The EMBO Journal vol.10 no.5 pp.1179-1186, 1991

Optimal DNA sequence recognition by the Ultrabithorax homeodomain of *Drosophila*

Stephen C.Ekker, Keith E.Young, 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 D.Nathans

The 61 amino acid homeodomain is conserved among members of a family of eukaryotic DNA-binding proteins that play regulatory roles in transcription and in development. We have refined a rapid method for determining optimal DNA binding sites and have applied it to a 72 amino acid peptide containing the homeodomain of the *Ultrabithorax* (*Ubx*) homeotic gene of *Drosophila*. The site (5′-TTAATGG-3′) is tightly bound (K_D $\sim 7\times10^{-11}$ M) by the Ubx homeodomain peptide; the four central TAAT bases of this sequence play a primary role in determining the affinity of binding, with significant secondary contributions deriving from the flanking bases. Although previously defined genomic sites contain multiple TAAT sequences with flanking bases distinct from those in the optimal binding site, we have found a new binding site with seven near-perfect repeats of the optimal sequence; this site is located in the promoter region of *decapentaplegic*, a probable Ubx regulatory target. The presence of a TAAT motif in the binding sites for most other homeodomain proteins suggests the existence of a conserved mechanism for recognition of this core sequence, with further specificity conferred by interactions with bases flanking this core.

Key words: development/DNA recognition/*Drosophila*/homeodomain/*Ultrabithorax*

Introduction

The homeodomain was first recognized as a 61 codon region of similarity in the sequences of several *Drosophila* genes that play important roles in embryonic development. These sequences are present in many other genes from higher and lower eukaryotes whose products also appear to function in transcriptional or developmental regulation (Gehring, 1987; Scott *et al.*, 1989). The homeodomain is sufficient for sequence-specific DNA binding activity, even without the context of a flanking polypeptide sequence (Mihara and Kaiser, 1988; Muller *et al.*, 1988). The C-terminal half of the homeodomain resembles the DNA-contacting helix-turn-helix portion of certain prokaryotic repressors and, indeed, recently determined structures for the Antennapedia (Antp) and engrailed (en) homeodomains of *Drosophila* confirm a helix-turn-helix conformation for this region (Qian *et al.*, 1989; Kissinger *et al.*, 1990). The functional significance of this structural homology is supported by the observation that DNA sequence preferences can be redirected by amino acid changes in portions of homeodomain proteins corresponding to the recognition helix of the prokaryotic helix-turn-helix motif (Hanes and Brent, 1989; Treisman *et al.*, 1989; Percival-Smith *et al.*, 1990). Functional studies of chimeric proteins in the *Drosophila* embryo indicate that regulatory specificity is determined, at least in part, by the identity of the homeodomain present within a particular protein (Gibson *et al.*, 1990; Kuziora and McGinnis, 1989; Mann and Hogness, 1990).

Our focus here is upon *Ultrabithorax* (*Ubx*), a homeotic gene within the lowest tier of the genetic control hierarchy that directs early *Drosophila* development. *Ubx* operates within the segmented framework established in the embryo by earlier-acting maternal effect and segmentation genes; its function is to specify the unique features of parasegments 5 and 6 (PS5 and PS6), which together constitute a contiguous region including the posterior thorax and a portion of the first abdominal segment (reviewed in Beachy, 1990). As a result of differential splicing, the *Ubx* transcription unit produces a family of closely related proteins (Beachy *et al.*, 1985; Kornfeld *et al.*, 1989; O'Connor *et al.*, 1988) which appear to function in the control of transcription (Johnson and Krasnow, 1990; Krasnow *et al.*, 1989; Samson *et al.*, 1989; Thali *et al.*, 1988). Since each resultant protein product contains the homeodomain, each is presumably capable of sequence-specific DNA binding; this property, however, has been directly studied for only one of the Ubx protein structures (Beachy *et al.*, 1988).

Although DNA binding sites for a variety of homeodomain proteins have been identified, optimal sites have not been defined; nor have the contributions of individual bases to the affinity of binding been systematically examined. Such information is crucial for understanding how homeodomain proteins discriminate between DNA binding sites, and in particular, how homeodomain proteins might bind cooperatively or competitively to DNA in situations where a number of different homeodomain regulatory proteins are present within a cell. Such information is also essential as a basis for interpreting three-dimensional structure data concerning the molecular mechanisms of DNA sequence recognition.

We have refined a method for determining the optimal binding site of a DNA-binding protein; this method provides a statistical indication of the relative importance of individual bases within the binding site. We have applied this method to a homeodomain peptide encoded by *Ubx* and have confirmed the validity of the statistical predictions *via* a biochemical analysis of several binding site sequence variants. Our results indicate that the Ubx homeodomain optimally recognizes a seven base pair sequence (5′-TTAATGG-3′), within which the central TAAT core plays a major role in determining the affinity of binding. Bases flanking the TAAT core further contribute to overall affinity and hence to sequence specificity. Since TAAT is a sequence common to many other homeodomain binding

sites, interactions with flanking bases may provide the basis for discrimination between binding sites by different homeodomain proteins.

Results

Selection of binding site oligonucleotides

Our strategy to define the optimal *Ubx* binding site is an amplification of the procedure described by Oliphant *et al.* (1989) and is similar to a method recently reported (Blackwell and Weintraub, 1990). We used an affinity matrix containing immobilized Ubx homeodomain peptide to select binding site sequences from a population of 70 base oligonucleotides with a random sequence core. As illustrated in Figure 1A, the 12 bp random sequence core within the 70mer was flanked by restriction sites for cloning and end sequences homologous to two 17 base primers which were used either for amplification in the polymerase chain reaction (PCR) or to make the 70mer double-stranded. The general scheme (Figure 1B) was to select specifically bound sequences by loading double-stranded 70mers on to the Ubx homeodomain affinity column at low salt concentration; tightly bound DNA was then isolated by washing the column at intermediate salt and eluting at high salt concentrations. The design of our oligonucleotides permitted amplification by PCR, thus facilitating handling and cloning of the small DNA quantities that remain after multiple rounds of selection.

The 72 amino acid polypeptide used to construct the affinity matrix was expressed in *E.coli* using pET3c, a plasmid expression vector containing a T7 promoter (Rosenberg *et al.*, 1987). The DNA fragment to be inserted in pET3c was generated by a PCR with a *Ubx* cDNA template. The primers for this PCR were designed such that appropriate initiation and termination codons and restriction sites for cloning into pET3c were added in a single operation (Figure 2A; see Materials and methods). The plasmid used for Ubx homeodomain expression (pUHD-72; Figure 2B) carried tandem repeats of the homeodomain coding fragment, both of which contained the 61 codons of the Ubx homeodomain with a 10 codon C-terminal extension (amino acid residues 295–365 in the UBX L11 open reading frame; Beachy, 1990; Kornfeld *et al.*, 1989). We do not know which of these repeats was expressed, although each open reading frame was initiated by a methionine codon and followed by an amber chain termination codon; we suspect that the particularly high level of homeodomain accumulation

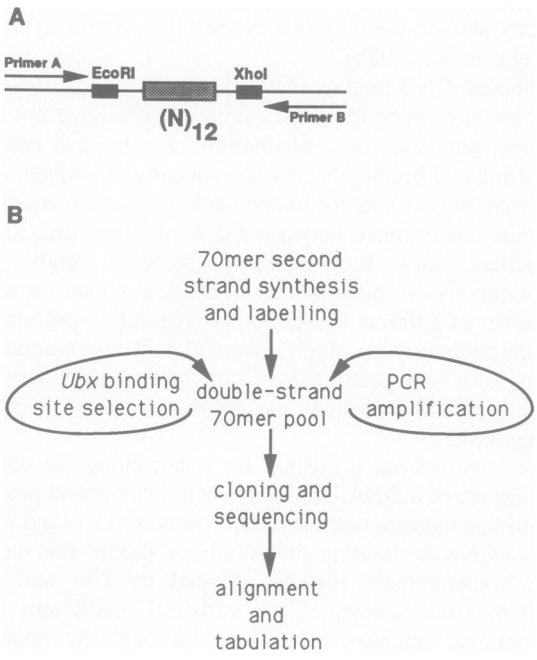

Fig. 1. General strategy for determination of optimal binding site sequences recognized by DNA-binding proteins. (**A**) Oligonucleotide used for binding site selection. The 70 base oligonucleotide contained a 12 base random core, $(N)_{12}$. The two restriction sites used for cloning and the two primers used for PCR are schematically represented (see Materials and methods for complete oligonucleotide sequences). (**B**) General scheme for the selection and analysis of double-stranded oligonucleotides containing binding sites.

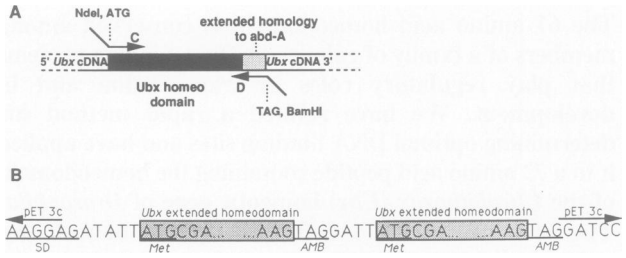

Fig. 2. Construct for expression of the Ubx homeodomain peptide. (**A**) PCR primers and template for generating the DNA fragment containing the Ubx homeodomain fragment. Primers C and D define the extent of the fragment and provide initiation and termination codons as well as restriction sites for cloning. The fragment contains the Ubx homeodomain plus a 10 codon C-terminal extension homologous to the *abd-A* homeotic gene (details in text and in Materials and methods). (**B**) Structure of the Ubx homeodomain expression plasmid. In the final construct, pUHD-72, the sequence encoding amino acids 295–365 in the UBX L11 open reading frame (Beachy, 1990) is present in two tandem repeats. The repeats include the Ubx homeodomain, a 10 codon C-terminal extension as well as initiation (Met) and termination (AMB) codons for each repeat and a Shine–Dalgarno (SD) sequence upstream of the first repeat. Details of the construction are given in Materials and methods.

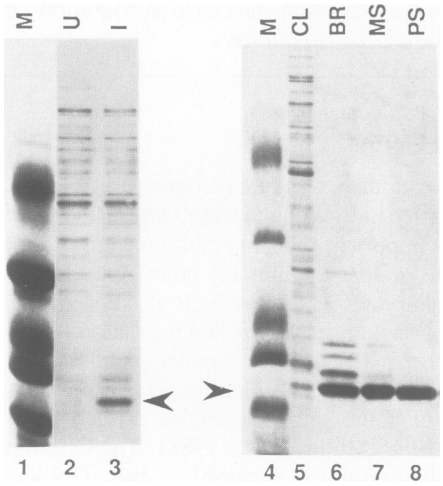

Fig. 3. Expression and purification of the Ubx homeodomain peptide. Protein samples from uninduced (U) and induced (I) cells containing the expression plasmid are shown in lanes 2 and 3; the polypeptide composition from various stages of the purification is shown in lanes 5–8 (CL, cleared lysate; BR, BioRex 70 pool; MS, Mono S pool; PS, Phenyl Superose pool; 5 μg protein are loaded in each lane). Markers (M) in lanes 1 and 4 are 43, 29, 18.4, 14.3 and 6.2 kd. Samples were electrophoresed on a 15% SDS–polyacrylamide gel and stained with Coomassie blue.

produced by this construct may have been due to expression of both repeats. The C-terminal extension was included because it encompasses conserved residues encoded by *Ubx*, the *abd-A* homeotic gene of *Drosophila*, and the leech homeodomain gene *Lox-2* (Wysocka-Diller *et al.*, 1989; Karch *et al.*, 1990).

The identity of the extended Ubx homeodomain peptide was confirmed by its immunoreactivity with a homeodomain-specific monoclonal antibody (data not shown) and by the concordance between predicted and observed mobilities in SDS–PAGE (expected M_r = 9169; see Figure 3A). Purification to homogeneity from *E.coli* extracts (details in Figure 3B and in Materials and methods) was achieved by a combination of conventional column chromatography (Bio-Rex 70) with FPLC (Mono S, Phenyl Superose). The purified peptide was immobilized by attachment to cyanogen bromide activated Sepharose and this affinity matrix was used for three rounds of selection designed to enrich for oligonucleotides containing Ubx homeodomain binding site sequences (details in Materials and methods). The precise degree of enrichment for specific sequences could not be determined because of differences in quantity of DNA loaded at each round and because the activity and specific binding capacity of the matrix-bound protein was unknown. The occurrence of enrichment between rounds one and two was suggested by the increase in percentage input DNA retained after the 0.25 M wash (see Table I). Enrichment between rounds two and three was more difficult to assess because of amplification by PCR after the second round, which greatly magnified the quantity of specific DNA being loaded on to the column. The occurrence of enrichment at this stage was suggested by the greater mass of DNA retained during round 3 after the 0.3 M wash. Amplification was again used after the third round of selection to facilitate cloning. The differing sequences of all but two of the oligonucleotides (62 and 63 in Figure 4A) indicate that the PCR amplification itself did not detectably skew the population toward a particular sequence.

Table I. Amount of DNA retained by the affinity column after washing with the indicated NaCl concentration

Round	Amount of DNA retained (as % input)			
	0.25 M Na^+[a]	0.3 M Na^+	0.35 M Na^+	1.0 M Na^+
1	5.4	–[b]	–	0
2	72.3	10.0	–	0
	(141)[c]	(19)		
3	–	(30)	(8.9)	0

[a]Other components of the wash buffer are given in Materials and methods.
[b]A dash indicates that the column was not washed at the indicated concentration during that selection round.
[c]Values in parenthesis are absolute amounts of DNA (in ng).

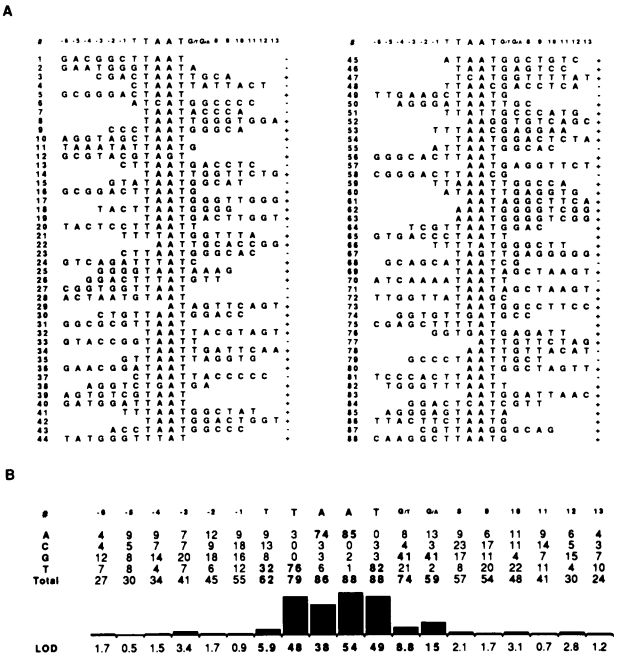

Fig. 4. Tabulation and analysis of Ubx homeodomain binding site sequences present within affinity-selected oligonucleotides. (**A**) Aligned oligonucleotide sequences. The nucleotide sequences of oligonucleotides cloned after the third round of selection are shown as aligned by the method described in the text. Only the random sequence core portion of each oligonucleotide is shown. The (+) and (–) respectively indicate that the *Eco*RI cloning site is present to the left or to the right of the sequence shown. (**B**) Statistical analysis of base usage by position . At each position (numbered as in panel A), the number of occurrences of the four bases is shown along with an estimate of the degree of skewing from random expectation. This estimate was derived as described in the text and is given as a probability in the form of a LOD score (A LOD of 5 indicates a probability of one in 10^5). Seven positions clustered in the center of the tabulation display LOD scores higher than those of the surrounding positions; the sixth and seventh positions show strong secondary preferences when the base of first preference is not present. The consensus binding site sequence thus derived is 5′-T-T-A-A-T-G/T-G/A-3′.

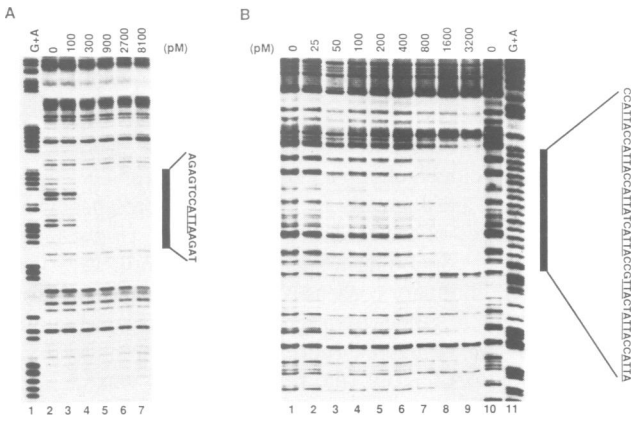

Fig. 5. DNase I protection by Ubx homeodomain of synthetic and naturally occurring DNA sequences containing the consensus binding site. DNase I footprinting was carried out as described in Materials and methods. Ubx homeodomain peptide concentrations are given in pM and protected sequences are indicated by the solids bars. Within protected sequences, the ATTA motif is underlined. (**A**) Footprint of a single consensus binding site sequence. A 15 base protection is centered over the consensus sequence present in oligonucleotide clone 54 (Figure 4A). The protein concentration required for half-maximal protection was between 1 and 3×10^{-10} M (lanes 3 and 4). (**B**) Protection of a *Drosophila* genomic DNA containing multiple repeats of the consensus binding site sequence. The protected region contains seven repeats of the near-perfect consensus sequence TAATGG. These repeats are located near the promoter of the *decapentaplegic* gene (nucleotides 987–1028 according to the numbering of Padgett *et al.*, 1987), a likely target for regulation by Ubx protein. The protein concentration required for half-maximal protection was $\sim8\times10^{-10}$ M (lane 7).

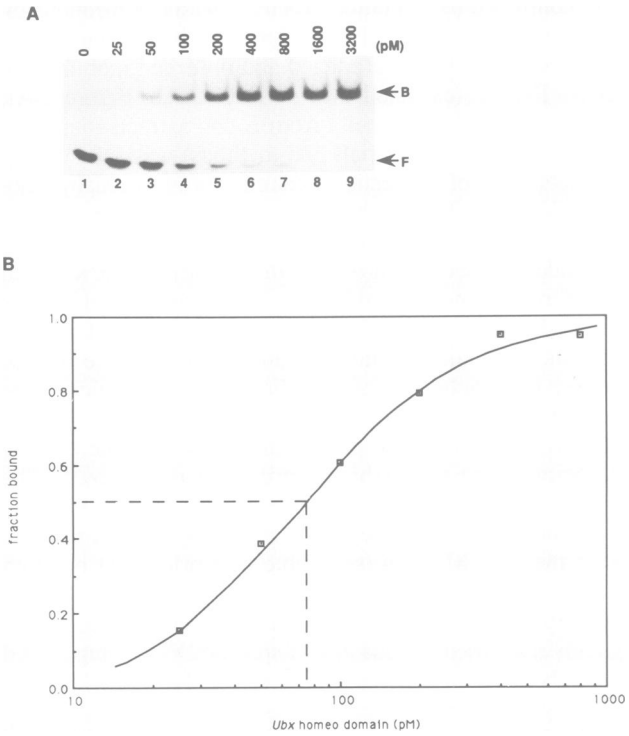

Fig. 6. Determination of equilibrium binding coefficients. (**A**) Equilibrium DNA binding as a function of homeodomain peptide concentration. DNA mobility shift assays were as described in Materials and methods using the indicated protein concentrations and 5 pM end-labeled DNA (from oligonucleotide clone 54, which contains the optimal binding site; see Figure 4A). B is bound, F is free DNA. (**B**) Quantitative analysis. Bands from (A) were excised, counted and the fraction of bound DNA was plotted as a function of the log of Ubx homeodomain peptide concentration. Under conditions of protein excess the concentration required for half-maximal binding (73 pM) may be considered an estimate of the equilibrium binding coefficient (see Table II).

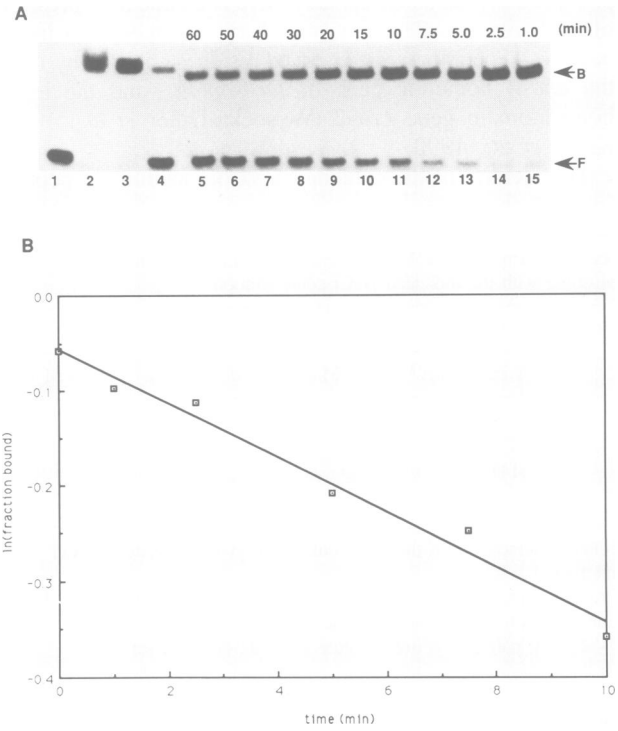

Fig. 7. Determination of dissociation rate constants. (**A**) Decay of protein–DNA complexes as a function of time. The stabilities of protein–DNA complexes were assayed by DNA mobility shift assays (see Materials and methods). Complexes pre-formed with labeled DNA were incubated in the presence of excess unlabeled competitor DNA; times of incubation are indicated in min above lanes 5–15. End-labeled DNA was from oligonucleotide clone 54, which contains the optimal binding site (see Figure 4A). Lane 1 shows the end-labeled DNA alone, lane 2 shows a binding reaction without added competitor, lane 3 shows a binding reaction loaded immediately after addition of competitor (time 0), and lane 4 shows a reaction where competitor was added before protein. B is bound, F is free DNA. (**B**) Quantitative analysis. Bands from (A) were excised, counted and ln(fraction DNA bound) was plotted as a function of time. The dissociation rate constant was determined from the formula: ln(fraction DNA bound) $= -k_d t$. Only early time points were used to minimize the effect of reassociation.

Determination of a consensus binding site sequence

Sequences of cloned oligonucleotides were optimally aligned by inspection, a task greatly facilitated by the presence within most sequences of the tetramer 5′-TAAT-3′ (Figure 4A). All sequences containing a single TAAT motif were used to derive a preliminary consensus; this preliminary consensus was used as a guide to align optimally the remaining sequences that contained more than one complete or an incomplete TAAT motif (the final consensus was identical to this preliminary consensus). In several instances the best alignment for a particular oligonucleotide produced a consensus region that contained one or more bases contributed by sequences flanking the random core of the 70mer. This occurred most frequently in certain specific alignments (e.g. 10 instances where downstream flanking sequences supplied a G at positions 6 and 7, and 12 instances where a T at position 1 was supplied by upstream flanking sequences). High frequencies of specific alignments involving flanking bases as part of the consensus region are an inherent feature of this type of approach; this feature could create distortions in the final consensus through a 'piggyback' effect where other adjacent bases within flanking sequences are made to appear important for binding. The effect was avoided by tabulating for statistical analysis only bases derived from within the random sequence region of the 70mer. As an estimate for the significance of skewing from random expectation, the number of occurrences of the four bases at each position within the tabulation was used to calculate the χ^2 statistic and a corresponding probability (expressed as a LOD score; Figure 4B). Because of the method of tabulation, the probability scores should not be taken as absolute measures of base preference but rather as approximate indications of the degree of constraint.

Of the 19 positions tabulated, a cluster of seven centrally located positions displayed a significantly higher degree of constraint with respect to base preference (Figure 4B). Two of these seven displayed strong secondary preferences when the base of first preference was not present. The consensus sequence thus derived is 5′-T-T-A-A-T-G/T-G/A-3′. Within this consensus, the central four positions (the TAAT core) appeared to be more highly constrained, with LOD scores ranging from 38 to 54 versus LOD scores from 5.9 to 15 for the three outer bases. Specific binding of the Ubx extended homeodomain peptide to a DNA restriction fragment containing the consensus sequence was demonstrated by the DNase I protection experiment shown in Figure 5A, with a 15 base protected region centered over the seven base consensus sequence. Under the conditions

Table II. *Ubx* homeodomain binding to variants of the optimal sequence

	1 T	2 T	3 A	4 A	5 T	6 G/T	7 G/A	K_D[a] (pM)	ΔG[a] (kcal)	K_D (rel)	$k_d \times 100$[a] (min^{-1})	$t_{1/2}$[a] (min)	K_d (rel)
A[b]	T	T	A	A	T	G	G	73, 67[c]	−13.7	1.00	2.3 ± 0.1, 2.2[c]	30.1	1.00
B	T	T	A	A	T	T	A	83	−13.6	1.19	3.1 ± 0.1	22.4	1.35
C	T	T	A	A	T	a	G	89	−13.5	1.27	3.3 ± 0.2	21.0	1.43
D	a	T	A	A	T	T	G	100	−13.4	1.43	4.2 ± 0.3	16.5	1.82
E	T	T	t	A	T	G	G	360	−12.7	5.14	5.0 ± 0.3	13.9	2.17
F	T	T	A	A	g	G	G	360	−12.7	5.14	5.9 ± 0.1	11.7	2.57
G	a	c	A	A	g	c	A	23 000	−10.3	329	190	0.37	83
LOD[d]	5.9	48	38	54	49	8.8	15						

[a]Equilibrium dissociation constants (K_D), dissociation rate constants (k_d), free energies (ΔG), and half-lives of the complexes ($t_{1/2}$) were determined as described in Materials and methods and illustrated in Figures 6 and 7. The k_d values (except for [c]A and G) are given as an average of two independent determinations ± the standard error.

[b]Sequences A, C, D, E and F respectively, are DNAs 54, 69, 50, 66 and 87 as given in Figure 4A. Sequences B (TAATTAGGGGGC$^+$) and G (GACAAGCACTC$^+$) were derived from other experiments and are aligned for maximum fit to the consensus. The small letters indicate differences from the optimal binding site sequence, with upper case letters indicating occurrence of a secondary preference.

[c]K_D and k_D values were also measured for a second DNA containing the optimal binding site sequence (DNA 41 as given in Figure 4A). An average K_D value of 70 pM was used for A to calculate free energy and the relative K_D.

[d]LOD scores are from Figure 4B.

of this experiment, where the homeodomain peptide was in molar excess relative to the DNA fragment, the protein concentration required to produce half-maximal protection ($1-3\times10^{-10}$ M) may be considered an estimate of the equilibrium dissociation coefficient (K_D).

The dependence of binding affinity upon sequence corresponds to the degree of constraint at a particular position

To quantify binding affinities more accurately, we used a DNA fragment mobility shift assay. As shown in Figure 6, the K_D measured for the consensus sequence was 7×10^{-11} M. By similarly measuring the affinities of DNA fragments from other oligonucleotide clones of known sequence, we were able to correlate binding affinities with changes at particular positions within the consensus binding site. A further indication of complex stability as a function of DNA sequence was obtained by measurement of the dissociation rate (k_d) for complexes between the Ubx extended homeodomain peptide and various oligonucleotides (Figure 7). Unlike the footprinting and mobility shift equilibrium binding experiments, which were all carried out at approximately physiological ionic strengths, these kinetic experiments were carried out at reduced ionic strength to slow down dissociation rates and thus bring them into a measurable range. (The half-life of the Ubx homeodomain complex with the consensus binding site was <1 min at physiological ionic strengths.) The results from the equilibrium and kinetic experiments are summarized in Table II.

The principal conclusion from our binding measurements is that alteration of the consensus binding site sequence at a given position affects overall affinity of binding (K_D) or complex stability (k_d) to an extent that correlates well with the degree of constraint observed in our statistical analysis (Table II). Thus, single substitutions at the highly constrained positions within the inner core (rows E and F) have greater impact on overall affinity and complex stability than do single (row C) or double substitutions (rows B and D) at positions outside the core. Likewise, substitution at a more highly constrained position within the inner core (row F) has greater impact than substitution at a less constrained position within the core (row E). We also find that effects of substitutions at multiple positions appear to be cumulative, as suggested by comparison of the overlapping single, double and multiple substitutions in rows C, D and G. One interesting exception to these general rules is illustrated in row B, where a double substitution at positions 6 and 7 has less impact than other single or double substitutions at positions outside the inner core (rows C and D). Even this exception, however, is consistent with our tabulation since the bases present at both substituted positions are the secondary preferences indicated by our statistical analysis (*i.e.* two favorable substitutions may have less impact than a single unfavorable substitution). Although we have not systematically tested all possible combinations of single, double and multiple substitutions, the agreement of these binding data with the expectations from the tabulation leads us to conclude that our statistical analysis has good predictive value and that the consensus sequence is the optimal binding site for the Ubx homeodomain peptide.

Finally, by comparing binding of Ubx homeodomain to sequences in rows A and G, we can assess the degree of discrimination between specific and non-specific binding sites. For complex stability at lower ionic strength, the difference was >80-fold, while at the higher ionic strength used for the equilibrium measurements, the difference appears to be >300-fold.

Multiple repeats of a near-optimal binding site at decapentaplegic, a probable Ubx regulatory target

We have identified a new genomic site, bound both by intact Ubx protein and the Ubx homeodomain peptide, which contains multiple repeats of a sequence closer to the optimal binding site than sequences at previously reported sites (Beachy *et al.*, 1988; see Discussion). This site is located near a promoter in the *short vein* region of the *decapentaplegic* (*dpp*) gene, which encodes a protein homologous to the vertebrate growth factor TGF-β and functions in a variety of developmental processes in the *Drosophila* embryo and larva (Padgett *et al.*, 1987; St Johnston *et al.*, 1990). This site contains seven near perfect tandem repeats of the sequence TAATGG, which is a close relative of the TTAATGG optimal binding sequence for the

Ubx homeodomain. As judged from the higher apparent K_D (Figure 5B), the *dpp* site is bound by Ubx homeodomain peptide with lower affinity than the optimal site, consistent with the absence of the first base of the *Ubx* consensus from the repeated motif at *dpp*. In contrast, binding affinity for the *dpp* motif is greater than that for previously defined sites containing multiple repeats of the more distantly related TAA and TAATCG motifs (D.von Kessler and P.A.Beachy, unpublished). While the functional significance of the *dpp* site has not been demonstrated, recent genetic studies indicate positive regulation of *dpp* expression by Ubx in the visceral mesoderm, where Ubx is present in a sharply defined stripe corresponding to parasegment 7 (Immergluck *et al.*, 1990; Reuter *et al.*, 1990). The presence near one of the *dpp* promoters of a sequence tightly bound by Ubx protein suggests that this regulation may be direct and provides an example of near-optimal Ubx binding sites occurring within the *Drosophila* genome.

Discussion

The optimal binding site we have defined differs from previously identified binding sites for intact Ubx protein in several respects, the most striking of which are the small size of the sequence and of its DNase I footprint (7 bp sequence; 15 nucleotide DNase I footprint). The naturally The naturally occurring sites, in contrast, range in size from 40 to 90 bp and consist primarily of multiple tandem repeats of simple sequence elements such as the triplet TAA or the hexanucleotide TAATCG (Beachy *et al.*, 1988). Underlying these apparent differences, however, are some fundamental similarities. We know, for example, that the homeodomain peptide binds the large naturally occurring sites with high affinity in DNAse I protection experiments (K.E.Young and P.A.Beachy, unpublished data). In addition, a site containing four TAA repeats is sufficient for specific binding by the intact UBX L11 protein (Beachy *et al.*, 1988), indicating that the large size of naturally occurring sites is not absolutely essential. It is of interest that the most important positions in the optimal site we report here correspond to a core TAAT tetramer; this motif occurs in multiple copies at all of the known naturally occurring sites due to the repeating nature of sequences at these sites. Indeed, U-A, one of the naturally occurring Ubx binding sites (Beachy *et al.*, 1990; see below), is capable of simultaneously binding multiple Ubx homeodomain peptides (S.C.Ekker and P.A.Beachy, unpublished data); this suggests that the large naturally occurring sites contain multiple core recognition sequences.

The TAAT motif is also present in binding sites reported for many other homeodomain proteins, although sequences outside the TAAT core vary (reviewed by Hayashi and Scott, 1990). These binding site sequences include many which have been shown to mediate a functional response to *Drosophila* homeodomain proteins, either *in vivo* or *in vitro* (Thali *et al.*, 1988; Muller *et al.*, 1989; Krasnow *et al.*, 1989; Winslow *et al.*, 1989; Samson *et al.*, 1989; Hanes and Brent, 1989; Biggin and Tjian, 1989; Johnson and Krasnow, 1990; D.von Kessler, S.C.Ekker and P.A.Beachy, unpublished data).Thus, the TAAT tetramer appears to constitute a common feature recognized by many homeodomain proteins and this notion is strongly supported by the crystallographically determined structure of an engrailed homeodomain–DNA complex (Kissinger *et al.*, 1990). This structure reveals major or minor groove contacts with each base of the TAAT motif. These base-specific contacts involve the side chains of amino acid residues which are conserved in engrailed, Ubx and indeed most other *Drosophila* and vertebrate homeodomains: Arg3 and Arg5 are part of an N-terminal arm and they contact the first two base pairs of the core in the minor groove, while Asn51 and Ile47 are part of the recognition helix and provide major groove contacts with the third and fourth base pairs of the TAAT core (Kissinger *et al.*, 1990).

It has been suggested that residue 50 (residue nine within the recognition helix) plays an important role in modulating the specificity of DNA sequence recognition by homeodomain proteins (Hanes and Brent, 1989; Treisman *et al.*, 1989). More recent work suggests that the side chain of residue 50 contacts a base(s) to the 3′ side of the TAAT core (Kissinger *et al.*, 1990; Otting *et al.*, 1990; Percival-Smith *et al.*, 1990). Since residue 50 in the Ubx homeodomain is glutamine, our tabulation suggests that Gln50 prefers to interact with one or both G-C base pairs following the TAAT core. Alternative modes of interaction by Gln50 may be possible, as suggested by the existence of secondary base preferences at these positions. Since the N-terminal and C-terminal ends of the engrailed homeodomain lie close to the DNA in the crystal structure of the complex (Kissinger *et al.*, 1990), adjacent regions of homeodomain proteins may also be involved in specifying sequence binding preferences. Whether the 10 amino acid C-terminal extension in the Ubx homeodomain is involved in base-specific contacts is not known, but it is interesting in this regard that differential splicing of the Ubx primary transcript (Beachy *et al.*, 1985; O'Connor *et al.*, 1988; Kornfeld *et al.*, 1989) yields structures with alternative amino acid sequences adjacent to the N-terminus of the homeodomain. We are using the method presented in this work to determine how N- and C-terminal residues in larger Ubx peptides might be involved in specific binding.

In light of the importance of the TAAT core for homeodomain binding, it is tempting to speculate that the repeated triplet or hexamer motifs found in many naturally occurring sequences may be capable of binding multiple homeodomain proteins through recognition of the TAAT motif, while the exact affinity of a protein for a particular site might be modulated by the identity of the bases occurring between TAAT motifs. The character of a protein complex associated with a specific site would then depend upon the identity and level of homeodomain proteins expressed within a particular cell and upon the bases between core TAAT elements. From our Ubx homeodomain binding studies, the contributions of bases between core TAAT elements sum to several-fold differences in complex stability or overall affinity. If one accepts the possibility for cooperative binding to many core sites arranged in tandem (through multimerization or other protein–protein interactions), the range of affinities of homeodomain proteins (alone or in combination) for these naturally occurring binding site clusters could be quite large, thus providing a highly specific mechanism for differential gene regulation. In support of this idea, the 45 bp U-A site near the *Ubx* promoter, originally identified as a binding site for Ubx protein (Beachy *et al.*, 1988), has been reported to bind homeodomain proteins encoded by *even-skipped* (Biggin and Tjian, 1989), *abdominal-A* (Samson *et al.*, 1989), and can probably be

bound by other *Drosophila* homeodomain proteins not yet tested. Given a knowledge of the nucleotide sequence at U-A or at other sites it would be interesting to know whether the affinities of single proteins or combinations of proteins are predictable from a knowledge of: (i) the optimal binding sites of individual proteins and; (ii) the energies of interactions between these proteins.

The method we present for determination of optimal binding site sequences is rapid and sensitive. One indication of its sensitivity is its success in correctly predicting the relative importance of bases which, when altered, affect complex stability by $<10\%$ (Table II). This sensitivity stems from the infinitely renewable population of oligonucleotides, which permit many rounds of selection and thus make it possible to pinpoint preferences for specific bases that make minor contributions to overall binding energies. This method therefore should prove useful for studying the binding specificities of groups of closely related DNA-binding proteins (e.g. other homeodomains) as well as for predicting the binding affinities of closely related sites for a particular protein. The three rounds of selection used in this work produced a fairly heterogeneous collection of oligonucleotide sequences which can be driven to a much higher degree of uniformity by further selection (D.von Kessler and P.A.Beachy, preliminary results). An advantage of using fewer selections, however, is the occurrence of intermediate degrees of constraint which permit assignments of relative importance to bases at particular positions.

The small size of the footprint and the lack of dyad symmetry in the optimal binding site sequence suggest that the Ubx homeodomain peptide binds DNA as a monomer; this is consistent with the peptide's behavior as as a monomer in gel filtration (data not shown) and with the monomer binding observed for the engrailed and Antennapedia homeodomains (Affolter *et al.*, 1990; Kissinger *et al.*, 1990; Otting *et al.*, 1990). From our mobility shift experiments the equilibrium dissociation coefficient (K_D) for Ubx homeodomain at physiological ionic strength appears to be $\sim 7\times10^{-11}$ M. Binding as a monomer is somewhat unusual for proteins of the helix−turn−helix class and this affinity is surprisingly high given that prokaryotic helix−turn−helix proteins or peptides bind to half-operator sites only weakly or not at all (see for example, Hollis *et al.*, 1988). The value is comparable, however, to values reported recently for the interaction with a specific DNA site of the Antennapedia homeodomain (Affolter *et al.*, 1990). The ability of homeodomain peptides to independently fold and bind DNA with such high affinity and sequence discrimination suggests that they constitute functional, autonomously acting units whose study should yield genuine insights into the properties of the larger proteins in which they function.

Materials and methods

PCR, sequence determination and plasmids

All PCR amplifications (Saiki *et al.*, 1988) were carried out in 10 mM Tris−HCl (pH 8.3), 50 mM KCl, 1.5 mM $MgCl_2$, 0.01% gelatin, 50 μM each of dNTP, 0.5 μM each of DNA primer and 0.025 units/μl *Taq* polymerase (Cetus). Temperature cycling was done with the PTC-100 (MJ Research) at maximum ramp speed. DNA sequencing was by the double-strand plasmid method of Hattori and Sakaki (1986), using a Sequenase kit (US Biochemical).

The 70 base oligonucleotide synthesized for use in binding site selection was 5′-GTAAAACGACGGCCAGTGAATTCAGATCT$(N)_{12}$GGATCCCTCGAGGTCGTGACTGGGAAAAC-3′ where $(N)_{12}$ indicates an equal mixture of the four bases during synthesis at 12 consecutive positions. Primers A (5′-GTAAAACGACGGCCAGT-3′) and B (5′-GTTTTCCCAGTCACGAC-3′) were used for amplification of the 70mer by cycling 20 times at 94°C (30 s), 58°C (30 s) and 72°C (10 s). Cycling was preceded by 3 min at 94°C and followed by 10 min at 72°C. Following selection for the presence of binding site sequences (see below), double-stranded 70mer was digested with *Xho*I and *Eco*RI and ligated into similarly digested Bluescript (Stratagene). Resulting clones contained 1, 3, or 5 inserts and sequence determination utilized the Bluescript T3, T7, or M13 primers.

Primers C (5′-ACGGCATATGCGAAGACGCGGCCGA-3′) and D (5′-GATTGGATCCTACTTCTCCTGTTCGTTCA-3′) were used for amplification from a *Ubx* cDNA template (p3712; Beachy *et al.*, 1985) to generate the extended homeodomain fragment: 30 cycles at 94°C (1 min), 62°C (45 s), and 72°C (2 min). Cycling was preceded by 9 min at 94°C. The resulting fragment was digested with *Nde*I and *Bam*HI and ligated into similarly digested pET3c (Rosenberg *et al.*, 1987). Plasmid pUHD-72 was selected from among many candidate clones because of its particularly high levels of Ubx homeodomain production. Restriction enzyme digestion and sequence analysis showed that pUHD-72 contained two intact copies of the extended homeodomain. We can not be certain about the genesis of this plasmid, but its structure is suggestive of a ligation event involving products which suffered some exonuclease digestion. (The sequencing of the pUHD-72 insert was done by recloning an *Xba*I−*Bam*HI fragment into Bluescript and using the T3 and T7 primers.)

Homeodomain purification

E.coli strain BL21(DE3) pLysS carrying pUHD-72 was grown in M9ZB at 37°C to OD_{600} = 0.7, then induced with IPTG as described (Rosenberg *et al.*, 1987). Growth continued for 2.5 h, followed by harvest of cells, lysis, clarification and removal of nucleic acids by polyethyleneimine precipitation as described in Beachy *et al.* (1988). Following the polyethyleneimine step, the homeodomain peptide was purified by chromatography with a BioRex 70 column and with a Pharmacia Mono S FPLC column (Figure 3, lanes 6 and 7) essentially as described in Muller *et al.* (1988). Pooled fractions from the Mono-S column (~0.7 M NaCl, 50 mM $NaPO_4$ (pH 7.5), 1 mM dithiothreitol, 10% glycerol) were adjusted to 1.7 M $(NH_4)_2SO_4$ and further fractionated by loading 1 mg portions on to a Pharmacia Phenyl Superose HR5/5 FPLC column. The column was developed with a 20 ml linear concentration gradient from 1.7−0.0 M $(NH_4)_2SO_4$, with homogeneous Ubx homeodomain eluting at ~1 M $(NH_4)_2SO_4$ (Figure 3, lane 8). Pooled fractions from the Phenyl Superose column were dialyzed against storage buffer (0.6 M NaCl, 50 mM $NaPO_4$ (pH 7.5), 1 mM dithiothreitol, 10% glycerol), aliquotted at a peptide concentration of 60 μM and stored at −80°C. Concentration of purified homeodomain peptide was determined using an e_{280} of 9600 M^{-1} cm^{-1}; the yield was ~1 mg/l culture.

Selection of oligonucleotides containing Ubx binding sites

Ubx homeodomain peptide was conjugated to cyanogen bromide activated Sepharose 4B (Pharmacia) in storage buffer and remaining active groups were blocked by treatment with 1.0 M Tris−HCl (pH 8.3). Coupling efficiency was >97%, and the final concentration of Ubx homeodomain was 100 μg/ml gel. Double-stranded ^{32}P-labeled 70mer (10 μg) was prepared to a specific activity of 10^6 c.p.m./μg by annealing to three-fold molar excess primer B and extending with the large fragment of DNA polymerase I in the presence of all four unlabeled dNTPs (500 μM each) and [α-^{32}P]dATP (50 μl final reaction volume). Affinity chromatography was carried out at ~22°C with 1 ml of Ubx−Sepharose resin in a column 0.7 cm in diameter and ~2.5 cm in height. All loadings of double-stranded 70mer were in 0.5 ml of buffer C (50 mM Tris−HCl pH 8.0, 1mM dithiothreitol, 10 μg/ml gelatin) supplemented with 0.1 M NaCl, 5 μg poly d(I-C), and 10 μg *E.coli* tRNA. After loading 7.4 μg of double-stranded 70mer, round 1 of selection proceeded with successive washes of the column with 3 ml buffer C containing 0.1 M, 0.25 M, 0.4 M and 1.0 M NaCl. Fractions from the 0.4 M and 1.0 M washes were pooled and of the 400 ng in these fractions (5.4% of input, see Table I), 195 ng were recovered after phenol extraction and ethanol precipitation. This material was loaded and round 2 of selection proceeded with successive washes with 3 ml buffer C containing 0.1 M, 0.25 M, 0.3 M and 1.0 M NaCl. Fractions from the 1.0 M wash were pooled, concentrated and amplified by PCR (see above). Approximately 2 μg of the amplified DNA was used for labeling according to the protocol for labeling 70mer described above except that primers A and B were both used. This material was loaded and round three of selection proceeded with successive washes with 3 ml buffer C containing 0.1 M, 0.3 M, 0.35 M and 1.0 M NaCl. Fractions from the 1.0 M wash were pooled, concentrated by ethanol precipitation, amplified by PCR, digested with *Eco*RI and *Xho*I and cloned into a Bluescript vector (see above).

DNase I footprinting

The DNA used for Figure 5A was a 260 bp *Not*I–*Ase*I fragment from a Bluescript clone of an oligonucleotide containing the optimal binding site sequence (Figure 4A, sequence 54; Table II,line A). This plasmid was digested with *Not*I, treated with the large fragment of DNA polymerase I in the presence of [α-^{32}P]dGTP and -dCTP, digested with *Ase*I and purified by agarose gel electrophoresis (Sambrook *et al.*, 1989). The DNA used in Figure 5B (supplied by R.Padgett and W.Gelbart) is a ~ 330 bp fragment from the *short vein* region of the *decapentaplegic* gene (St Johnston *et al.*, 1990) that extends from nucleotide 896 to ~1230 in the numbering of Padgett *et al.* (1987). This fragment was 3′ end-labeled with [α-^{32}P]dATP at an *Eco*RI site in flanking vector sequence and purified following *Bam*HI (also present in flanking vector sequence) digestion and agarose gel electrophoresis.

Binding of homeodomain peptide to labeled DNA fragments (at a concentration of ~18 pM) proceeded for 20 min at 22°C in 300 μl binding buffer (10 mM HEPES (pH 7.0), 150 mM potassium acetate, 2 mM $MgCl_2$, 0.1 mM EDTA and 0.1 mM dithiothreitol). To each reaction was added 40 μl of a solution containing 21 mM $CaCl_2$ and 37 mM $MgCl_2$ and this was followed immediately by addition of 4 ng pancreatic DNase I. DNase I digestion was stopped after 30 s at 22°C by addition of 60 μl of stop solution (1.5 M NaCl, 250 mM Tris–HCl pH 8.0, 100 mM EDTA, 5% Na-lauroyl sarcosine and 175 μg/ml *E.coli* tRNA). After extraction with 400 μl phenol–chloroform (1:1), nucleic acids were concentrated by ethanol precipitation and electrophoresed through 0.4 mm 8% polyacrylamide sequencing gels (Sambrook *et al.*, 1989).

Equilibrium and kinetic DNA binding measurements

DNAs used for mobility shift assays were 34 bp *Eco*RI–*Xho*I fragments from selected single insert Bluescript clones of 70mer oligonucleotides (see above). Fragments were separated in 10% polyacrylamide gels and ethanol precipitated following elution from gel slices (Sambrook *et al.*, 1989). Fragments were then 3′ end-labeled with [α-^{32}P]dATP and -dTTP by incubation with the large fragment of DNA polymerase I (Sambrook *et al.*, 1989). Each fragment was purified with a NACS column (Bethesda Research Labs), ethanol precipitated, extracted with phenol–chloroform (1:1), and further purified with a NICK column (Pharmacia). The column purifications were essential for reproducibility of binding measurements. All binding experiments were carried out with freshly-thawed protein aliquots.

Binding reactions for equilibrium experiments proceeded for 20 min at 22°C in 50 μl footprinting binding buffer (see above) modified to also contain 50 μg/ml bovine serum albumin (Miles, Pentex Fraction V) and 10% glycerol. Reactions contained the specified amounts of homeodomain protein and end-labeled DNA fragments at concentrations $\leq$5 pM.

Binding reactions for kinetic experiments contained Ubx homeodomain and an end-labeled DNA fragment at 10 nM and ~0.5 nM concentrations respectively, all in low salt binding buffer (20 mM Tris–HCl pH 7.6, 75 mM KCl, 1 mM dithiothreitol, 50 μg/ml BSA, 10% glycerol). Following binding for a minimum of 30 min at 22°C, 20 μl aliquots were removed and mixed with 2 μl of a solution containing annealed oligonucleotides E (5′-TAATGGTAATGGTAATGGTAATGGTAATGG-3′) and F (5′-AGACAGCCATTACCATTACCATTACCATTACCATTACCAT-TA-3′) at a concentration of 330 nM duplex DNA (final concentration, 30 nM). Reactions were incubated with competitor at 22°C, for the times indicated, before loading on gels.

Samples from equilibrium and kinetic experiments were electrophoresed at room temperature for 1 h at 400 V on 7.5% polyacrylamide gels (30:0.8, acrylamide to bis-acrylamide) containing 0.5×TBE and 3% glycerol. Dried gels were subjected to autoradiography at −80°C with intensifying screens. Bands corresponding to free and bound DNA were excised, placed directly into Safety Solve™ (Research Products International) liquid scintillation cocktail and counted on a Beckman LS 6800 scintillation counter. Free energies were calculated from the formula $\Delta G = -RTln(K_{eq})$, where R is the universal gas constant and T is absolute temperature. $t_{1/2}$ was calculated as the time required for half the complexes to dissociate: $t_{1/2} = -ln(0.5)/k_d$.

Acknowledgements

The authors wish to thank J.Corden, J.Nathans, C.Pabo and C.Wolberger for helpful suggestions at various stages during this work, R.Padgett and W.Gelbart for supplying *dpp* DNA, R.Weinzierl and R.White for their generous gift of anti-homeodomain antibody and A.Collector and C.Wendling for oligonucleotide synthesis. S.C.E. is a predoctoral fellow of the March of Dimes Birth Defects Foundation and P.A.B. is a Neuroscience Fellow of the Alfred P.Sloan Foundation.

References

Affolter,M., Percival-Smith,A., Muller,M., Leupin,W. and Gehring,W.J. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 4093–4097.
Beachy,P.A. (1990) *Trends Genet.*, **6**, 46–51.
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.
Biggin,M.D. and Tjian,R. (1989) *Cell*, **58**, 433–440.
Blackwell,T. K. and Weintraub,H. (1990) *Science*, **250**, 1104–1110.
Gehring,W.J. (1987) *Science*, **236**, 1245–1252.
Gibson,G., Schier,A., Lemotte,P. and Gehring,W. J. (1990) *Cell*, **62**, 1087–1103.
Hanes,S.D. and Brent,R. (1989) *Cell*, **57**, 1275–1283.
Hattori,M. and Sakaki,Y. (1986) *Anal. Biochem.*, **152**, 232–238.
Hayashi,S. and Scott,M.P. (1990) *Cell*, **63**, 883–894.
Hollis,M., Valenzuela,D., Pioli,D., Wharton,R. and Ptashne,M. (1988) *Proc. Natl. Acad. Sci. USA*, 85, 5834–5838.
Immergluck,K., Lawrence,P.A. and Bienz,M. (1990) *Cell*, **62**, 261–268.
Johnson,F.B. and Krasnow,M.A. (1990) *Genes Dev.*, **4**, 1044–1052.
Karch,F.,Bender,W. and Weiffenbach B. (1990) *Genes Dev.*, **4**, 1573–1587.
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.
Krasnow,M.A., Saffman,E.E., Kornfeld,K. and Hogness,D.S. (1989) *Cell*, **57**, 1031–1043.
Kuziora,M.A. and McGinnis,W. (1989) *Cell*, **59**, 563–571.
Mann,R.S. and Hogness,D.S. (1990) *Cell*, 60, 597–610.
Mihara,H. and Kaiser,E.T. (1988) *Science*, **242**, 925–927.
Muller,M., Affolter,M., Leupin,W., Otting,G., Wutrich,K. and Gehring,W.J. (1988) *EMBO J.*, **7**, 4299–4304.
Muller,M.,Thuringer,F.,Biggin,M.,Zust, B. and Bienz, M. (1989) *EMBO J.*, **8**, 4143–4151.
O'Connor,M.B., Binari,R., Perkins,L.A. and Bender,W. (1988) *EMBO J.*, **7**, 435–445.
Oliphant,A.R., Brandl,C.J. and Struhl,K. (1989) *Mol. Cell. Biol.*, **9**, 2944–2949.
Otting,G., Qian,Y.Q., Billeter,M., Müller,M., Affolter,M., Gehring,W. J. and Wüthrich,K. (1990) *EMBO J.*, **9**, 3085–3092.
Padgett,R.W., St. Johnston,R.D. and Gelbart,W.M. (1987) *Nature*, **325**, 81–84.
Percival-Smith,A., Müller,M., Affolter,M. and Gehring,W. J. (1990) *EMBO J.*, **9**, 3967–3974.
Qian,Y.Q., Billeter,M., Otting,M., Muller,M., Gehring,W.J. and Wuthrich,K. (1989) *Cell*, **59**, 573–580.
Reuter,R.,Panganiban,G.E.F.,Hoffmann,F.M. and Scott,M.P. (1990) *Development*, **110**, 1031–1040.
Rosenberg,A.H., Lade,B.N., Chui,D.-s., Lin,S.-W., Dunn,J.J. and Studier,F.W. (1987) *Gene*, **56**, 125–135.
Saiki,R.K., Gelfand,D.H., Stoffel,S., Scharf,S.J., Higuchi,R., Horn,G.T., Mullis,K.B. and Erlich,H.A. (1988) *Science*, **239**, 487–491.
St Johnston,R.D., Hoffmann,F.M., Blackman,R.K., Segal,D., Grimaila,R., Padgett,R.W., Irick,H.A. and Gelbart,W.M. (1990) *Genes Dev.*, **4**, 1114–1127.
Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
Samson,M.-L., Jackson-Grusby,L. and Brent,R. (1989) *Cell*, **57**, 1045–1052.
Scott,M.P., Tamkun,J.W. and Hartzell,G.W. (1989) *Biochim. Biophys. Acta*, **989**, 25–48.
Thali,M., Muller,M.M., DeLorenzi,M., Mattias,P. and Bienz,M. (1988) *Nature*, **336**, 598–601.
Treisman,J., Gonczy,P., Vashishtha,M., Harris,E. and Desplan,C. (1989) *Cell*, **59**, 553–562.
Winslow,G.M.,Hayashi,S.,Krasnow,M.,Hogness,D.S. and Scott,M.P. (1989) *Cell*, **57**, 1017–1030.
Wysocka-Diller,J.W., Aisemberg,G.O., Baumgarten,M., Levine,M. and Macagno,E.R. (1989) *Nature*, **341**, 760–763.

Received on January 3, 1991; revised on February 7,1991