the three homeodomains are correspondingly shaded. The asterisk denotes uncertainty in the location of the base(s) contacted by residue 54 and the question mark indicates a contact proposed for amino acid residue 7 (see text). The numbering schemes are as in Figure 1A for amino acid residues and as in Figure 3 for DNA binding site sequences. Data for *ftz* (*fushi tarazu*) are from Florence *et al.* (1991). (B) Sequence alignment of amino- and carboxy-terminal residues from *Ubx*, *Dfd* and *ftz* homeodomains. (C) *Drosophila* homeodomain proteins likely to bind preferentially to a TAAT core. Some *Drosophila* homeodomains containing R3, R5, I47, N51 and M54 residues are listed. These residues should be diagnostic for TAAT-preferring homeodomains (see text).

Curiously, despite the shared presence of Q50 and M54, the binding site sequences for the three homeodomains tabulated in Figure 6 show distinct dinucleotide preferences at positions 6 and 7. In addition, preferences at positions 8 and 9 differ. From our selection experiments with *Ubx* and *Dfd* chimeric homeodomains, we know that the carboxy-terminal region functionally responsible for these preferences includes several amino acid differences within the homeodomain and several additional differences present in the extension (Figure 6B). We do not know which of these residues are responsible for the differences in base preference; one possibility is that base preferences at positions 6 and 7 are modulated through an effect upon positioning of the Q50 or M54 side chains, thus influencing the dinucleotide preference.

The structural studies suggest no obvious mechanism by which preferences at positions 8 and 9 might be specified, although it is clear from binding studies and from the yeast assays that these positions play some role in recognition. In addition, the $\alpha 2$ binding site includes a functionally important base at a position corresponding to position 8 in our numbering (Wolberger *et al.*, 1991). One possibility is that backbone contacts occurring at these positions provide an indirect readout of the base sequence, thus providing some specificity. Alternatively, residues carboxy-terminal to the homeodomain may be involved in specifying base preferences at these positions. This region has not been resolved in any of the existing structural models, but the evolutionary conservation of these residues within related groups of homeodomains suggest that they play an important role in some aspect of homeodomain protein function. A third possibility is that hydrogen bonds mediated by water molecules play a role in base recognition at these positions (Otwinowski *et al.*, 1988); such water-mediated hydrogen bonding would not have been seen due to the limits of resolution in the current structure determinations.

The yeast homeodomain structure provides a specific suggestion to account for differences in base preference to the 5' side of the TAAT core. In $\alpha 2$, the R7 residue within the amino-terminal arm makes minor groove contacts with bases at positions -1 and 1 (Wolberger *et al.*, 1991). Residue 7 is a threonine in *Ubx* and in *fushi tarazu*, which show a common preference for a T at position 1, while *Dfd*, which prefers T and C equally at position 1, contains an alanine. Base preferences at position 1 of the binding site also correlate with the identity of residue 6, which is a glutamine in *Ubx* and *fushi tarazu* and a threonine in *Dfd*. In the *engrailed* structure, T6 appears to play a role in positioning the amino-terminal arm by forming a hydrogen bond with a phosphodiester oxygen; the side chain of the A7 residue, however, is too short to make contacts in the minor groove. The general possibility would be that the backbone contact of residue 6 serves to anchor residue 7 which, depending upon its side chain, may then form a contact in the minor groove at position(s) 5' of the TAAT core. That residues 6 and 7 function as a pair conferring sequence specificity is further consistent with their wide divergence among all homeodomains, but with very strong conservation as a pair within related groups (Laughon, 1991). We note further that these two residues are included in the seven residue region from *Ubx* which is sufficient to switch the regulatory specificity of a *Dfd/Ubx* chimera so that it targets the *Antennapedia* promoter (Lin and McGinnis, 1992; see above).

The mechanistic role of differential DNA sequence recognition in the biological specificity of homeotic gene action

Homeodomain proteins appear able to recognize various DNA sites with a wide range of affinities (Percival-Smith *et al.*, 1990; Ekker *et al.*, 1991; Florence *et al.*, 1991; this study). This flexibility suggests the possibility that expression of many target genes could be differentially modulated by a single protein. The cooperative binding of *Ubx* protein to multiple sites (Beachy *et al.*, in preparation; see Figure 5) further suggests a mechanism by which small differences in affinity for individual sites might be summed to give large overall differences in binding and, consequently, specific target regulation. The homeodomain protein encoded by *bicoid* provides a well-described example of regulation that may involve cooperative binding: both the number and quality of sites near the *hunchback* promoter are integrated to form a discrete on/off switch that is very sensitive to the concentration of *bicoid* protein (Struhl *et al.*, 1989; Driever and Nüsslein-Volhard, 1989).

The presence of TAAT or other related core sequences in the binding sites for many homeodomain proteins suggests the possibility that a number of proteins might cooperate or compete in binding to a particular regulatory region. By analogy to MCM1, the specificity-enhancing partner of the yeast $\alpha 2$ homeodomains (Smith and Johnson, 1992), we must also consider the possibility of interactions with other non-homeodomain proteins. Knowledge both of DNA sequence preferences and of how binding to multiple sites is integrated, including the role of any protein partners in this process, will be essential for fundamental understanding of differential target activation by homeotic gene products during the course of development.

Materials and methods

Plasmid constructions

Plasmid pDHD72 was constructed as follows: primers DfdHD-A (5'-ACGGCATATGCCAAAACGCCAACGCACC-3') and DfdHD-B (5'-GATTGGATCTACTTCTTGGCGACGCCCT-3') were used for PCR with 1 μ g *Drosophila* genomic DNA as template; the resulting product was digested with *NdeI* and *BamHI* and cloned into correspondingly digested pET3c (Rosenberg *et al.*, 1987). Plasmids pDDU72 and pUUD72 were constructed using pUHD72 (Ekker *et al.*, 1991), pDHD72, the internal *BglII* site common to both protein sequences and a common vector restriction enzyme site. Plasmid pUDD72 was constructed in a fashion similar to pDHD72, using pDHD72 as template in a PCR reaction using the following primers: primer C (Ekker *et al.*, 1991; 5'-ACGGCATATGCCAAG-ACGCGGCCGA-3'), a bridging primer (5'-CCAGGGTCTGGTAGCG-GGTGTATGTCTGCGGCCGCGTCTTCGCA-3') and primer DfdHD-B (see above). The structure of all these constructs was verified by double-stranded sequence analysis; plasmid pUDU72 was made from pUDD72 using the common internal *MluI* site and a common vector restriction enzyme site.

The yeast inducible regulatory plasmids were constructed by inserting the open reading frame (ORF) sequences from the following expression constructs: pAR3040Dfd (Jack *et al.*, 1988) for *Dfd*; pAR3040Dfd/*Ubx* (Dessain *et al.*, 1992) for DFD/*UBX*HD; and pET3-*UBX* Ib (see below) for *UBX* Ib into the unique *XhoI* site of plasmid pRS129 (R. Sikorski and P. Hieter, 1989). Plasmid pET3-*UBX* Ib was constructed by insertion of a PCR generated *NdeI*-*PstI* fragment of the *UBX* Ib ORF (the *NdeI* site was introduced at the ATG via PCR mutagenesis at the 5' end of the ORF) as well as a *PstI*-*BamHI* *UBX* Ib ORF fragment from p3712 (Beachy *et al.*, 1985) into pET3c (Rosenberg *et al.*, 1987).

Plasmid pSEAD1' was constructed by inserting an *XhoI*-*AatII* fragment containing a basal promoter upstream of about one-fifth of a *GAL1*-*lacZ* gene (from plasmid pLRAD1; West *et al.*, 1984) and an *AatII*-*Apal* fragment containing approximately the last four fifths of *lacZ* (from plasmid pPD16.43; Fire *et al.*, 1990) into a *XhoI*/*Apal* digested yeast shuttle vector pRS315 (Sikorski and Hieter, 1989). The resulting plasmid (pSEAD1'; see Figure 4)

contains seven unique cloning sites in the polylinker upstream of the promoter: 5'-*Sac*II, *Not*I, *Xba*I, *Sma*I, *Pst*I, *Sal*I and *Xho*I-3'.

The yeast target plasmids were constructed by insertion of annealed oligonucleotides (top strand only shown) into the *Xho*I site of plasmid pSEΔ1': targets a-j and l were 5'-TCGAG(N1-N9)(N1-N9)ATG(N1-N9)(N1-N9)AC-3' where N1-N9 are given in Table II, target k was 5'-TCGAG(N1-N9)(N1-N9)GTT(N1-N9)(N1-N9)TC-3', and target n (the LexA-operator) was 5'-TCGAGCTTTTATGCTGTATA-TAAAACAGTGGTTATATGTACAGTATTATTTC-3'. In all targets the 3' end of the strand indicated above was inserted closest to the promoter.

All cloned PCR products and oligonucleotides were sequenced as described by Hattori and Sakaki (1986) using Sequenase 2.0 (US Biochemical).

Purification of homeodomain proteins

Plasmids pDHD72, pDDU72, pUDD72 and pUDU72 were transformed into *E. coli* strain BL21(DE3) pLysE (Rosenberg *et al.*, 1987) and pUUD72 was transformed into *E. coli* strain BL21(DE3)pLysS (Rosenberg *et al.*, 1987). Induction, harvest and purification by chromatography were essentially as described for UbxHD (Egger *et al.*, 1991) except that the Mono-S and Phenyl Superose fractionations were replaced by chromatography using phosphocellulose. DDU required an additional DNA affinity column purification step, performed as described for UBX Ib (Beachy, P.A. *et al.*, in preparation). DfdHD, DDU, UUD, UDD and UDU proteins eluted from the phosphocellulose runs in peaks around 1.06 M, 0.96 M, 1.12 M, 1.12 M and 1.06 M NaCl respectively (in a buffer containing 5% glycerol, 1 mM DTT, 0.5 mM EDTA and 25 mM NaPO₄, pH 7.5). Protein concentrations were measured using the absorbance at 280 and 205 nm (Scopes, 1987).

Structure, amplification and sequence analysis of the selection oligonucleotide

The 64-base oligonucleotide used for binding site selections was 5'-GTAAAACGACGGCCAGTGGATCCNNNNNNNATTANNNNNN-NGCGGCCCGCGTACTGGGAAAAC-3' where N indicates the use of all four bases in equal parts during synthesis at those positions. Primers 64A (5'-GTAAAACGACGGCCAGTGGAT-3') and 64B (5'-GTTTTTCCCAGTACGGCGGCC-3') were used for amplification and second strand synthesis and labelling. Amplification via PCR was performed using the cycle profile of 94°C, 30 s; 62°C, 30 s; 72°C, 30 s for 20 cycles preceded by 94°C for 9 min according to the manufacturer's instructions for AmpliTaq DNA polymerase (US Biochemical) in 100 μl reactions. 5 μl of the resulting reaction was added to 45 μl of fresh buffer which contained a single primer (64A) and was cycled an additional 10 times. The product was purified on a 2% SeaPlaque (FMC) agarose gel or a 7% polyacrylamide (19:1 acrylamide:bis-acrylamide)/7 M urea gel according to standard protocols (Ausubel *et al.*, 1991). Labelling was performed by the addition of ~2-fold molar excess of primer 64B, extension with the large fragment of DNA polymerase I in the presence of 100 μM dATP, dGTP and dTTP and limiting amounts of [α -³²P]dCTP for 15 min. This was followed by addition of 100 μM cold dCTP and another 15 min incubation. The reaction was extracted with phenol-chloroform (1:1) and purified using a NICK column (Pharmacia). Sequence analysis after each round of amplification was performed with Sequenase 2.0 and MnCl₂ (US Biochemical) using either ³²P-labelled primer 64A or [α -³²P]dCTP.

Selections using the homeodomain peptides and full-length UBX Ib

Three rounds of selection with each homeodomain peptide were performed as follows: protein (final concentration 10 nM) was added to labelled, double-stranded 64mer (final concentration ~1 nM) in binding buffer U [75 mM KCl, 20 mM Tris-HCl pH 7.6, 1 mM dithiothreitol, 50 μg/ml bovine serum albumin (Sigma) and 10% glycerol] for UbxHD, UDD and UDU, or binding buffer D (as buffer U, except 115 mM KCl) for DfdHD, DDU and UUD. Following incubation at 22°C for 20 min, 20 μl aliquots were removed and mixed with 2 μl of competitor DNA to yield a final concentration of 50 nM. The competitor consisted of annealed oligonucleotides Comp A (5'-GAATTCAGATCTTAATGGACTCTAGGATCCC-3') and Comp B (5'-CTCGAGGGATCCTAGAGTCCATTAAGATCTG-3') suspended in binding buffer. Following a 30 min incubation with competitor, samples were electrophoresed for 2.5 h at 400 V through a 15% polyacrylamide gel (30:0.8 acrylamide:bis-acrylamide) containing 0.5 × TBE and 3% glycerol and using a water-cooled apparatus (Hoefer SE 600). Following autoradiography of the dried gels, the bands corresponding to bound DNA were excised, rehydrated, amplified via PCR and subjected to sequence analysis.

Three rounds of selection for sequences specific for protein UBX Ib (Beachy *et al.*, 1988) were performed as follows: UBX Ib (final concentration 10 nM) was added to labelled, double-stranded 64mer (final concentration ~1 nM) in UBX Ib binding buffer [10 mM HEPES, 150 mM potassium

acetate (pH 7.5) 2 mM MgCl₂, 0.1 mM EDTA and 1 mM dithiothreitol] and allowed to bind at 22°C for 20 min. Following addition of competitor to 50 nM (see above), 100 μl aliquots were filtered through nitrocellulose as described (Ausubel *et al.*, 1991). DNA retained after 15 min incubation with competitor was recovered by phenol extraction of the nitrocellulose filter and ethanol precipitation, followed by amplification via PCR and sequence analysis.

Quantitative sequence analysis of selected oligonucleotides

A Phosphorimager and storage phosphor screen (Molecular Dynamics) were used to acquire a digitized image of the gel to be analyzed. Quantification involved generation of a line graph for each lane, definition of width windows for each peak, and integration of the pixel values within this window. Peak widths were defined on the unselected DNA lanes for each position and then applied across the gel. Integration of defined peaks was performed by summing the pixel values within each peak, which yielded values corresponding to the intensities of each band. Peak values at positions -5 and -4 for Figure 2 (or -3, -2 and -1 for Figure 3B) were used to normalize the values in each lane to the corresponding unselected lanes for positions 5' of the TAAT; positions 10 and 11 were similarly used for normalization 3' of the TAAT. The ratios of each normalized peak value to the peak value in the corresponding unselected lane was used to construct a preference index. This index is scaled so that the sum of the index for all four bases at a particular position is always equal to 4; the value for a particular base is >1 if its presence is favorable for selection and <1 if its presence is unfavorable. The histograms in Figure 3 utilize the preference index -1. A value greater than or less than zero thus indicates selection for or against, respectively, the particular base at that position.

Dissociation rate constant measurements

DNA sequences used in the dissociation rate constant studies were 5'-AATTCAGATCTT(N1-N9)ATGGATCCCTCGA-3' where N1-N9 are the bases shown in positions 1-9 in Table I for sequences a-l and 5'-TCGATAAGC(N1-N9)GTTCCAGCCGCAATT-3' for sequence m. This DNA either came from pBluescript clones (b and k) or was generated using synthetic oligos. Sequences c-j and l were made double-stranded by extension with the large fragment of DNA polymerase I of 34 base oligonucleotides annealed to a common primer (5'-AATTCAGATCTTTAAT-3'). Sequences a and m were generated by annealing two complementary 34 base oligonucleotides. DNA was purified on a 20% polyacrylamide gel (19:1 acrylamide:bisacrylamide), eluted and further purified using a NACS column (Bethesda Research Labs). DNA was 5' end-labelled using T₄ polynucleotide kinase and [γ -³²P]ATP, the ends filled in with the large fragment of DNA polymerase I (Ausubel *et al.*, 1991), extracted with phenol-chloroform (2:3) and purified with a NICK column (Pharmacia).

Measurements were performed on coded DNA samples whose identities remained unknown to the experimenter until analysis was completed. Binding reactions contained 10 nM protein in Buffer U or D, respectively, for UbxHD and DfdHD (see above) and ~50 pM labelled DNA. A higher salt concentration for experiments with DfdHD was found to bring complex half-lives into a shorter and more accurately measurable range; for this reason, no direct comparisons of binding between UbxHD and DfdHD are made (see text). Binding proceeded for a minimum of 20 min at 22°C, and 20 μl aliquots were removed and mixed with 2 μl of 330 nM competitor DNA (see above). Reactions were incubated at various times before loading onto a 7.5% polyacrylamide gel (30:0.8 acrylamide:bisacrylamide) containing 0.5 × TBE and 3% glycerol. Gels were dried and exposed to film for autoradiography after 45 min of room temperature electrophoresis at 360 V. Storage phosphor screens and a Phosphorimager (Molecular Dynamics) were used to quantify bands corresponding to the bound and free DNA complexes. The dissociation rate constants were determined by plotting ln(fraction DNA bound) as a function of time and using the formula ln(fraction DNA bound) = -k_dt. Half-lives of the complexes (t_{1/2}) were calculated as the time required for half the complexes to dissociate: t_{1/2} = -ln(0.5)/k_d. Only early time points (up to about the first half-life) were used to minimize any effect of reassociation.

Dissociation rate constant measurements with UBX Ib protein

The DNA molecules used for Figure 5A were 444 base *Xba*I-*Msc*I fragments from the yeast target vectors b and k (Table II). These targets were digested with *Xba*I, 3' end-labelled with [α -³²P]dCTP and T₄ DNA polymerase (Ausubel *et al.*, 1991). Following digestion with *Msc*I, the desired fragment was purified by agarose gel electrophoresis (Ausubel *et al.*, 1991). DNA was recovered from the agarose by successive extractions with phenol, phenol-chloroform (2:3) and chloroform and further purified using a NICK column (Pharmacia).

Binding was performed using 10 nM UBX Ib and -40 pM labelled DNA

in UBX Ib binding buffer (see above) modified to contain 20 µg/ml BSA (Sigma, Fraction V) in addition. Binding proceeded for 20 min at 22°C and competitor DNA (see above) was added (final concentration, 50 nM). Incubation with competitor proceeded for the indicated times prior to treatment with DNase I as described (Ekker et al., 1991). An autoradiogram of the resulting gel is shown in Figure 5A.

Quantification was performed using a Phosphorimager and a storage phosphor screen (Molecular Dynamics); the indicated regions were quantified via integration of the digitized image. The initial dissociation rate and the resulting half-lives were calculated as described above.

Correlation between in vitro binding and in vivo transactivation

To determine the Spearman rank-order correlation coefficient (r_s), the 11 sequences shown in Table II were ordered for each protein based upon binding values (for UbxHD and DfdHD from Table I) or transactivation values (for UBX Ib, DFD, and DFD/UBXHD from Table II). Sequences that were the most tightly bound or had the highest activation value were given a rank of 1, and the most poorly bound or activated a rank of 11. Between any two sets of rankings, the Spearman rank-order correlation coefficient (r_s)

is calculated using the following formula: $r_s = 1 - [6/(N^3 - N)] \sum_{i=1}^N d_i^2$,

where d is the difference in ranks by each protein for a particular sequence and N is the number of sequences in the sample (Siegel and Castellan, 1988). The significance of r_s was calculated using a one-tailed test for positive association between any two sets of rankings. Probabilities were taken from a table showing critical values calculated for r_s (Table Q of Siegel and Castellan, 1988).

Yeast strains

Yeast strains were constructed by co-transfection of regulator and target plasmids into strain YPH 500 (Sikorski and Hieter, 1989) using LiAc (Ausubel et al., 1991). Cells were plated on or grown in minimal media [1.7 g/l yeast nitrogen base (Sigma) and 5 g/l ammonium sulfate] supplemented with 4g/l adenine, 2 g/l histidine, 3 g/l lysine and 2 g/l uracil. 2% (w/v) glucose or 2% (w/v) raffinose was included as a carbon source as necessary.

Each yeast strain was grown in liquid selective media containing 2% glucose to an OD₆₀₀ of 1–2. Aliquots of these cells were washed and resuspended in liquid selective media containing 2% raffinose and allowed to grow to an OD₆₀₀ of ~1. These cells were diluted to an OD₆₀₀ of 0.05–0.08, allowed to grow to an OD₆₀₀ of ~0.25 and then induced with 0.1 vol 20% galactose for 4 h. β-galactosidase activity was measured in a kinetic assay using multiple end-point determinations (Ausubel et al., 1991) to generate a plot of β-galactosidase activity as a function of time; the slope of the generated line is the derived activity value (Miller, 1972).

Acknowledgements

We wish to thank S.Dessain and W.McGinnis for generously giving pAR3040Dfd and pAR3040Dfd/Ubx expression plasmids, R.Sikorski and P.Hieter for their yeast plasmids pRS129 and pRS315 and yeast strain YPH500, R.West for plasmid pLR1Δ1, A.Fire for plasmid pPD16.43, and A.Collector and C.Wendling for oligonucleotide synthesis. We also thank N.Craig, R.Johnson, G.Kato, S.Parks, C.Thummel and C.Wolberger for critical readings of the manuscript. S.C.E. is a predoctoral fellow of the March of Dimes Birth Defects Foundation.

References

- Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A. and Struhl, K. (1991) *Current Protocols in Molecular Biology*. Greene Publishing Associates and Wiley Interscience, New York.
- Beachy, P.A., Helfand, S.L. and Hogness, D.S. (1985) *Nature*, **313**, 545–551.
- Beachy, P.A., Krasnow, M.A., Gavis, E.R. and Hogness, D.S. (1988) *Cell*, **55**, 1069–1081.
- Bergson, C. and McGinnis, W. (1990) *EMBO J.*, **9**, 4287–4297.
- Blackwell, T.K. and Weintraub, H. (1990) *Science*, **250**, 1104–1110.
- Cribbs, D.L., Pultz, M.A., Johnson, D., Mazzulla, M. and Kaufman, T.C. (1992) *EMBO J.*, **11**, 1437–1449.
- Dearolf, C.R., Topol, J. and Parker, C.S. (1989) *Nature*, **341**, 340–343.
- Desplan, C., Theis, J. and O'Farrell, P.H. (1988) *Cell*, **54**, 1081–1090.
- Dessain, S., Gross, C.T., Kuziora, M.A. and McGinnis, W. (1992) *EMBO J.*, **11**, 991–1002.
- Diederich, R.J., Merrill, V.K.L., Pultz, M.A. and Kaufman, T.C. (1989) *Genes Dev.*, **3**, 399–414.
- Driever, W. and Nusslein-Volhard, C. (1989) *Nature*, **337**, 138–143.
- Ekker, S.C., Young, K.E., von Kessler, D.P. and Beachy, P.A. (1991) *EMBO J.*, **10**, 1179–1186.
- Fire, A., Harrison, S.W. and Dixon, D. (1990) *Gene*, **93**, 189–198.
- Fitzpatrick, V.D. and Ingles, C.J. (1989) *Nature*, **337**, 666–668.
- Florence, B., Handrow, R. and Laughon, A. (1991) *Mol. Cell. Biol.*, **11**, 3613–3623.
- Gavis, E.R. and Hogness, D.S. (1991) *Development*, **112**, 1077–1093.
- Gibson, G., Schier, A., Lemotte, P. and Gehring, W. (1990) *Cell*, **62**, 1087–1103.
- Graham, A., Papalopulu, N. and Krumlauf, R. (1989) *Cell*, **57**, 367–378.
- Guazzi, S., Price, M., De Felice, M., Damante, G., Mattei, M. and Di Lauro, R. (1990) *EMBO J.*, **9**, 3631–3639.
- Hafen, E., Levine, M. and Gehring, W.J. (1984) *Nature*, **307**, 287–289.
- Hanes, S. and Brent, R. (1991) *Science*, **251**, 426–430.
- Hattori, M. and Sakaki, Y. (1986) *Anal. Biochem.*, **152**, 232–238.
- Hayashi, S. and Scott, M.P. (1990) *Cell*, **63**, 883–894.
- Hoey, T. and Levine, M. (1988) *Nature*, **332**, 858–861.
- Jack, T., Regulski, M. and McGinnis, W. (1988) *Genes Dev.*, **2**, 635–651.
- Kenyon, C. and Wang, B. (1991) *Science*, **253**, 516–517.
- Kissinger, C.R., Liu, B., Martin-Blanco, E., Kornberg, T.B. and Pabo, C.O. (1990) *Cell*, **63**, 579–590.
- Kornfeld, K., Saint, R.B., Beachy, P.A., Harte, P.J., Peattie, D.A. and Hogness, D.S. (1989) *Genes Dev.*, **3**, 243–258.
- Kuziora, M.A. and McGinnis, W. (1989) *Cell*, **59**, 563–571.
- Kuziora, M.A. and McGinnis, W. (1991) *Mech. Dev.*, **33**, 83–94.
- Laughon, A. (1991) *Biochemistry*, **30**, 11357–11367.
- Lin, L. and McGinnis, W. (1992) *Genes Dev.*, **6**, 1071–1081.
- Lopez, A.J. and Hogness, D.S. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 9924–9928.
- Malicki, J., Schughart, K. and McGinnis, W. (1990) *Cell*, **63**, 961–967.
- Mann, R.S. and Hogness, D.S. (1990) *Cell*, **60**, 597–610.
- McGinnis, W. and Krumlauf, R. (1992) *Cell*, **68**, 283–302.
- McGinnis, N., Kuziora, M.A. and McGinnis, W. (1990) *Cell*, **63**, 969–976.
- Miller, J.H. (1972) *Experiments in Molecular Genetics*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- O'Connor, M.B., Binari, R., Perkins, L.A. and Bender, W. (1988) *EMBO J.*, **7**, 435–445.
- Otting, G., Qian, Y.-q., Muller, M., Affolter, M., Gehring, W. and Wutrich, K. (1988) *EMBO J.*, **7**, 4305–4309.
- Otwinowski, Z., Schevitz, R.W., Zhang, R.G., Lawson, C.L., Joachimiak, A., Marmorstein, R.Q., Luisi, B.F. and Sigler, P.B. (1988) *Nature*, **335**, 321–329.
- Percival-Smith, A., Müller, M., Affolter, M. and Gehring, W.J. (1990) *EMBO J.*, **9**, 3967–3974.
- Ptashne, M. (1988) *Nature*, **335**, 683–689.
- Regulski, M., Harding, K., Kostriken, R., Karch, F., Levine, M. and McGinnis, W. (1985) *Cell*, **43**, 71–80.
- Regulski, M., McGinnis, N., Chadwick, R. and McGinnis, W. (1987) *EMBO J.*, **6**, 767–777.
- Regulski, M., Dessain, S., McGinnis, N. and McGinnis, W. (1991) *Genes Dev.*, **5**, 278–286.
- Rosenberg, A.H., Lade, B.N., Chui, D.-s., Lin, S.-W., Dunn, J.J. and Studier, F.W. (1987) *Gene*, **56**, 125–135.
- Rosenfeld, M.G. (1991) *Genes Dev.*, **5**, 897–907.
- Samson, M.-L., Jackson-Grusby, L. and Brent, R. (1989) *Cell*, **57**, 1045–1052.
- Sauer, R.T., Smith, D.L. and Johnson, A.D. (1988) *Genes Dev.*, **2**, 807–816.
- Scopes, R.K. (1987) *Protein Purification: Principles and Practice*. Springer-Verlag, New York.
- Scott, M.P., Tamkun, J.W. and Hartzell, G.W. (1989) *Biochim. Biophys. Acta*, **989**, 25–48.
- Siegel, S. and Castellan, N.J., Jr (1988) *Nonparametric Statistics for the Behavioral Sciences*. McGraw Hill, New York.
- Sikorski, R.S. and Hieter, P. (1989) *Genetics*, **122**, 19–27.
- Smith, D.L. and Johnson, A.D. (1992) *Cell*, **68**, 133–142.
- Struhl, G., Struhl, K. and Macdonald, P.M. (1989) *Cell*, **57**, 1269–1273.
- Treisman, J., Gonczy, P., Vashishtha, M., Harris, E. and Desplan, C. (1989) *Cell*, **59**, 553–562.
- Treisman, J., Harris, E. and Desplan, C. (1991) *Genes Dev.*, **5**, 594–604.
- West, R.W., Yocum, R.R. and Ptashne, M. (1984) *Mol. Cell. Biol.*, **4**, 2467–2478.
- Wolberger, C., Vershon, A.K., Liu, B., Johnson, A.D. and Pabo, C.O. (1991) *Cell*, **67**, 517–528.
- Wysocka-Diller, J.W., Aisemberg, G.O., Baumgarten, M., Levine, M. and Macagno, E.R. (1989) *Nature*, **341**, 760–763.

Received on June 22, 1992