Introduction

For every complex problem there is a solution that is simple, neat and wrong.
- H. L. Mencken.

Transcription Factor specificity.

Understanding of how complex organisms develop from a single cell has progressed from simple anatomical descriptions through the recognition of conserved tissue and cell types to the identification of individual factors that govern the developmental fate of each cell. With the advent of molecular biology it has become apparent that the regulation of patterns of gene expression is essential to establish and maintain cell identities. This regulation is commonly achieved by affecting the frequency with which mRNA synthesis is initiated at each promoter. Since the proteins comprising the transcriptional machinery are ubiquitous, variation in cis-acting DNA sequences must be responsible for variations in transcription levels between genes. In prokaryotes, at least, DNA variations that affect the binding affinity of RNA polymerase are responsible for substantial differences in the level of basal transcription (Gilbert, 1976; Reznikoff and Abelson, 1978). Where induction and repression of transcription are mediated by trans-acting factors, it is also evident that variations in their DNA targets have profound effects on the resulting frequency of initiation (Gerster and Roeder, 1988; Ptashne, 1992). Consequently the regulation of transcriptional initiation is considered largely dependant on the presence of the appropriate protein binding sites in the DNA of each gene. One potential problem presented by this reliance on DNA sequences, is the enormous amount of alternative DNA present in the genome which must *not* be bound, if genes are to be specifically regulated.

Transcription factor specificity has been studied in prokaryotes for some time now, and the dynamics of transcription initiation are perhaps best understood in this system. To assess the preference for operator over random DNA, competition

experiments have been performed using the *lac* repressor (von Hippel et al., 1974; Lin and Riggs, 1975). These experiments suggested a huge preference for the specific site, such that the entire *E. coli* genome was calculated to sequester repressor with a 20-fold lower affinity than a single operator. Even with only about 10 molecules of repressor present per cell, this level of specificity is theoretically ample to give >99% saturation of operator despite the presence of the rest of the genome as a potential competitor (Lin and Riggs, 1975). These authors consider the prospects for eukaryotic regulation by similar factors and note that because of the huge increase in genome size it would be necessary to either increase the concentration of repressor or effective operator sites, decrease the level of competing DNA by some form of masking, or to increase the specificity and binding strength of the eukaryotic regulatory proteins.

The concentration of transcription factors in eukaryotes has been shown to be substantially higher (50000 molecules per cell is not unusual (Krause et al., 1988), and it seems likely that chromosomal compaction reduces the amount of competing DNA to some degree (Vargaweisz and Becker, 1995). However, the specificity of binding of eukaryotic regulators is dramatically reduced by the small size of their binding sites (Faisst and Meyer, 1992). The prokaryotic regulator binding sites mentioned above are sufficiently large that it has been reasonable to assume that similar sites are unlikely to randomly appear in the rest of the genome. For example, the exact dyad site for *trpR* should occur at random once in 4^{20} bp, or once in about 10^6 bacterial genomes (Haran et al., 1992). The equivalent calculation for an eukaryotic regulator with a 10bp site predicts a perfect, high affinity site about every megabase. Consequently, high affinity binding to many inappropriate sites as well as low-level binding to random sequences is likely to reduce regulator occupancy of its target sites in eukaryotes.

A further complication in eukaryotic regulation has arisen from studies of regulator gene families that have indicated that several members of the family may bind similar sites (Hoey and Levine, 1988; Blackwell et al., 1993; Haas et al., 1995). This

not only increases the number of high affinity inappropriate sites, but also raises the possibility of competition between proteins for the target site. It is of great interest to discover how the exquisite spatial and temporal control of transcription required during development can be generated by transcription factors that must overcome these formidable hurdles of specificity. The role of specificity in development has been most extensively studied using the homeo domain family of proteins, which will be the focus of this review.

Homeo domain proteins.

The homeo domain was identified as a highly conserved 61 amino acid motif found in widely diverged species (McGinnis et al., 1984; Scott and Wiener, 1984). Subsequently, proteins containing this motif have been shown to be required for a wide variety of developmental patterning and differentiation events (for reviews see Scott et al., 1989; McGinnis and Krumlauf, 1992). The domain was shown to bind DNA (Desplan et al., 1985) and homeo domain proteins are known to act as transcriptional activators, repressors or both (reviewed by Hayashi and Scott, 1990). Several classes of homeo domain proteins have been identified by sequence comparisons, and in some cases, conservation of residues has been correlated with conservation of function within a class and across species (Scott et al., 1989; Laughon, 1991). The most dramatic example of this is the homeotic complex from *Drosophila* and the Hox cluster from mammals, in which homologous homeobox genes share the same relative position, orientation and exhibit related timing and position of expression in these two highly diverged species (Duboule and Dolle, 1989; Graham et al., 1989).

Several conclusions relevant to specificity can be drawn from these findings. Firstly, there are many homeo domain proteins in most metazoans, and their DNA binding domains are closely related. Secondly, the functions of these homeo domain proteins as determined by their expression patterns and mutant phenotypes, vary widely. If this diversity of phenotypes reflects a diverse set of target genes (which seems likely

but is not yet proven), then these apparently similar DNA binding domains must each have unique target selection properties. Two lines of inquiry have addressed this issue of how homeo domain proteins identify their correct targets. The first examined the relationship between amino acid variation in the homeo domain and DNA binding site preferences *in vitro*, while the second tested which parts of homeo domain proteins were required for their correct function *in vivo*.

Homeo domain DNA-binding specificity *in vitro*.

The first identification of the sequences to which a homeo domain might bind was carried out by using immunoprecipitation of Engrailed (En) protein to isolate bound fragments from essentially random DNA (Desplan et al., 1985). It should be noted that these studies found a surprisingly high frequency of moderate affinity sites (at least 14 from the λ genome), which suggested a site at random about every 3kb and thus a binding target of 5-6 bases. The sequences of the higher affinity sites, determined by footprinting, generated a 10bp consensus that was highly represented amongst the bound sites (Desplan et al., 1988; Hoey and Levine, 1988). When four divergent homeo domain proteins from *Drosophila* were tested for their ability to bind such sites, it was observed that all four could bind sites resembling this consensus, (Hoey and Levine, 1988). One protein, Even-skipped (Eve), also bound a divergent site, and the affinities for the Engrailed consensus varied, but the results strongly suggested that the conservation of amino acids in homeo domains might be reflected in a conservation of their target sites.

To explain how homeo domains might select different targets, work was then concentrated on identifying the highest affinity binding site for each homeo domain *in vitro* and correlating this with its amino acid sequence (reviewed by Laughon, 1991). It became clear that most homeo domains (except the Abdominal-B class) bound preferentially *in vitro* to a DNA site containing the sequence ATTA with varying preferences for the preceding two bases (Hanes and Brent, 1989; Florence et al., 1991;

Wilson et al., 1993; Ekker et al., 1994; Pellerin et al., 1994). The three-dimensional structures of several homeo domains bound to DNA were also determined (Kissinger et al., 1990; Otting et al., 1990; Wolberger et al., 1991), confirming the interactions with the ATTA base pairs and allowing predictions to be made regarding the effect of individual residues on binding specificity. The homeo domain forms a three helix structure with the second two resembling the Helix-Turn-Helix found in prokaryotic regulators such as the λ repressor, although the similarity does not include the points of DNA contact (discussed in Laughon, 1991). The N-terminal arm of the homeodomain was found to be relatively unstructured and to make minor groove contacts with the last two bases of the ATTA 'core' site (see Figure 1.1). The third or 'recognition' helix was shown to make major groove contacts with the previous four bases including the two that varied in the *in vitro* selected binding sites for different homeo domains (Figure 1.1). The residue potentially contacting these two bases was a glutamine (Q_{50}) in the Antennapedia and Engrailed homeo domains for which the structure had been determined (Kissinger et al., 1990; Otting et al., 1990), but this residue varied in other homeo domains, suggesting that this one residue might be responsible for differing target selection between classes of homeo domain.

Several groups pursued this possibility both *in vivo* (see below) and *in vitro* (Hanes and Brent, 1989; Percival-Smith et al., 1990; Florence et al., 1991; Wilson et al., 1993; Ades and Sauer, 1994; Sun et al., 1995). Most of this work focussed on the observation that the Bicoid (Bcd) homeo domain had a lysine at position 50 and preferred a site of GGATTA (Driever and Nusslein-Volhard, 1989), whereas Q_{50} homeo domain protein such as Fushi Tarazu (Ftz) preferred CAATTA (Florence et al., 1991) and bound GGATTA poorly (Percival-Smith et al., 1990). Alterations of Q_{50} to lysine in En, Ftz, and Ultrabithorax (Ubx), generated proteins that now bound GGATTA in preference to CAATTA (Percival-Smith et al., 1990; Ades and Sauer, 1994; Sun et al., 1995), and the opposite substitution of $K_{50} \rightarrow Q$ in Bcd made it now bind CAATTA in preference to GGATTA (Hanes and Brent, 1991). The initial interpretation of these results was that

variations in residue 50 of the homeodomain played a crucial role in determining which of the NNATTA sites might be bound and thus generated homeo domain specificity (Hanes and Brent, 1989).

Several reasons exist for modifying this interpretation. Firstly, the Ftz protein has been shown to bind effectively to 7 of the 16 possible NNATTA sites, suggesting that for at least this protein, residue 50 is not imparting significant specificity (discussed in Laughon, 1991). Secondly, alteration of $Q_{50}\rightarrow K$ in Ubx does not dramatically affect its affinity for wild type sites as measured by its ability to activate from such sites in yeast (Sun et al., 1995), indicating that a glutamine at position 50 is not necessary to direct Ubx to its targets. It should be noted that similar work using Bicoid showed that $BcdK_{50}$ could also activate from multiple CAATTA sites, but in this case a glutamine at position 50 gave significantly higher activation from such sites (Hanes and Brent, 1989; Hanes and Brent, 1991), so in some contexts Q_{50} may improve binding of a Ftz-like target. Thirdly, examination of the constraints at residue 50 in the Engrailed homeo domain revealed that affinity for its optimal site TAATTA was not dramatically altered by changing Q_{50} to lysine or even alanine (which should not contact the DNA) (Ades and Sauer, 1994). Despite having no known contacts with the first two nucleotides of the site, EnA_{50} and $BcdA_{50}$ (Hanes et al., 1994) preferred a TAATTA site to a GGATTA site, clearly indicating that specificity for those two bases exists, but is not generated by the residue at position 50 (Figure 1.2). EnK_{50} was shown to have a much higher affinity for its preferred site GGATTA than EnQ_{50} had for TAATTA, consistent with the examples cited above where a lysine switched preferences toward the GG dinucleotide (Ades and Sauer, 1994).

The simplest interpretation of these results is that homeo domains can discriminate to some extent between the NNATTA sites, but that the preferred site is not dictated by the residue known to contact the NN dinucleotide unless it is a lysine, which mediates a very high affinity interaction with a GGATTA site (Figure 1.2). A consequence of this is

Figure 1.1 Homeodomain base contacts.

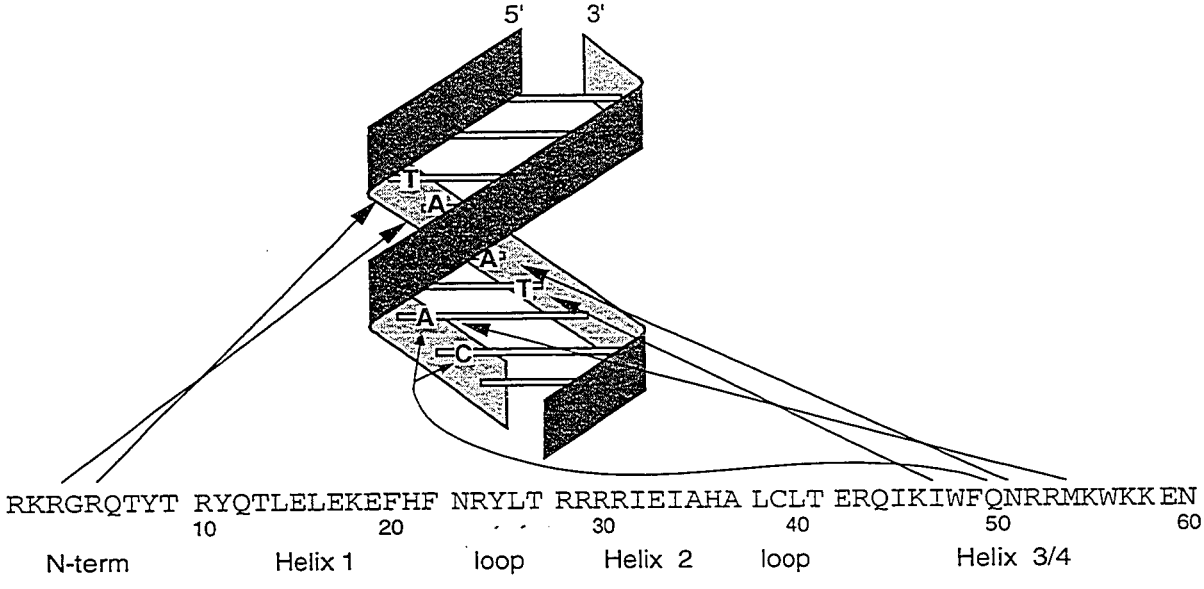


Figure 1.1 Arrows indicate the position of Engrailed homeo domain amino acid contacts with nucleosides in a CAATTA binding site. Residues 3 and 5 make contacts in the minor groove and residues 47, 50, 51 and 54 make contacts in the major groove. In addition, residues 6,25, 31,48,53,55 and 57 make contacts to the phosphate backbone (not shown). Diagram adapted from (Mann, 1995, Fig 2).

that it is still not usually possible to predict what the preferred binding site of a homeo domain will be from its amino acid sequence. For example, Ubx, Ftz and Deformed (Dfd) have identical recognition helices but Ubx preferentially binds CCATTA (Ekker et al., 1991), Ftz prefers CAATTA (Florence et al., 1991) and Dfd prefers TCATTA (Ekker et al., 1992). Homeo domain substitution experiments revealed that the slight difference between Ubx and Dfd preferences was mainly conferred by the third helix plus 13 C-terminal residues (Ekker et al., 1992). Much more structural information on variant proteins binding variant sites is likely to be required before these preferences can be fully explained.

Although the determinants of specificity remain to be identified for the large class of Q_{50} homeo domains, several of them do preferentially bind different sites *in vitro*, so it is reasonable to suppose that they might identify different targets *in vivo* if the protein concentration were limiting. On the other hand, at least two *Drosophila* homeo domain proteins that are co-expressed in some cells (Ubx and Antennapedia, Antp) have indistinguishable preferences for binding sites *in vitro* (Ekker et al., 1994). This is perhaps not surprising since they differ at only 6 positions across the whole domain, but loss of expression generates a dramatically different phenotype for each gene (discussed by Mann and Hogness, 1990). Two non-exclusive explanations are that the *in vitro* assays are not capable of identifying the differences between sites targeted *in vivo* and/or that these proteins give different effects from binding the same targets. This issue will be discussed further below, where evidence is presented suggesting that both of these explanations are correct.

Implications for gene regulation

To return to more general considerations, it seems that a perfect, optimal homeo domain site consists of *at most* nine bases with some redundancy (Ekker et al., 1994). Such a site will occur in a random DNA sequence more than once every 100kb, so even if 99.9% of the human genome were masked by other stably bound proteins, any

Figure 1.2 Effect of Q_{50} mutations on En specificity

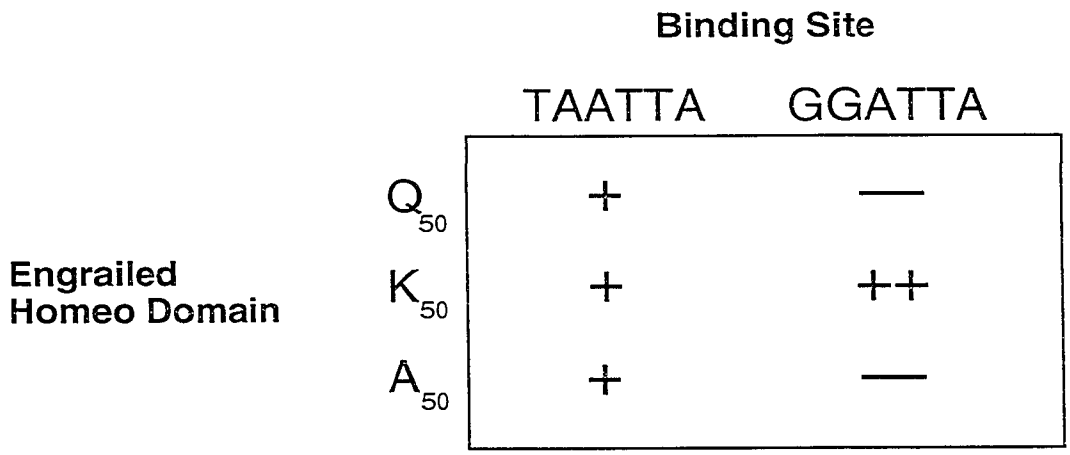


Figure 1.2 The binding affinity for two alternative sites by each Engrailed variant is indicated by +: $8x10^{-11}$- $3x10^{-10}$, ++: $8x10^{-12}$, –: $>5x10^{-9}$. Each variant can bind TAATTA, but only K_{50}binds GGATTA effectively. Site specificity is retained by A_{50}although that residue does not contact the DNA. Data from (Ades and Sauer, 1994).

appropriate target sites would be competed by dozens of fortuitous sites with equal or greater affinity. Further competition would be expected from those non-random sites used by other homeo domains with the same site preferences. Proteins such as Ftz, which apparently only recognizes a redundant six base site (Florence et al., 1991), must presumably contend with many hundreds of fortuitous competitors. Such competition would be reduced if NNATTA were significantly underrepresented in random genomic DNA, but this does not seem to be the case (Gross and Gruss, 1995). These calculations are consistent with studies that have measured the incidence of effective homeo domain sites in essentially random DNA (Desplan et al., 1985; Walter et al., 1994) - the *in vivo* relevance of this will be discussed later.

The relative affinity of several homeo domains for their specific site over random DNA has been estimated at about 100-fold preference for the specific site (Affolter et al., 1990; Ekker et al., 1991; Florence et al., 1991). This represents a surprisingly high affinity for non-specific DNA, particularly when compared with prokaryotic HTH proteins such as *lacI* with a specificity of $>10^6$-fold (Lin and Riggs, 1975). Non-specific DNA will also compete for binding of homeo domains (with 100 fold lower affinity), so the genome considered above should be equivalent to thousands of alternative high affinity sites, and we should expect that most of the protein will be bound to non-specific DNA most of the time (Lin and Riggs, 1975). When this is combined with the occurrence of fortuitous sites mentioned above, such calculations predict that any individual homeo domain target site should never achieve better than a fractional percentage occupancy by its correct protein; a model that would seem to preclude effective gene regulation.

Several possible explanations exist for these *in vitro* results. It is possible that any enhancer regulated by homeo domain proteins could consist of many potential binding sites, not all of which need be occupied at once to give effective regulation. This is consistent with observations that some homeo domain responsive enhancer regions

contain multiple potential binding sites that have a cumulative effect on transcriptional regulation when mutated (Jiang et al., 1991; Appel and Sakonju, 1993; Schier and Gehring, 1993; Zeng et al., 1994; McCormick et al., 1995; Sun et al., 1995). These studies found very few individual potential binding sites that could confer a unique effect on the spatio/temporal regulation of transcription. Consequently it is possible that a low level of individual site occupancy may be partially offset by the presence of many effectively equivalent sites (Figure 1.3). Although this could help to explain how homeo domains could effectively regulate target genes, it does not address the question of how different homeo domain proteins might identify different targets *in vivo*.

Functional specificity *in vivo*.

The second method used to uncover the determinants of differential regulation was to systematically mutate a homeo domain protein, then express it *in vivo* to see the phenotypic effect. Detailed interpretation of these phenotypes was possible for some homeo domains due to the extensive characterisation of homeotic genes (Lewis, 1978). This set of homeo domain proteins is involved in the processes of body plan specification and differentiation, and many of them are expressed in overlapping patterns during development (McGinnis and Krumlauf, 1992). Mutant phenotypes in *Drosophila* suggested that a unique combination of homeo domain proteins in each segment was responsible for its unique fate (Lewis, 1978; Wakimoto and Kaufman, 1981; Sanchez-Herrero et al., 1985). Ectopic expression of such proteins can alter some epithelial cells to resemble tissues where the protein is normally found (Schneuwly et al., 1987; Kuziora and McGinnis, 1988; Gibson et al., 1990; Mann and Hogness, 1990; Lamka et al., 1992; Halder et al., 1995). The difference in phenotype observed when different homeo domain proteins are ubiquitously expressed is strong evidence that they regulate different targets and has been used as the basis for assessing their functional specificity. These results are highly significant since they suggest that the expression of a single homeo domain protein may be sufficient in some cases to direct the differentiation of an entire body part. Caution should be used in interpreting overexpression studies, however,

Figure 1.3 Multiple equivalent binding site model

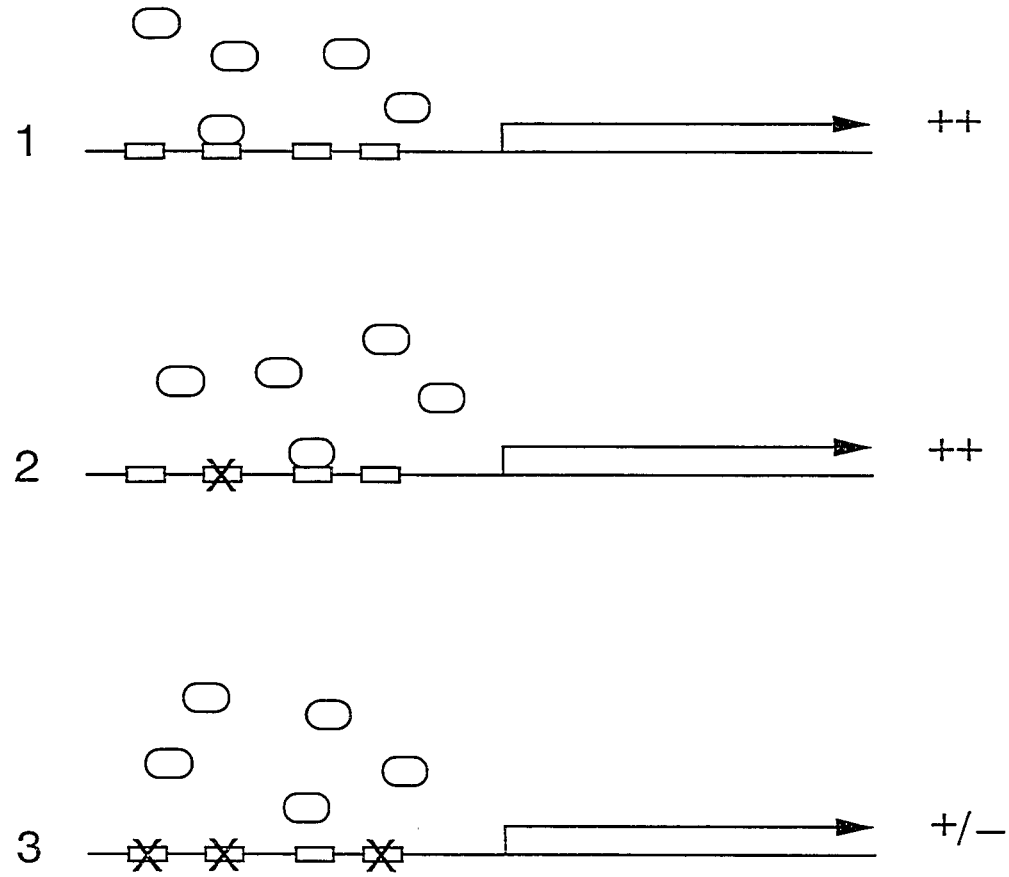


Figure 1.3 (1) Transcriptional control of a promoter may be regulated by multiple DNA elements that can all bind the same protein factor to give the same effect. (2) Mutation of any individual site has little effect. (3) Mutation of more sites has a progressively greater effect on control. This model is proposed to compensate for the lack of site binding affinity of homeo domains and is consistent with several mutagenesis studies (see text).

since the exact time, place and amount of protein expression have been found to affect whether transformation is observed (McGinnis and Krumlauf, 1992).

Because ectopic expression phenotypes resembled a duplication of tissues normally expressing the protein, it was hypothesised that the protein was performing its normal function in these new tissues. Many homeo domain proteins have been mutated then overexpressed in *Drosophila* to establish which parts of the protein are responsible for its ectopic and, by inference, its normal function (Kuziora and McGinnis, 1989; Gibson et al., 1990; Mann and Hogness, 1990; Furukubo-Tokunaga et al., 1993; John et al., 1995). In these experiments, parts of one homeo domain protein were replaced by the equivalent region of another to determine which chimeric proteins retained the *in vivo* function of each parent protein. In each case, replacement of the homeo domain resulted in phenotypes that approximately resembled overexpression of the homeo domain donor protein (see Figure 1.4). Deletion series confirmed that the homeo domain was essential to generate any transformed phenotype (Gibson et al., 1990; Mann and Hogness, 1990). Taken together, these results were interpreted to mean that residues in or near the homeo domain contained the main determinants of functional specificity (Lin and McGinnis, 1992).

Attention was then directed at assessing which residues within the homeo domain might be responsible, using the same approach of ectopic expression (Lin and McGinnis, 1992; Chan and Mann, 1993; Furukubo-Tokunaga et al., 1993; Zeng et al., 1993). These studies concluded that the region of the domain N-terminal to the first helix contained residues that largely determined which transformation would occur upon overexpression. The structural analyses indicated that this region was relatively unstructured but that some residues (3 and 5) lay in the DNA minor groove and could contact the last two bases of the ATTA core site (Kissinger et al., 1990; Otting et al., 1990). The N-terminal region is conserved within a homeo domain class, but varies widely between classes (Scott et al., 1989). This obviously made the N terminal residues

Figure 1.4 Overexpression of Homeodomain Chimeras

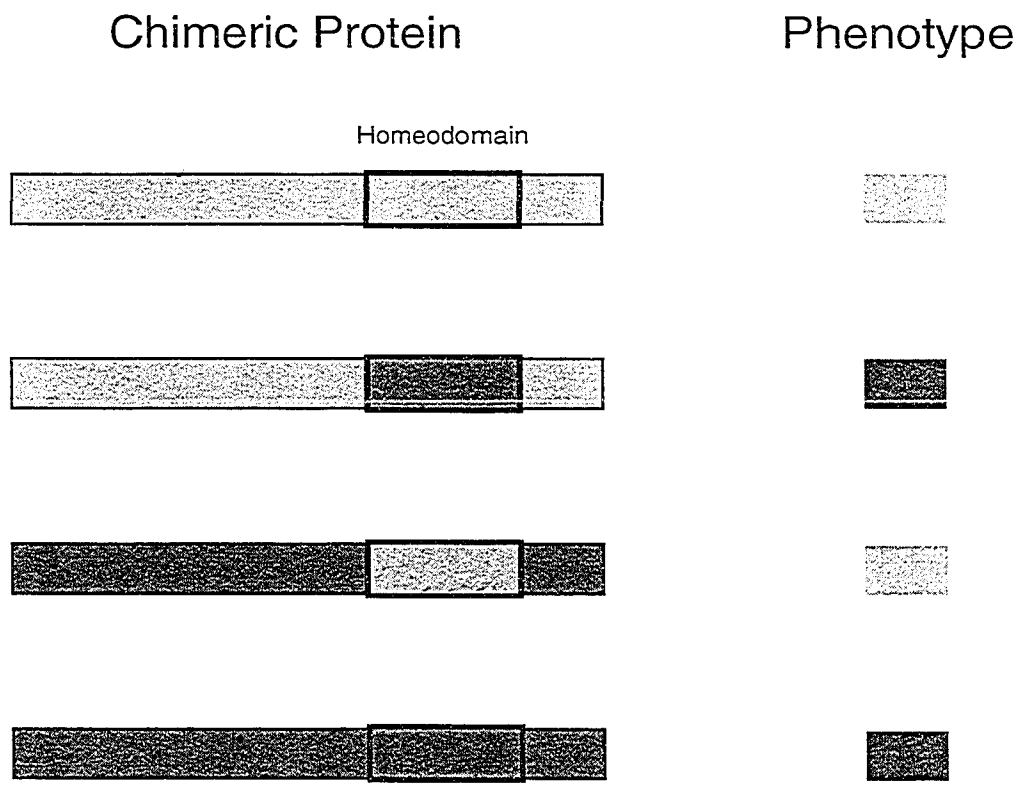


Figure 1.4 When homeotic proteins are ubiquitously overexpressed in *Drosophila* embyos, the result is usually a duplication of the tissues in which the proteins is normally expressed. If the homeo domain is swapped with that of a different homeotic protein, the overexpression phenotype of the chimera generally resembles that of the homeo domain donor protein. The homeo domain clearly has a dramatic impact on function in vivo, whether by directing specific DNA binding or by altering protein contacts.

excellent candidates for the determinants of homeo domain functional specificity. The only N-terminal residues known to make DNA contacts (R_3 and R_5) are highly conserved between homeo domains and the bases contacted are also common to most homeo domain binding sites (Laughon, 1991). This implies that if the N-terminal region has a role to play in generating specificity, it does not come about by influencing site binding preferences. Further analysis suggested that the N-terminal region was necessary but insufficient, with significant effects on specificity from residues C-terminal of the recognition helix (Chan and Mann, 1993; Heberlein et al., 1994). These residues also vary widely between homeo domain classes and are not known to contact the DNA. Since many of the residues implicated in homeo domain specificity are predicted to face away from the DNA, it has been hypothesised that they may mediate protein-protein rather than protein-DNA interactions (Furukubo-Tokunaga et al., 1992; Lin and McGinnis, 1992; Chan and Mann, 1993; Zeng et al., 1993). To summarise, ectopic expression of chimeric homeo domain proteins has revealed that the homeo domain and nearby residues play a critical role in determining the phenotypic effects of the protein. The residues implicated suggest that differing effects may reflect different protein contacts rather than different DNA binding preferences.

There are several features of ectopic expression which have limited its usefulness in characterising homeo domain specificity. Firstly, the proteins are produced from a heat shock responsive promoter, which generates very high levels of protein for a limited period (Lis et al., 1983). This is in complete contrast to most homeotic promoters that produce comparatively modest amounts of protein, but do so indefinitely once induced (Morata and García-Bellido, 1976). The amount of homeo domain protein present can clearly affect its regulatory ability because several homeotic proteins (Scr, Ubx, Abd-B) give mutant phenotypes when produced at half-normal levels (Kaufman et al., 1980). The exact timing and temperature of heat shock induction is known to be essential to be able to generate reproducible transformations (Gibson et al., 1990; Mann and Hogness, 1990). The experimental details of heat shock induced expression may thus affect the

resultant transformation: overexpression of Ubx either does (Dessain et al., 1992) or does not (Gonzales-Reyes and Morata, 1990) repress Dfd expression, depending presumably on the amount of protein produced. Chimeric homeo domain proteins produced from their normal promoter have revealed dose-dependant functional determinants quite different from those seen with heat shock induction (Lockett et al., 1993; Heberlein et al., 1994). Thus, although the phenotypes produced by overexpressed homeotic proteins are generally interpretable, it is still a matter of speculation as to how closely these effects may reflect their normal functions.

The second difficulty in interpreting ectopic expression studies comes from uncertainty over which effects are direct and which are mediated by unknown cross regulatory events. For example, when Antp is overexpressed it has a variety of effects primarily in the head. If its homeo domain is replaced with Sex combs reduced (Scr) residues, it now has effects in the thorax similar to those seen for Scr overexpression (Gibson et al., 1990; Zeng et al., 1993). This could be because the chimera now has Scr functional specificity, or instead it could be because it has a novel specificity that directs it to activate Scr, resulting in an Scr overexpression phenotype. These possibilities could be discriminated by testing Scr levels of course, but the same result would be observed if the chimera activated any other protein that might regulate Scr target genes in the same way as Scr does when overexpressed (Figure 1.5). Consequently, it would be imprudent to interpret the phenotype of overexpressed Antp/ScrHD as having Scr functional specificity. Detailed characterisation of homeo domain target enhancers will clarify this issue, but few have been identified and their regulation is still very poorly understood (Zeng et al., 1994; McCormick et al., 1995; Sun et al., 1995). Thus indirect activation by chimeric homeo domain proteins cannot yet be ruled out as an explanation for their phenotypic effects.

Efforts have been made to exclude the possibility of indirect effects by introducing a $Q_{50} \rightarrow K$ mutation in the homeo domain and altering its proposed targets to GGATTA,

Figure 1.5 Direct vs Indirect Regulation by Chimeric Homeodomains

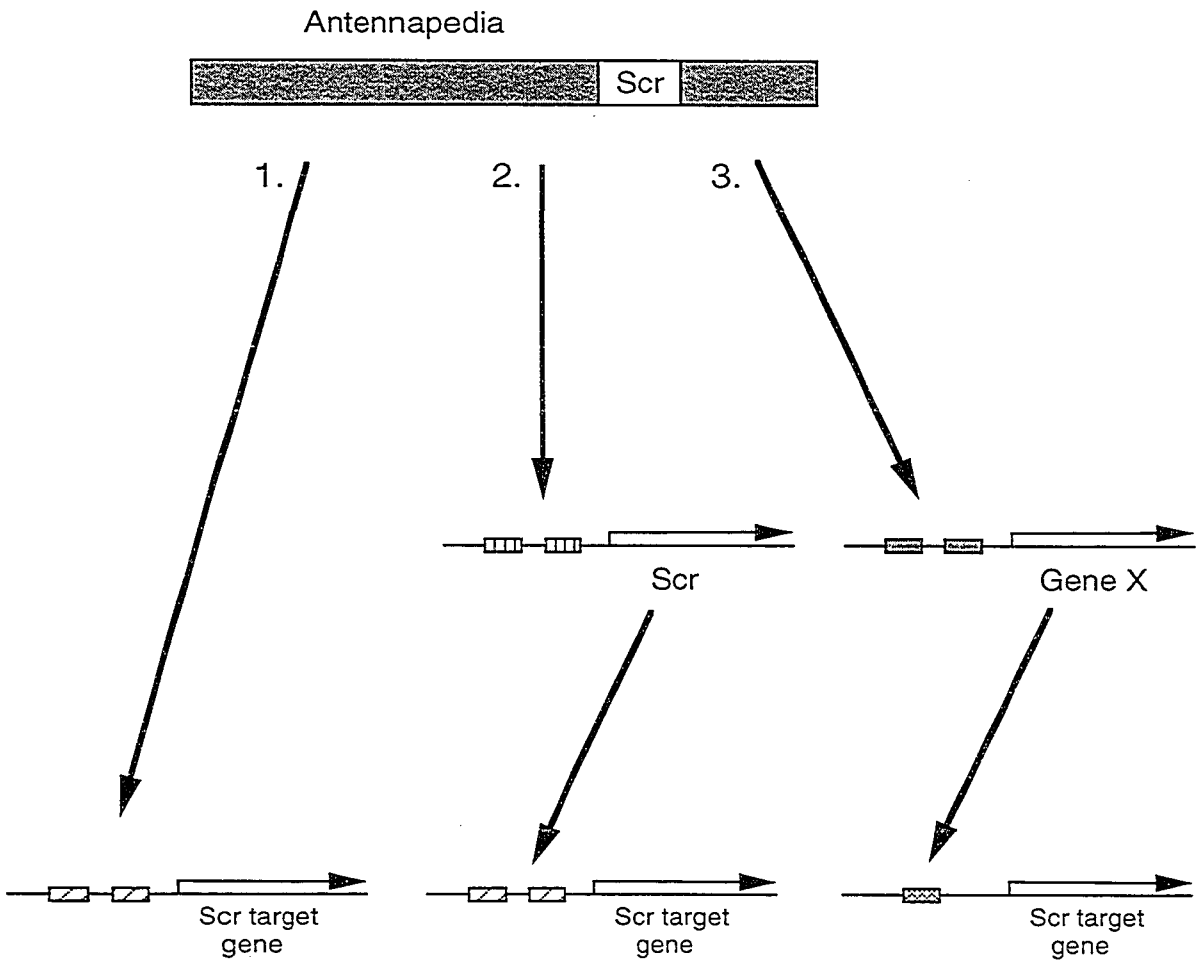


Figure 1.5 The result of replacing parts of the Antennapedia homeo domain with residues from Scr is to generate an Scr-like over expression phenotype (see Fig 1.4). This may occur because: 1. The chimeric protein has Scr binding specificity and thus regulates Scr target genes. 2. The chimera has changed binding specificity allowing it to activate Scr, which generates the Scr phenotype. 3. The chimera may have an entirely novel specificity which causes it to regulate unknown genes which in turn activate Scr targets.
An Scr phenotype therefore does not necessarily imply Scr binding specificity.

which only K_{50} homeo domains prefer (Capovilla et al., 1994; Sun et al., 1995). The results were consistent with some direct interaction between $UbxK_{50}$ and the altered *decapentaplegic* midgut enhancer. However only weak or ectopic activity was observed, suggesting that part of the Ubx regulation of these sites is normally mediated indirectly.

The need for cautious interpretation should be reinforced by the observation that several homeo domain chimeras have effects that are not seen from either parental protein (Gibson et al., 1990; Lin and McGinnis, 1992; Zeng et al., 1993). These results strongly suggest that homeo domain chimeras gain a novel specificity that may overlap with that of the parental proteins, but which also includes additional unknown targets. This does not detract from the real significance of the overexpression studies, which is that alteration of a very small number of residues, particularly in the N-terminal arm of the homeo domain, can dramatically change functional specificity. As with the *in vitro* studies, it is not yet possible to predict the *in vivo* effects of even closely related homeo domain proteins from their sequence, although some of the regions likely to be involved in specificity have been identified.

One important feature of *in vivo* studies is that they generally cannot distinguish between a change in target choice and a change in target regulation. In each case where different effects are seen for variant homeo domains, it is theoretically possible that they can all bind with high affinity to the same binding sites in the same target genes, the only difference being whether they activate transcription. Consistent with this possibility, *in vitro* studies with Ubx have shown that the DNA-binding specificity of the full length protein is the same as that of the homeo domain alone (Ekker et al., 1992), but Ubx splicing variants containing the same homeo domain have variant effects on PNS morphogenesis when overexpressed (Mann and Hogness, 1990). This suggests that functional specificity need not correlate with *in vitro* DNA-binding specificity. This has also been observed in Ubx/Dfd chimeras and between Ubx and Antp, which have

different *in vivo* effects, but indistinguishable DNA-binding specificities (Lin and McGinnis, 1992; Ekker et al., 1994). It remains possible that homeo domain proteins are modified *in vivo* to recognise unique sites, and indeed post-translational modifications have been detected, but these have little effect on affinity and no known effect on specificity of binding (Bourbon et al., 1995). An alternative is that binding to a common DNA site is a necessary but insufficient step in homeo domain transcriptional regulation. The additional requirement may be for an interaction with another DNA-binding protein through the exposed residues known to be important for functional specificity.

Cooperative binding

Prokaryotic regulators are frequently found as multimers that bind a dyad symmetrical site (Hochschild and Ptashne, 1986; Aggarwal et al., 1988; Haran et al., 1992). The affinity with which the second half-site is bound is increased many-fold by protein binding to the first, leading to the formation of a high affinity complex from relatively low affinity components (discussed in Ptashne, 1992). Cooperative binding is also observed in eukaryotes where DNA binding proteins of relatively low affinity can form high affinity homo- and heteromeric complexes (Kouzarides and Ziff, 1989; Murre et al., 1989; Tsai et al., 1989; Xiao et al., 1991). Homeo domains studied in isolation can clearly bind with high affinity as monomers to a monomeric site (Affolter et al., 1990; Florence et al., 1991; Beachy et al., 1993), but this does not preclude the possibility that the complete protein may interact with other DNA-binding proteins to generate cooperative binding. There is now evidence to suggest that this may occur through interactions between homeo domain proteins (Dranginis, 1990; Ingraham et al., 1990; Beachy et al., 1993; Wilson et al., 1993; Chan et al., 1994; van Dijk and Murre, 1994). There is some evidence to suggest that such interactions are common to many related homeo domains (Zappavigna et al., 1994). Homeo domain interactions have also been observed with other classes of DNA binding domains in the same polypeptide (Ingraham et al., 1990; Fortini et al., 1991; Verrijzer et al., 1992) or in other proteins (Keleher et al., 1988; Grueneberg et al., 1992). These interactions can lead to the formation of very

stable DNA complexes (see especially Fortini et al., 1991; Beachy et al., 1993). This suggests a model where homeo domain proteins bind to many inappropriate sites in the genome, but these reactions are only stabilised where there are nearby sites for interactor proteins. The implicit assumption is that homeo domain proteins in such stable complexes are functionally different from those which might bind individually. This model incorporates both explanations for the *in vitro* results described on page 7. The binding site assays cannot identify sites that are effective *in vivo* because they do not include the appropriate cofactors. Secondly, many homeo domains may bind identical sites, but the result is different depending on their ability to make protein-protein contacts.

Evidence consistent with such a model comes from studies that suggest that *in vivo*, homeo domain proteins can be found on a wide range of 'inappropriate' sites as well as their genetically expected targets (Walter et al., 1994). These experiments however, did not establish whether the observed binding was to fortuitous specific sites, or the cumulative effect of lower affinity sites across a region, or even indirect DNA association through other DNA binding proteins. With little information about the binding dynamics to these unexpected regions it is too early to conclude that they will have no regulatory significance, although that is the authors' prediction (Walter et al., 1994). The functional importance of co-factor binding can be inferred from studies in yeast where MATα2 binding specificity is quite different in haploid cells in which it binds with MCM1 to asg operators, to diploid cells in which it binds with MATa1 to hsg operator sites (Johnson, 1993). Comparable examples in higher eukaryotes are the effects of SRF on Phox1 binding specificity (Grueneberg et al., 1995) and the effects of Extradenticle (Exd) on Ubx and En specificity (reviewed by Mann, 1995). In all of these cases, the protein-protein interaction has a significant effect on the specificity and affinity of the homeo domain-DNA interaction.

Since these complexes require more bases to form an effective dual site, the potential problem of fortuitous competitor sites may be alleviated, but it should be noted

that non-specific competition will remain extreme unless the complex can be demonstrated to have a greatly increased affinity for specific over non-specific DNA. This has yet to be shown for any homeo domain complex, although it may be tentatively inferred from a dramatic increase in dissociation time for Ubx when several sites are bound cooperatively (Beachy et al., 1988).

If these examples can be generalised to account for the functional specificity of most homeo domain proteins, then the regulation of their target genes should show an obvious requirement for multiple sites and interacting protein factors. A great deal of effort has gone into studying the regulation of the few known homeo domain target genes, but the size and complexity of their promoter regions has confined analysis to a small number of enhancer elements (Jiang et al., 1991; Appel and Sakonju, 1993; Schier and Gehring, 1993; Zeng et al., 1994; McCormick et al., 1995; Sun et al., 1995). The most striking result from all of these studies is that whereas the requirement for each homeo domain site and what might bind there are open to debate, there is no doubt that multiple sites and other unknown protein factors are required to generate any given homeo domain-dependant enhancer expression pattern. Characterisation of these sites and factors is under way, but may require the extremely painstaking approach of nuclear extract cross-competitions (Yuh et al., 1994) to maximise the likelihood of identifying all the relevant binding sites.

The existence of proteins such as Exd, MCM1 and SRF, which may alter the functional specificity of homeo domain proteins, underlines the difficulty of assessing what will form a functional homeo domain site *in vivo*. One approach to this problem has been to introduce into yeast a library of random binding sites upstream of a reporter gene and to screen for those sites that make the reporter responsive to the expression of a homeo domain protein (Gross and Gruss, 1995). This screen did not isolate any sites that bound their homeo domain protein (Hoxa-7), instead it selected sites that apparently bound endogenous yeast factors. These factors could bind without Hoxa-7, but required

Hoxa-7 to activate transcription. No activation was observed if the homeo domain was mutated to prevent DNA-binding, and binding assays with nuclear extracts suggested that Hoxa-7 was bound elsewhere on the promoter (Gross and Gruss, 1995). It should be noted that promoter contained at least three optimal Hoxa-7 binding sites, yet its transcriptional response to Hoxa-7 was barely detectable in the absence of the cofactor binding site. Although this study did not formally prove that these homeo domain sites were necessary for activation, it was consistent with a model where transient homeo domain interactions with several potential binding sites affected transcription only when the binding site for an unknown cofactor was present.

If regulation by homeo domain proteins is dependant on the presence of other protein factors it is entirely possible that an optimal homeo domain binding site may be unnecessary. This has been suggested by several studies where *in vitro* defined optimal binding sites have been mutated without adverse effect on enhancer function *in vivo*, as well as the converse where lower affinity sites were absolutely required (Jiang et al., 1991; Appel and Sakonju, 1993; Zeng et al., 1994; Sun et al., 1995). This is consistent with a model whereby homeo domain specificity is primarily conferred by residues that mediate protein-protein interactions, rather than by those which contact the DNA.

Further evidence to support this model comes from parallel investigations of En binding sites *in vivo* and *in vitro* (Vincent et al., 1990; Kalionis and O'Farrell, 1993). In the first study, multimers of an En consensus binding site were placed upstream of a reporter gene that was then introduced into the fly genome. The reporter gene was not expressed if a *ftz* or *en* promoter was used and was expressed in glial cells in a *hsp70* promoter context (Vincent et al., 1990). In complete contrast to cell culture results (Jaynes, 1988), these constructs were clearly not responsive to En or Ftz or any other known homeo domain protein that binds a CAATTA site, instead they responded to an unknown factor found in glial cells that required both the CAATTA site and other unknown features specific to the *hsp70* promoter. From this we may conclude that the

functional specificity of En and Ftz is not dictated by their *in vitro* DNA binding specificity and that homeo domain site binding proteins require other sites to be effective. The second study assessed what amino acid restrictions there might be on homeo domains that bound the same site by screening a *Drosophila* cDNA expression library with the En binding site (Kalionis and O'Farrell, 1993). The remarkable result of this screen was the isolation of 15 highly diverse homeo domain proteins (not including En, Ftz or Ubx) which all preferentially bound a CAATTA site. On the assumption that not all of these proteins regulate the same target genes, it is again apparent that the diversity of their functions must be generated by the variation in residues that have no bearing on their DNA binding specificity. Consequently it seems likely that an understanding of homeo domain specificity cannot be reached by studying them in isolation, but instead will require the identification and characterisation of those protein factors with which they interact on the DNA. Proteins such as Exd and SRF provide some explanation for the specificity of the Ubx and Paired class homeo domains with which they interact. It remains to be seen what factors may account for the specificity of the remaining classes of homeo domains.

In addition to the varied homeo domain proteins isolated by the En site binding screen, one potential homeo domain interactor was also found (Kalionis and O'Farrell, 1993). This clone, bk60, produced a product that bound several variant sites with similar preferences to those seen for Q_{50} homeo domain proteins, but preliminary sequencing did not detect a homeo domain. *In situ* hybridisation of this clone to mRNA in *Drosophila* embryos revealed a dynamic expression pattern including ubiquitous maternal deposition followed by progressively more restricted expression in a gap gene pattern, the mesoderm and finally a variety of apparently unrelated organs in the mature embryo (B. Kalionis, personal communication). The DNA binding ability and expression pattern of this clone suggested that it came from a gene which was likely to be involved in transcriptional control and tissue differentiation. Its site binding preferences suggested that this protein

could bind the same targets as homeo domain proteins and thus had the obvious potential to interact with them either by competition or cooperation.

Dead ringer

This thesis describes the characterisation of the *Drosophila* gene corresponding to bk60. This gene has been named *dead ringer* (*dri*) to reflect its mutant phenotype, expression pattern and sequence conservation (see Results). Specific aims were firstly to identify its novel DNA binding domain then more closely define its *in vitro* specificity to assess the *bona fides* of its interaction with homeo domain sites. Secondly, a more detailed examination of its developmental expression pattern and mutant phenotype analysis were undertaken to assess the requirements for *dri* function *in vivo* and to provide further directions for enquiry. Experiments designed to address these issues are described in three Results chapters, followed by a discussion of the outcome of this work and its implications for further study.

Materials

"One long bent thing with a sort of ... lump on the end"

- The International Christmas Pudding The Goon Show

Abbreviations and Acronyms

AbdB:	Abdominal-B (HD protein)
ADH2	Alcohol Dehydrogenase II (yeast enzyme)
Antp:	Antennapedia (HD protein)
Bcd:	Bicoid (HD protein)
bk60:	A λgt11 cDNA clone from which Dri was isolated
Bright:	B-cell restricted Immunoglobulin-H transcription factor (Dri homologous)
Brm	Brahma (trx group protein)
Bp	Base-pair
BX-C:	Bithorax complex
cDNA:	complementary DNA (from RNA)
CNS:	Central nervous system
Dfd:	Deformed (HD protein)
DNA:	Deoxyribose nucleic acid
Dri:	Dead ringer
En:	Engrailed (HD protein)
Eve:	Even-skipped (HD protein)
Exd:	Extradenticle (HD protein)
Ftz:	Fushi Tarazu (HD protein)
GCG:	Genetics Computer Group (Chicago)
GST:	Glutathione-s-Transferase
Hb	Hunchback (Zn finger protein)
HD:	homeo domain
Hox:	mammalian homologues of *Drosophila* homeotic genes
Hsp70:	Heat shock protein (70KD)
HTH:	Helix-Turn-Helix DNA binding domain
Kb	Kilobase-pair
LacI:	Lactose operon repressor (HTH protein)
LacZ:	β-Galactosidase
MAR	Matrix attachment region
MRF:	Modulator response factor (Dri homologous)
mRNA:	messenger ribo nucleic acid

NP:	Near palindromic En binding site (GATCAATTAAAT)
ORF:	Open reading frame
p18-13	p18-13(1)B, a homeotic reponse element-LacZ insertion.
PAGE:	Polyacrylamide gel electrophoresis
PCR:	Polymerase chain reaction (Proliferation of Complete Rubbish?)
PNS:	Peripheral nervous system
Pfu	Plaque forming units (ie phage)
Q_{50}:	Glutamine residue in the third helix (position 50) in a homeo domain
Rb	Retinoblastoma protein
RBP:	Retinoblastoma (protein) binding protein (Dri homologous)
RF:	Reading frame
RT	Room temperature (21-27°C)
Scr:	Sex combs reduced (HD protein)
Snr1	SNF5-related1 protein
SRF:	Serum response factor (MADS domain protein)
TrpR:	Tryptophane operon repressor
Ubx:	Ultrabithorax (HD protein)
2D-IEF SDS PAGE:	Two-dimensional isoelectric focussing Sodium Dodecy Sulfonate Polyacrylamide gel electrophoresis (my favourite acronym)

Bacterial strains

All subcloning and protein expression studies were carried out using DH5α (Hanahan, 1983). Phage were propagated in Y1090 (Clontech)

Chemicals

All chemicals were of analytical grade or the highest purity available. Except where indicated, all solid chemicals were obtained from Sigma and all liquid chemicals from May and Baker Ltd.

Acrylamide	Bio-Rad
Agarose	Seakem
BCIG	Boehringer Mannheim
BCIP	Boehringer Mannheim
digoxigenin-11-dUTP	Boehringer Mannheim
dNTPs	Boehringer Mannheim
NBT	Boehringer Mannheim
nitrocellulose	Schleicher and Schuell

Phenol	BDH Labs, Aust.
Sepharose CL6b	Pharmacia
TEMED	Kodak

Fly strains

Unless otherwise noted, all fly strains are as described by (Lindsley and Zimm, 1992) and were obtained from the Indiana Stock Centre, Bloomington.

Hindgut expressing enhancer traps:

pA350.1M2 CyO	Genes&Dev. 3: 1288
pA308.1M3 ry	Genes&Dev. 3: 1288
p18-13	a dpp45 insertion strain (Manak et al., 1994)
p18-11	"
p18-7	"
p90	Random 2nd chromosome enhancer trap insertions screened by S. Parkhurst et al., Fred Hutchinson Centre for Cancer Research, Seattle
p355	"
p716	"
p1647	"

Balancer chromosomes used:

CyO	
SM6a	
CyO *wg-lacZ*	N. Patel, Carnegie Inst. of Washington, Baltimore
CyO *ftz-lacZ*	"

Mutants around 59F:

l(2)02535 (dri^{P1})	
l(2)05096 (dri^{P2})	
Df(2R)bw^{5}	Distal break at 59F1-2
Df(2R)bw^{s46}	Deleted across 59E-60A
Df(2R)*tid*	(Kurzig-Dumke et al., 1992), proximal break at 59F1
Df(2R)*x32*	(Kurzig-Dumke et al., 1992), proximal break at 59F2-3

Libraries

Four plasmid vector cDNA libraries constructed by N. Brown were screened. These were made from 0-4, 4-8, 8-12 and 12-24hr *Drosophila* embryonic RNA using a poly-T primer for first strand synthesis (Brown and Kafatos, 1988). The complexity of these libraries was approximately $1x10^6$. The 9-12hr cDNA library in λgt11 was constructed in the laboratory of C. Goodman (Zinn et al., 1988). The 0-18 hr cDNA library in λgt11 was purchased from Clontech, product number IL1010b. This library had a complexity of $2x10^6$, and had been size selected for inserts between 0.6 and 5kb that were generated with random as well as poly-T primers.

Media and Buffers

All buffers and media were made with deionised water and sterilized by autoclaving, except heat labile reagents which were filter sterilised. All bacterial strains were grown in L-broth or on L-agar plates (1.5% bacto-agar):

L-Broth:	1%	amine A
	0.5%	yeast extract
	1%	NaCl, pH7

When required for selection, ampicillin was added to a final concentration of 100μg/ml. All Drosophila strains were grown on fly media:

Fly media:	10%	treacle
	20%	yeast
	1%	agar
	10%	polenta
	2.5%	tegosept
	1.5%	propionic acid

Commonly used buffers were:

PBS:	7.5 mM Na_2HPO_4
	2.5 mM NaH_2PO_4
	145 mM NaCl
PSB:	10 mM Tris-Cl pH 7.4
	10 mM NaCl
	100 mM $MgCl_2$
SSC:	150 mM NaCl
	15 mM Na citrate
TE:	10 mM Tris-Cl pH 7.4

0.1 mM EDTA

TAE: 40 mM Tris-acetate pH 8.2
1 mM EDTA

TBE: 50 mM Tris-borate pH 7.4
1 mM EDTA

10x Agarose gel loading buffer: 50% glycerol
10mM EDTA
0.2% bromophenol blue
0.1% xylene cyanol

Molecular Weight Markers

DNA: λ DNA digested with BstEII and SalI.

RNA: 0.24 - 9.5 kb RNA ladder (Gibco BRL)

Protein: Prestained high molecular weight markers (Gibco BRL).

Oligos

Capitals signify sequences found in *dead ringer*. Restriction sites are underlined.

Dri 1 5': AAAGGCCACCGAGTTGT
Dri 2 5': CGAATGTTGCGGTTGAT
Dri 3 3': AGGTGGTTCGTTTTTG
Dri F: ACAAGGAAAGGAATACC
PCR1 (gt11): caagcttcggtggcgacgactcctgg
RACE: cgggatccccccccccccc
Check: TTTCCACTGCCCCACAACTC
Inside: GATTCTTTTTCTCGCACTCG
RT: GCTCCATCCATTATTCGTCT
Dri A: GATCTCCTGCTTGACCA
Dri B: TGCTGCGGCGAGGTGTG
Dri C: CGTGGACCAGGATGACA
Drif/lNco5': catgccATGgAACTGCGAGTGCACCC
Drif/lBam5': agggatccaATGCAACTGCGAGTGCACC
Drif/lBam3': cgggatCCGGTTCGTCGCGTTATCCTTC
5'ARID: agggatcCGCAGCAGAATAATGGATGGA
3'ARID: ggaattCGTCATCGGCATCATCTGGTTGTG
XARID5': gggaagcagtTCGCAGCAGAATAATGGA
XARID3': gctctagatcaCGTCATCGGCATCTGGTT
5'ARID#2: agaaTTCCTGGACGACTTG
3'ARID#2: gctctagactaCTCGTACGGGTACAG
5'inside: ggaattcATAATGGCCAAATCG
3'inside: gctctagatcaGGTGATGCTGGAGGG

$NP_3$5': tcaattaaatgatcaattaaatgatcaattaaatga
$NP_3$3': tcatttaattgatcatttaattgatcatttaattga
$TTA_9$5': ttattattattattattattattatta
$TTA_9$3': taataataataataataataataataa

ABS5': ccgattaatcccccgattaatcccccgattaatccc
ABS3': gggattaatcgggggattaatcgggggattaatcgg
MBS5': ccgattgatcccccgattgatcccccgattgatccc
MBS3': gggatcaatcgggggatcaatcgggggatcaatcgg

R20: cgggatccgtgactgaggnnnnnnnnnnnnnnnnnnnnttgatgccgaggatcccg
R20start: cgggatccgtgactgagg
R20end: cgggatcctcggcatcaa

Plasmid Vectors

pBluescript (Stratagene)

pGEX1,2 and 3 (Smith and Johnson, 1988)

pMALc-2 (New England Biolabs)

Radiochemicals

α-^{32}P-dATP (3000Ci/mmole)	BRESATEC
α-^{35}S-dATP (1500Ci/mmole)	"
γ-^{32}P-dATP (4000Ci/mmole)	"

Methods:

To make Gosky Patties:

Take a Pig, three or four years of age, and tie him by the off hind leg to a post. Place 5lb of currants, 3 of sugar, 2 pecks of peas, 18 roast chestnuts, a candle and 6 bushels of turnips within his reach; if he eats these, constantly provide him with more.

Then procure some cream, some slices of Cheshire cheese, four quires of foolscap paper and a packet of black pins. Work the whole into a paste and spread it out to dry on a sheet of clean brown waterproof linen.

When the paste is perfectly dry, but not before, proceed to beat the Pig violently, with the handle of a large broom. If he squeals, beat him again.

Visit the paste and beat the Pig alternately for some days, and ascertain if at the end of that period the whole is about to turn into Gosky Patties.

If it does not then, it never will; and in that case the Pig may be let loose, and the whole process may be considered as finished.

- *Nonsense Cookery* Edward Lear.

Antibody stains & dissection.

Protein expression was detected in fixed embryos using polyclonal antibodies and immunostaining. About 50µl of embryos were placed in PBT (PBS, 1% Triton X100,

0.2% BSA), then washed 2x5' in PBT then 2x20' in PBT, rocking. Embryos were then placed in 100 µl PBTG (PBT+ 5% Goat serum) 30', then 100µl 1° antibody diluted in PBTG was added for 2hr RT or o/n at 4°C. Antibody was washed off in PBT as above, then 2° antibody was added in 100µl PBTG for 2hrs RT, then washed in PBT as above. Antibody was detected by adding 10µl of DAB stock (10mg/ml Diaminobenzidine in PBT) and 2µl of 8% NiCl in 200µl PBT for 10'. Then 3µl of 3% H_2O_2 was added and allowed to react for up to 15', stopping by washing with PBT. Embryos were cleared by layering onto 70% glycerol in PBS. In some cases embryos were dehydrated in several washes of ethanol before layering onto methyl salicylate (no good for X-gal or AP). Intact embryonic hindguts were dissected under a 50x objective in 70% glycerol/PBS after immunostaining, then manipulated, covered and photographed at 1000x.

Brownies

Line baking tin with buttered baking paper. Preheat oven to 180°C. Melt 125g unsalted butter over low heat. Remove from heat and stir in 1 cup caster sugar, 1tsp vanilla essence, 1/2 cup cocoa and 4 beaten eggs. Mix in 1/3 cup flour and 1/4 cup ground pecans, then stir through 200g chopped quality dark chocolate and 1/2 cup chopped pecan nuts. Pour mixture into tin and bake for ~40min, checking at 30min. Remove from oven and cut into squares. Serve with cream and ice-cream. Yum.

Colony Cracking - screening for recombinant clones

Colonies resulting from transformation of a plasmid ligation were routinely screened for the presence of the correctly sized insert by colony cracking. A yellow tip was used to touch each colony, subculture each onto an appropriate plate and was then placed in 15µl of cracking solution (25µl 2mM NaOH, 50µl 10% SDS, 10µl 500mM EDTA, 125µl 80% glycerol, 785 µl H_2O and a touch of bromophenol blue). Each tip was swirled, then the tube and tip incubated 65°C 15'. The solution was expelled from each tip, which was swirled then discarded, incubating a further 10' 65°C. Each tube was microfuged briefly then the samples loaded on a 1% agarose gel with exposed wells,

run in for 15' at 30V, then covered with TAE and run at 90V for 30' to resolve the plasmid sizes.

DNA preparations: plasmid, phage, and fly genomic

Plasmid minipreps were carried out from 3ml overnight cultures using the boiled lysis method (Murphy and Kavanagh, 1988), resuspending in 20µl TE after isopropanol precipitation. Larger scale plasmid DNA isolation was carried out using Qiagen midipreps according to the manufacturers specifications.

Phage DNA was isolated by using a fresh overnight culture of Y1090r⁻ to inoculate 50ml L-Broth at 1/50 and growing at 37°C to an OD_{600} of 0.4-0.8. To this was added ~10^{10} pfu of λgt11 (or 500µl of a 5ml small plate lysate), and the culture was grown until lysed or for up to 6 hours. At this stage 2ml $CHCl_3$ was added and vigorously shaken, then left overnight at 4°C if necessary. Debris was removed by spinning 10' at 7000g then DNAse and RNAse were added to 10µg/ml and incubated 37°C or rt for 30'. Debris was removed by spinning at 12000g 15', then to the supernatant was added PEG8000 to 5% and NaCl to 0.75mM (2.5g each). After >2hrs at 4°C, phage were spun down for 15' at 12000g and resuspended in 500µl PSB, carefully removing the PEG. 500µl $CHCl_3$ was used to extract remaining PEG, then 20µl 500mM EDTA, 10µl 10% SDS and 5µl 5mg/mg proteinase K were added and incubated at 65°C for 30'. Two phenol/chloroform extractions were carried out, then the DNA was precipitated with 50µl 3.5MKOAc pH5.5 and 1ml Ethanol and microfuged immediately for 15'. The pellet was washed once with 70% ethanol then resuspended in 100µl TE.

Embryo fixing

Embyos were washed off collection plates with water into baskets then dechorionated in 50% bleach for 2-3', washed well in water, then put into scint vials containing 5ml heptane, 3.5ml water, 0.5ml 10xPBS, 1ml 20% formaldehyde then shaken 10-15'. The aqueous (bottom) layer was removed and replaced with 5ml

methanol then the vial was vigorously shaken 1' to devitellinize. Embryos were removed and washed several times in methanol then ethanol and stored at -20°C.

Gel Retardations

Electrophoretic Mobility Shift Assays were performed by incubating 0.1ng of ^{32}P end-labelled double-stranded trimer of the consensus Engrailed binding site TCAATTAAATGA (NP_3) with approximately 10ng of fusion protein in 20µl of a buffer containing 10mM Tris pH7.5, 1mMEDTA, 100mM KCl, 0.1mM dithiothreitol, 5% glycerol, 50µg/ml BSA and 100ng of herring sperm DNA as non-specific competitor. Specific competitor oligonucleotide was added as indicated, using 1, 10 or 100ng of NP_3 or TAA_9. Incubation was for 20 minutes at 25°C, followed by electrophoresis on a 6% polyacrylamide 10% glycerol gel in 0.5x TBE (Tris Borate EDTA) buffer. To purify the fusion protein, glutathione-agarose purification was performed as described in Smith and Johnson (1988) then, with the fusion protein still bound to the beads, binding reactions to labelled NP_3 were carried out as described above. Unbound probe was removed by pelleting the beads, removing the liquid then resuspending the beads in buffer. After two washes, the amount of probe still bound was detected by scintillation counting. The results obtained using glutathione purified proteins in this method were consistent with those obtained by gel retardation of crude lysate.

Imprecise excision crosses and P-element reversion

To generate deletions from the P-element insertion in *dri*^{P1}, the following crosses were used:

$$\frac{dri^{P1}}{CyO}\,;\,ry \quad X \quad \frac{Sp}{CyO}\,;\,\frac{\Delta 2\text{-}3\ Sb}{TM3}$$

$$\Downarrow$$

Select for transposase

$$\frac{dri^{P1}}{Sp/CyO}\,;\,\frac{\Delta 2\text{-}3\ Sb}{ry} \quad X \quad \frac{dri^{P1}}{CyO}\,;\,ry$$

$$\Downarrow$$

Select for loss of P element (*rosy*$^+$)
-gives reversion frequency.

$\frac{dri^X ; ry}{CyO}$ X $\frac{dri^{P1} ; ry}{CyO}$

⇓

Select for dri^X/dri^{P1} lethality
-gives deletion frequency.

$\frac{dri^X ; ry}{CyO}$

In situ hybridisation

In situ hybridisation using a digoxygenin-11-dUTP labelled probe was carried out as described in (Patel and Goodman, 1992). using a probe containing *dri* sequence from position 500 to 2900.

Library screens

Phage libraries were plated on 14mm plates at a density of 3-5x10^4 pfu/plate, grown, transferred to duplicate sets of Plaquescreen (NEN) filters, denatured and hybridised with a radiolabelled probe according to the specifications of the manufacturer (Clontech). Plasmid libraries were plated at a density of approximately 1.5x10^4 colonies/14mm plate. They were transferred to duplicate nitrocellulose filters, lysed and denatured as described by (Brown and Kafatos, 1988). Probes were labelled by random-primed Klenow catalysed α-^{32}P-dATP incorporation using a Megaprime kit (Amersham). Unincorporated mucleotides were removed by Sepharose CL6b spun column chromatography as described by (Murphy and Kavanagh, 1988). Filters were typically washed with two changes of 2xSSC at RT then 0.1xSSC, 0.1%SDS at 65°C for 30'. Filters were then autoradiographed with X-Omat AR film (Kodak) and calcium tungstate intensifying screen at -80°C or using Fuji phosphorimager cassettes.

Northern Analysis

RNA was extracted from staged embryos by homogenization in 6M guanidine-HCl, 0.1M NaAc and pelleting at 37krpm in a SW41 rotor for 16hr through a 4.8M CsCl pad (MacDonald et al., 1987). Pellets were resuspended and ethanol precipitated twice before resuspension in water. 20µg of total RNA was fractionated on low formaldehyde, 1.2% agarose gels (Ausubel et al., 1987). RNA was immobilised by blotting onto

Nytran-N (Schleicher and Schuell) with HETS buffer (Cinna/Biotec). Transfer efficiency was assayed by staining the membrane with methylene blue. Filters were hybridised with radiolabelled probe, washed and autoradiographed as for library screens.

PCR Standard Reaction

DNA amplification was carried out using a FTS-1S capillary thermal cycler (Corbett Research). Standard reaction conditions were 50ng each primer, 1µl 2mM dNTPs, 2U Amplitaq Polymerase (Amersham), 2µl 10x PCR Buffer (100mM Tris-Cl pH 8.3, 500mM KCl, 15mM $MgCl_2$) and up to 10ng template DNA in a 20µl reaction which was denatured at 94°C for 15", then given 25 cycles of 94° 5", 55° 10", 72° 10", finishing with 72° 1'.

Photography

Fluorescent and brightfield microscopy were performed on a Zeiss Axiophot microscope equipped for Nomarski and epifluorescence. Objectives used were Plan-Neofluar 20x/0.5, 40x/0.75 and 100x/1.3 oil immersion. Photographs were taken with a Zeiss Microphot system and recorded on Ektachrome 160T film (Kodak). Slides were scanned with a Kodak RFS 2035 Film Scanner at >500dpi. Adobe Photoshop 3.0.4 was used for image preparation. Colour prints were obtained using a Kodak XLT7720 Digital Continuous Tone Printer.

Protein expression using pGex and pMal.

Proteins were expressed and crude lysates extracted essentially as described by (Smith, 1988) for pGEX fusions or according to the specifications of the supplier (NEB) for pMAL fusions. Protein production was analysed by SDS-polyacrylamide gel electrophoresis to ensure that equivalent amounts of each protein were used in subsequent assays.

Regulatory considerations

All manipulations involving recombinant DNA were carried out in accordance with the regulations and with the approval of the Genetic Manipulation Advisory Commitee and the University Council of the University of Adelaide.

Sequence analysis

Sequence analysis and database searches were carried out using the facilities provided by the Australian National Genome Information Service (ANGIS) and those provided by the Baylor College of Medicine at http://kiwi.imgen.bcm.tmc.edu. Codon preferences were compared to those compiled from a file of highly expressed Drosophila genes (M. Ashburner, Cambridge) using the program CODONPREFERENCES available through ANGIS. Protein structural predictions were generated by the PHD program described by (Rost and Sandler, 1994) as well as the algorithm of Chou and Fasman (Chou and Fasman, 1978) available though ANGIS. PCR reactions and oligonucleotide design were optimised using the OLIGO 4.04 program (National Biosciences). Routine restriction enzyme diagnostics and contig assembly were performed using Genejockey 1.2 (Biosoft).

Sequencing

Approximately 9µg of plasmid DNA was RNAse treated, alkali denatured and purified on a Sepharose CL-6b spun column as described by (Murphy and Kavanagh, 1988). 3µg of this was annealed to 10ng primer at 37°C for 1hr. Sequencing was carried out by the dideoxy method (Sanger et al., 1977) using α-^{35}S-dATP and a Sequenase Kit (USB).

Site selection

Dri target sequences were isolated as described by (Wilson et al., 1993), with the following randomer: CGGGATCCGTGACTGAGGN$_{20}$TTGATGCCGAGGATCCCG with amplification primers made to the first (top strand) and last (bottom strand) 18 bases. Binding was carried out as described for Gel Retardations, with the addition of 2μg/ml herring sperm DNA as non-specific competitor. Amplification was carried out with an annealing temperature of 50°C for twenty cycles. Following eight cycles of selection and amplification, the products were digested with BamHI, cloned into pBluescript (Stratagene) and sequenced. Of the oligonucleotides selected, only one sequence was found twice, all others were unique. Three sequences were isolated that could not be aligned with the consensus nor with each other, indicating a low background and high diversity in the final pool of selected oligonucleotides.

Southern Analysis

After visualisation with ethidium bromide, agarose gels were soaked in 0.25M HCl for 10', then 0.5M NaOH, 1.5M NaCl for 10' then 1M Tris-Cl pH7.4, 1.5M NaCl for 15'. The gel was then placed on three sheets of 3MM paper soaked in 20XSSC; on top of it were laid a wet sheet of nitrocellulose, three sheets of 3mm paper, a 5cm stack of paper towels and a 0.5kg weight, in that order. Transfer was carried out for >6hr, then the DNA was crosslinked onto the filter using a UV crosslinker (Stratagene) and was hybridised with a radiolabeled probe as for library screens.

Transformation of plasmids

Cells competent for transformation were prepared by inoculating 500ml L-Broth 1/100 with a fresh overnight culture of DH5α and growing at 37°C with vigorous shaking to an OD_{600} of 0.8. The culture was chilled on ice 15' then centrifuged at 4000g for 15'. All subsequent steps were chilled. The supernatant was removed and the pellet resuspended in 500ml cold water before centrifuging as above. This process of washing was repeated using less liquid and slightly higher speeds, twice, then the pellet was resuspended in 10ml cold 10% glycerol. This was spun down and finally resuspended in

1ml cold 10% glycerol before dividing into 50µl aliquots, snap freezing in a dry ice/ethanol bath and storing at -80°C. Cells were transformed using a BioRad electroporator with 2mm wide cuvettes at 2.5kV, 200Ω, 25µF, rescued in 1ml SOC at 37°C for 30' then plated on appropriate selective media.

Chapter 3: Characterisation of the *dri* transcript

And if you cannot work with love but only with distaste it is better that you should leave your work and sit at the gate of the temple and take alms of those who work with joy.

- *The Prophet* Kahlil Gilbran.

Introduction:

To be able to explain the surprising DNA binding characteristics of clone bk60, it was necessary to identify a DNA binding domain[1], confirm its *in vitro* activity and to assess its role in the *in vivo* function of the protein. Experimental work toward achieving these three objectives is presented in the three Results chapters. This first chapter addresses the need for preliminary characterisation of the message from which bk60 was generated. The 2.4kb bk60 cDNA clone was isolated from a *Drosophila* 9-12hr embryonic λgt11 library that expressed each clone as a *lacZ* fusion (Kalionis and O'Farrell, 1993). Preliminary sequencing had revealed a short open reading frame of 116 amino acids, with no evidence of a homeo domain, and a very long 3' untranslated region (B. Kalionis, pers. comm.). The 5' end of the coding region was absent since all clones were C-terminal fusions with *lacZ*. The first objective in the characterisation of the *dead ringer* gene was to isolate and sequence the complete coding region. This would allow identification of any evolutionarily conserved regions that might suggest functional domains. A complete sequence would also facilitate bacterial production of full-length or partial proteins for antibody generation and DNA binding assays. The first approach taken was to screen *Drosophila* embryonic cDNA libraries using parts of bk60 as probes.

[1]Domain is used throughout to describe the shortest known polypeptide which can confer an assayable function. The word 'motif' is used where the function of a highly conserved region is unknown.

Characterisation of *dri* cDNAs.

The libraries initially available were 0-4hr, 8-12hr, 12-24hr embryonic cDNA plasmid libraries, and the 9-12hr λgt11 library from which bk60 had been isolated (see Materials). These libraries were screened using the oligo-labelled insert from clone bk60 (a 2.4kb EcoRI fragment). The results from these experiments are presented in Figure 3.1.

Figure 3.1

Library	Clones Screened	3rd Round +ves	Dri clones
0-4hr plasmid	150,000	9	2
8-12hr plasmid	75,000	2	0
12-24hr plasmid	150,000	6	0
9-12hr phage	250,000	4	2
0-18hr phage	150,000	10	8

A surprisingly high background of hybridisation was encountered when using the entire clone as a probe. This led to the isolation of several clones that appeared to hybridise through three rounds of screening but were not highly homologous to the bk60 sequence as shown by Southern blots and sequence analysis (results not shown). Such clones frequently contained tandem repeats of CAG or related trinucleotides, which were also present in the bk60 sequence in two areas (see Figure 3.2A). These are known as *opa* repeats and although their function is unknown, they have been found in many developmental regulators (Wharton et al., 1985). Hybridisation of the 2.4kb bk60 clone to a Southern Blot of *Drosophila* genomic DNA at low stringency revealed many bands, consistent with repetitive DNA hybridisation (Figure 3.2B). To avoid spurious hybridisation by these repetitive regions, probes were then generated which contained only the repeat-free 5' or 3' ends of bk60. Several clones that extended further 3' were obtained, and sequencing of these revealed a common 3' end of transcription with a

Figure 3.2 The *dead ringer* cDNA

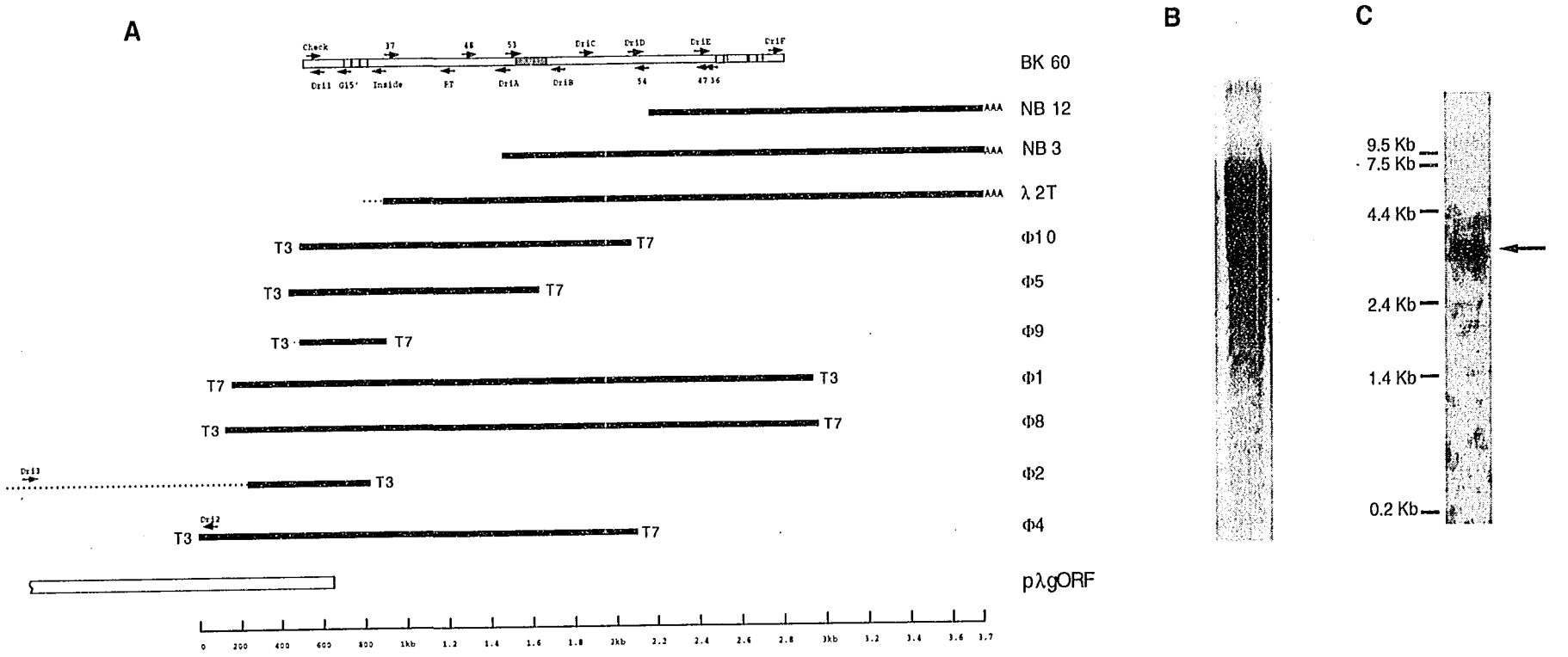


Figure 3.2 **A.** The position of each isolated*dri* cDNA is marked relative to the consensus 3.7 kb contig. The location of each primer used in sequencing is indicated by an arrowhead (not to scale). Chimeric ends isolated with*dri* cDNAs are indicated by dotted line extensions. The original cDNA BK60 is indicated by an open box with shading to show the extent of repetitive sequences. Vertical bars represent tandem repeats of the trinucleotide GGN, stippling represents repeats of ACG (opa repeats). The 3' end od the genomic clone pλgORF is shown as a broken box. **B.** Southern blot of*Drosophila* genomic DNA probed with the BK60 cDNA and washed at reduced stringency (2xSSC, 45°C). Many bands are apparent, consistent with hybridisation of repetitive sequences. **C.** Northern Blot of 0-4hr embryonic RNA probed with non-repetitive*dri* sequences. The hybridising band corresponds to a transcript of approximately 3.7 kb.

putative polyadenylation signal (Figure 3.2). Repeated library screening with a probe containing the 5' most 180 bases of bk60 yielded clones that did not extend the open reading frame further 5'. Alternative methods were then pursued to obtain the 5' end of the transcript.

A genomic clone hybridising to the 5' end of bk60 was kindly provided by R.D. Kortschak, and this clone (pλgORF) was mapped and subcloned as a source of sequences 5' to the available cDNA clones (Figure 3.2A). Sequencing into subclones revealed an extension of the open reading frame for at least 100 bases (results not shown). The *Drosophila* splicing acceptor site has a broad consensus (essentially ag|GG or ag|GT) so intron positions, and hence the open reading frame, could not be confidently predicted in this region (Breathnach and Chambon, 1981). Efforts to confirm the 5' end of the ORF included PCR amplification from DNA derived from a 0-4hr cDNA library, using one vector and one *dri* specific primer, and rapid amplification of cDNA ends (RACE) PCR using embryonic RNA as a template (see Methods). While these experiments were in progress they became unnecessary due to the isolation of clones from a random primed 0-18hr embryonic cDNA library in λgt11.

It was possible that earlier library screening attempts to obtain the 5' end had been hindered by the length of the transcript. Any useful clone from these poly-T primed libraries required successful reverse transcription of 3kb from the 3' end. Unlike the other libraries, the newly obtained 0-18hr library contained random primed inserts, so it was anticipated that it would contain more, although shorter, clones covering the 5' end. A probe was generated from a ClaI/ApaI fragment from pλgORF that contained the 5' end of bk60 and 0.5kb of genomic sequence extending further 5'. This probe hybridised to many potential clones in the 0-18hr library of which 10 were selected for further study. Restriction mapping and end sequencing of these clones generated the alignment shown in Figure 3.2A. This 'contig' of cDNA clones spanned 3.7kb with an alternative 5' end (Φ2) extending it to over 4kb. It was unclear whether the two clones with divergent 5'

ends (2T and Φ2) represented splicing variants of *dri*, or artifactual chimeric inserts generated during construction of the libraries. Neither contain consensus splice junctions (results not shown), so the latter interpretation seems probable. Hybridisation of *dri* sequences to 2-4hr embryonic RNA on a Northern blot revealed a band at 3.7kb, consistent with the contig of Φ4 and NB3 corresponding to the primary *dri* transcript (Figure 3.2C).

Open Reading Frames in *dri*

At this stage, comprehensive sequencing across the entire contiguous cDNA was undertaken to clearly define the open reading frame. A series of 18 *dri* specific primers were synthesised for this purpose (Figure 3.2A) as well as the two vector end primers to generate the sequence of both strands across the 3.7kb contig (Figure 3.3). In this process a number of frame shifts were detected relative to the ORF detected by Kalionis and O'Farrell (1993). This resulted in an extension of their predicted coding frame (Reading Frame 3) from 116 to 176 codons. More significantly, however, a new, very large ORF of 901 codons was detected in the frame -1 relative to the predicted frame (RF2). Most translation would be expected to start where indicated for RF2 (Figure 3.3) since the in frame ATG 24bp upstream has a very poor fit to the *Drosophila* consensus start of translation (Cavener, 1987). It was necessary to determine whether either or both ORFs were translated *in vivo*, and to determine which of them generated the DNA binding activity observed for the product of bk60.

Several reasons existed for expecting that the large ORF in RF2 was the frame expressed *in vivo*. In the process of sequencing clones from different libraries, several polymorphisms were detected that fell within both ORFs. Numbering from the start of bk60 these were:

Figure 3.3 Dead ringer cDNA sequence

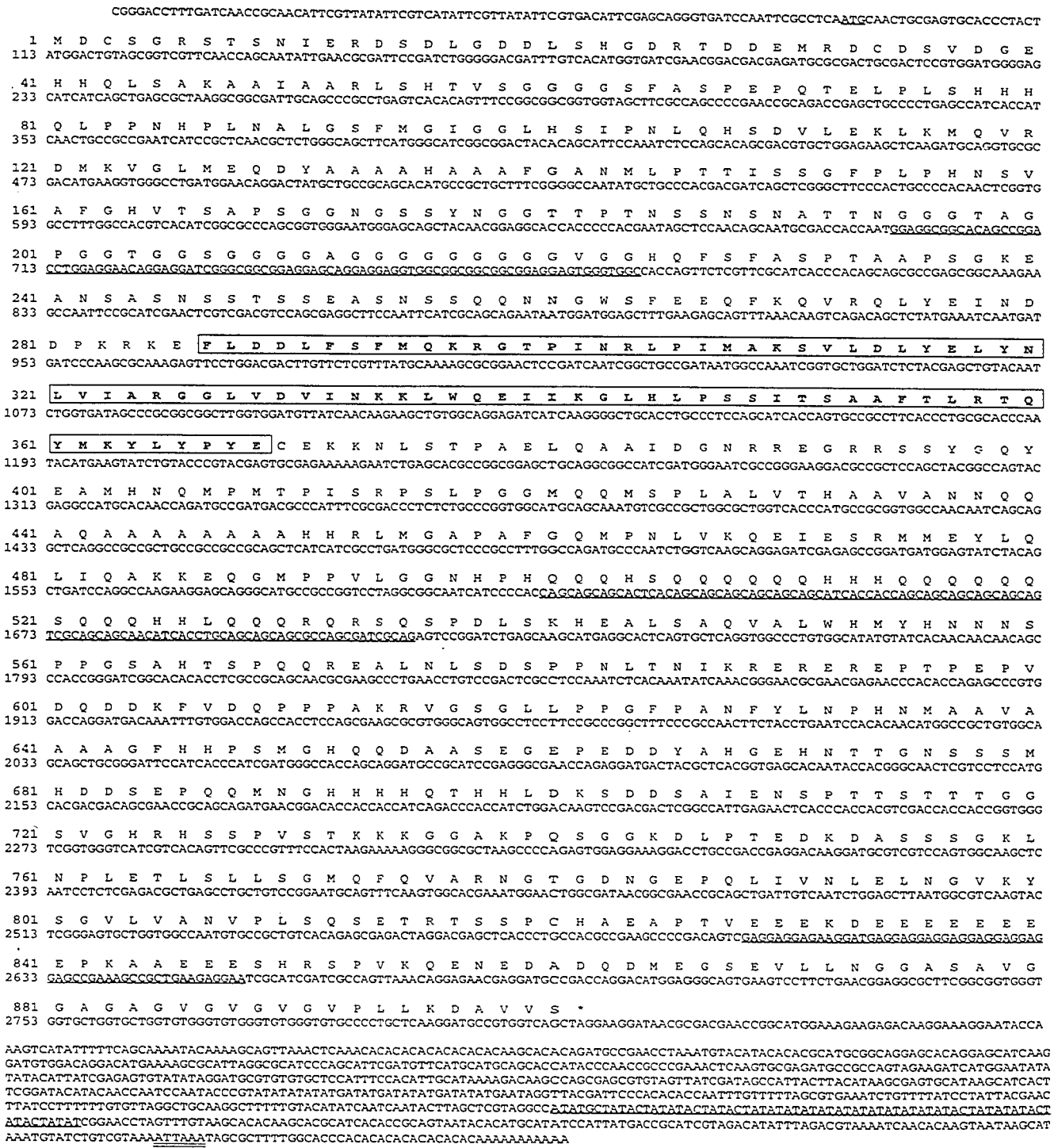


Figure 3.3 Contiguous nucleotide sequence from overlapping cDNA clones including the complete dead ringer open reading frame and 3' untranslated region, with derived amino acid sequence for reading frame two written in standard one letter code. The alternative start of translation in this frame is underlined. The boxed amino acids indicate the most widely conserved motif. Highly repetitive regions in the nucleotide sequence are underlined. A putative polyadenylation signal (ATTAAA) is double underlined.

9-12hr	35TTTCC	110AATGGA	131GAACAG
0-18hr	C	C	T
RF2	F→F	N→N	N→N
RF3	F→S	M→T	T→I

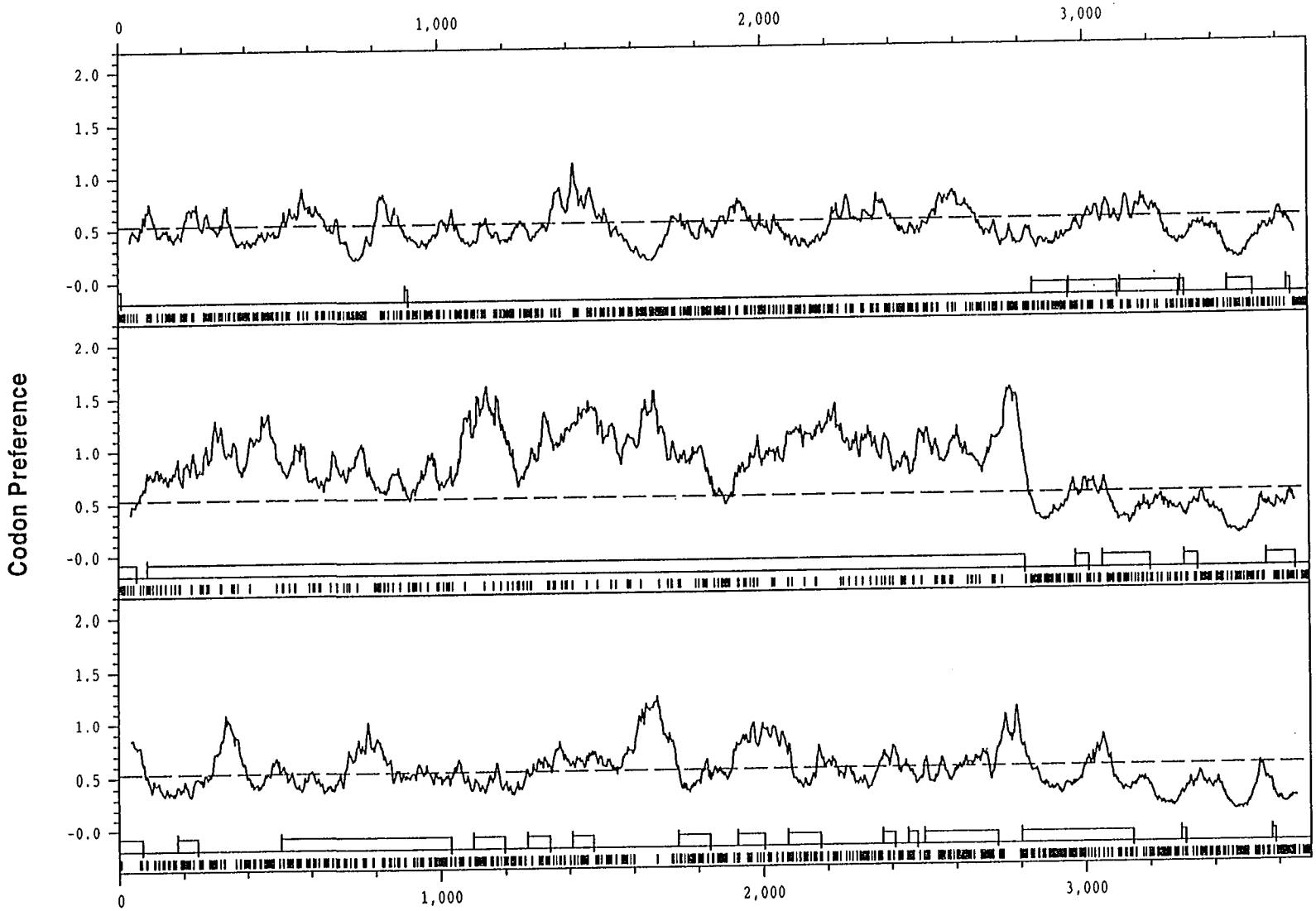
As can be seen, these are all silent polymorphisms in RF2, whereas in RF 3 they are all non-conservative substitutions with the potential to affect protein function. This suggests that RF2 is the correct ORF, although these variations could simply represent random errors introduced during reverse transcription when the libraries were constructed. The codon usage of each frame was then compared with that of highly expressed *Drosophila* genes using the GCG program CODONPREFERENCES (see Methods). This analysis showed that RF2 contained far fewer rare codons than either alternative frame (Figure 3.4). The average coding probability for RF2 was estimated at 0.74, compared with 0.47 and 0.5 for the alternative reading frame and 0.53 for a random sequence of this length. Although this did not rule out a function for RF3, it strongly suggested that the long ORF in RF2 was translated *in vivo*.

To assess this possibility, embryos were immunostained with antibodies raised against a fusion protein expressing RF2 (kindly provided by R.D. Kortschak). These stainings indicated that this ORF was expressed in *Drosophila* in a pattern closely related to that of *dri* mRNA (see Chapter 5). In contrast, antibodies raised against a RF3 fusion protein gave no consistent staining in embryos (B. Kalionis, pers. comm). These results indicated that the long ORF in RF2 was likely to code for the protein produced by *dri*.

The clone bk60 was isolated as a DNA binding fusion protein, and sequencing had indicated that the fusion would express RF3 (Kalionis and O'Farrell, 1993). Since that frame did not seem to be expressed in embryos, either the DNA binding activity was artifactual or there had been a frame assignment error. The frame had been assigned by sequencing the ends of an EcoRI subclone of the bk60 phage clone. It was possible that the phage clone to which the DNA binding function had been assigned contained more

Figure 3.4 Usage of rare codons and predicted likelihood of expression in each reading frame of the *dri* cDNA. The three X axes represent the contiguous cDNA sequence notionally translated in the three possible reading frames. Along these axes open reading frames are marked by boxes and the incidence of codons rarely used in highly expressed *Drosophila* genes is shown by vertical dashes. The codon usage was assessed in windows of 25 codons, then for each window, its usage was compared with that compiled from highly expressed genes to give a measure of the bias toward preferentially expressed codons. This bias is plotted as a continuous function across the cDNA with 0.5 corresponding to the average expected for random DNA. Comparison of the codon usage with the potentially expressed ORFs reveals that the large ORF in reading frame 2 (113-2819) has codon usage characteristic of highly expressed Drosophila genes and that no other ORF shows extended preferential coding.

Figure 3.4 Codon preferences and ORFs in dri cDNA



than one EcoRI insert, with only one being subcloned and sequenced. In this case the subclone ends would not necessarily indicate which frame was expressed in the phage. To test this possibility, a single plaque of bk60 phage was used to provide the template for PCR amplification using a vector primer (PCR1) and a *dri* primer (RT). The 700bp product from this reaction covered the 5' end of the bk60 clone and would necessarily indicate which frame was fused to *lacZ* in bk60. Subcloning and sequencing of this product indicated that there had been no error in the frame assignment - the predicted reading frame of *lacZ* ran into RF3 (Figure 3.5). This result suggested that the *in vitro* work of Kalionis and O'Farrell had identified a short peptide that was not expressed in *Drosophila* but which apparently possessed a sequence specific DNA binding activity. Chapter 4 describes experiments which demonstrated that the DNA binding activity was not in fact associated with this reading frame, and provides alternative explanations for the sequencing result.

Analysis of sequence similarities

Extensive homology searches were undertaken with both ORFs to determine if they contained any conserved regions that might suggest a function[2]. As previously mentioned, both ORFs contained opa repeats that were considered insignificant in homology searches due to their abundance and lack of known function. A match was considered significant only if the sequence was non-repetitive and if a match of that length would be expected at random from a database of that size with a probability of <0.05 (see Methods). The shorter ORF in RF3 contained no such significant homology to any sequences in the database. Searches using the long ORF in RF2 (hereafter called the *dri* ORF) detected one section of about 90 codons with significant homology to two human proteins: MRF and RBP1. These two proteins were about 40% identical to *dri* and to each other across this region (Figure 3.6). These constituted highly significant matches which would be expected at random from a database the size of Genbank (rel92.0) with a

[2] In this context 'homology' is used to denote 'degree of sequence relatedness' rather than 'evolutionary conservation of body parts'. This usage comes from *logos* - word rather than *logos* - ratio.

Figure 3.5 Predicted coding frame for BK60.

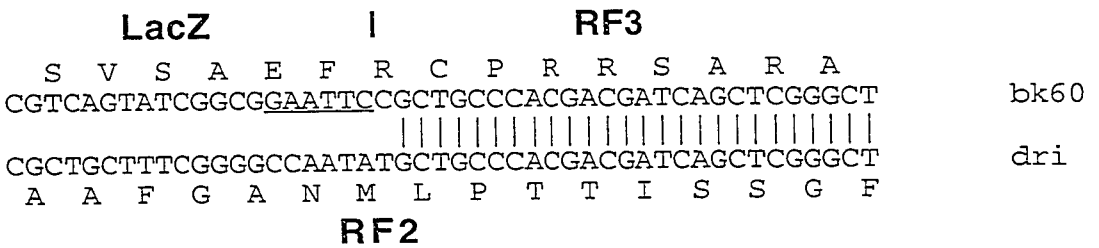


Figure 3.5 The sequence across the junction between *lacZ* and the *dri* cDNA insert in clone bk60 indicates that a LacZ-Dri fusion protein would be predicted to express the insert in reading frame 3 (RF3). The sequence of bk60 was confirmed by sequencing across the junction in both directions; the *lacZ* reading frame indicated represents the frame which generates functional β-galactosidase. The predicted fusion protein is written in one letter code above the bk60 DNA sequence, the amino acid sequence of *dri* in the long open reading frame expected to be translated (RF2) is written below. The EcoRI cloning site that marks the junction between vector and insert in bk60 is underlined.

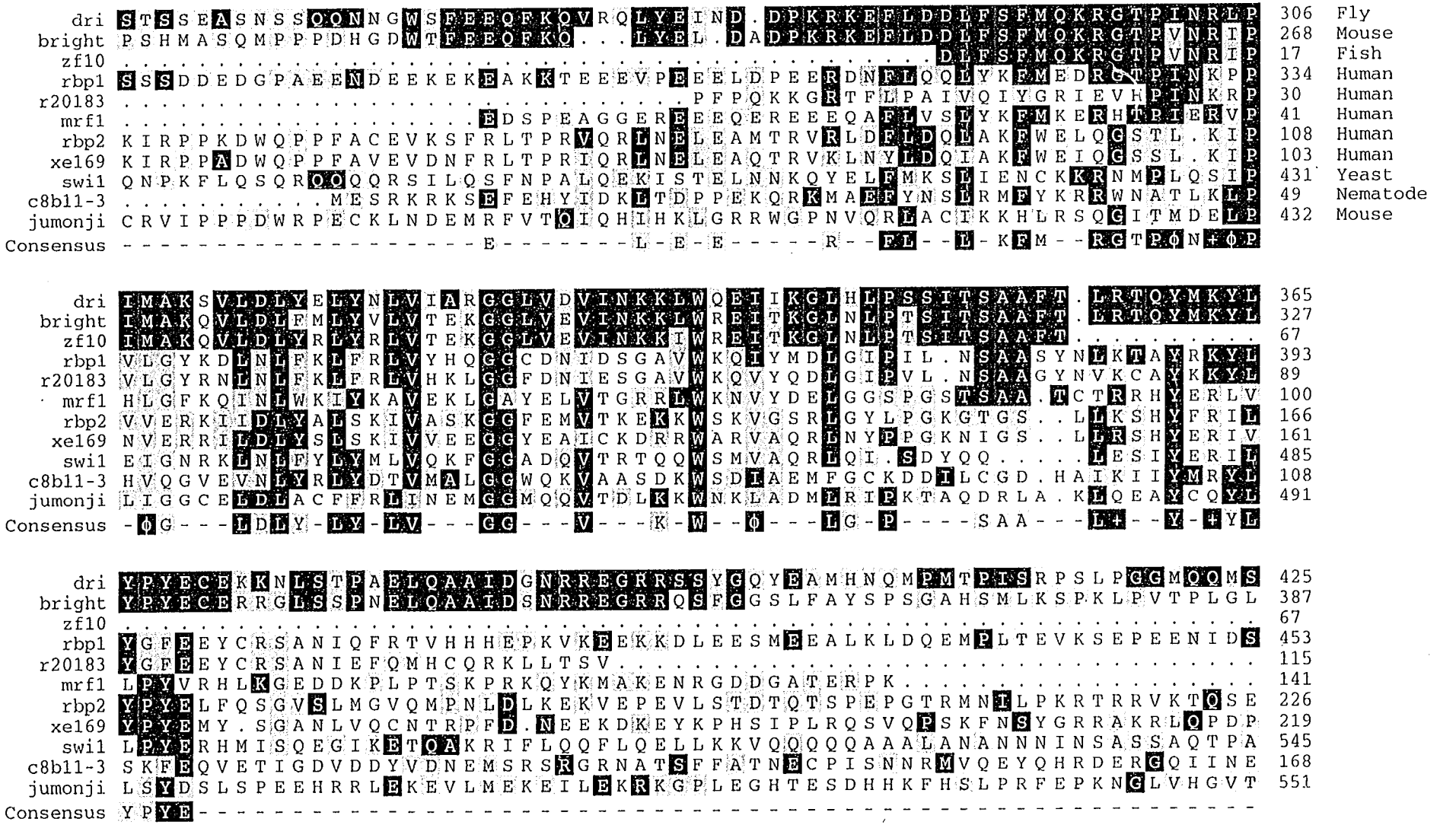
(rel92.0) with a probability of $<10^{-8}$. No other regions of significant homology were observed with these proteins or any others in the database.

Alignment of these sequences allowed the identification of highly conserved residues, which were used as a motif for further searches. When this was first carried out, three proteins were identified with the motif. As new sequences have been added to the database, more homologous proteins have been detected. As of December 1995, there are 13 proteins known to contain this motif, coming from organisms as diverse as yeast, nematodes, insects and mammals (Figure 3.6). The dark shading for highly conserved residues clearly delineates two boundaries for the homology. The core conserved region begins at 287 (FLDDLFS) in Dri and ends at 369 (KYLYPYE). The level of identity with Dri across this 83 residue region ranges from 83% for Bright to 24% for Jumonji. An extended region of homology exists between Dri and Bright, spanning 132 amino acids with 75% identity. Extended homology also exists between RBP2 and XE169 which continues to the N-terminus of both proteins. Phylogenetic analysis of this motif has been limited by the small number of sequences available; as yet there is no indication of distinct classes other than those suggested by the extended homologies (Fig. 3.6, result not shown).

None of the proteins in the database had previously been identified as sharing this motif, nor had this region been identified as a functional domain. It was therefore necessary to correlate the few known functions of these peptides to provide a tentative explanation for this high level of sequence conservation across such an evolutionary distance. The most obvious common feature of these proteins was that they bound DNA but contained no known DNA-binding domain. Dri, Bright and the two MRFs were all isolated in DNA binding screens, and RBP1 and 2 were also known to bind DNA (Herrscher et al., ; Whitson et al., ; Fattaey et al., 1993). SWI1, XE169 and SMCX were all implicated in chromatin mediated effects on transcription and Jumonji was predicted to be a transcription factor (Peterson and Herskowitz, 1992; Wu et al., 1994;

Figure 3.6 Alignment of the conserved region in Dri with related proteins. Proteins are listed in order of similarity to dead ringer. Black shading indicates residues identical to Dri, grey shading indicates residues that are related to Dri. Residues are considered related if they both fall into one of the following groups: M/L/I/V, D/E/N/Q, R/K/H, Y/F/W, G/S/A/T/P. The consensus at each position is formed from five or more identical residues or nine or more related residues. The consensus is black where seven or more of the proteins have residues identical to the consensus (highly conserved positions). The most widely conserved 83 residue motif corresponds to the section containing residues marked black in the consensus (287-369 in Dri). The symbols used are the standard one letter amino acid code with + for positively charged and ϕ for hydrophobic positions in the consensus.

Figure 3.6 Alignment of Dri related proteins



1994; Takeuchi et al., 1995). This obviously suggested a role for the conserved region as a DNA binding domain. This possibility was directly tested by experiments described in Chapter 4.

The secondary structure of this region was predicted from its sequence by several methods to determine whether some structural similarity might exist with known DNA-binding domains. The relatively trivial method of Chou and Fasman (1978) predicted a predominantly α-helical structure with three helices separated by short turn and sheet regions across the core motif (result not shown). On the assumption that all of the available instances of this motif would fold similarly, their sequences were compared to assess structural constraints using the PHD program (Rost and Sandler, 1994). This method has a three-state prediction accuracy of about 72% for this number of peptides. The predicted structure across the core motif was for four helices separated by short loops and no extended β-sheet regions (Fig. 3.7). This was consistent with the prediction generated using the Discrete State-space Model algorithm of Stultz *et al.* (1993). These predictions give only a tentative model for Dri structure, but they strongly suggest that this motif does not fold to form anything resembling the characteristic tri-helical structure of a homeo domain (Otting et al., 1990).

Summary

The objective of the work described in this chapter was to characterise the *dri* transcript and identify any regions likely to contribute to its function. Contiguous cDNA clones covering 3.7kb of *dri* message were isolated and sequenced. This contig contained a complete open reading frame of 901 codons as well as a shorter ORF of 176 codons which had been predicted to generate the DNA binding activity of bk60. Sequence polymorphisms and codon usage suggested that only the longer ORF was translated and this was confirmed by immunohistochemistry. Database searches revealed that Dri contained a novel, highly conserved motif found in proteins from a wide variety

Figure 3.7 Predicted secondary structure of conserved motif

```
                 ....,....1....,....2....,....3....,....4....,....5....,....6
        Dri      FLDDLFSFMQKRGTPINRLPIMAKSVLDLYELYNLVIARGGLVDVINKKLWQEIIKGLHL
                    HHHHHHHH          HHHHHHHHHHHHHHHHHHHH    EEEHHHHHHHHHHHH
        Rel      9865999987238856752389999999999999999918925432699999999994799

detail:
        prH      0027899888531000123589999999999999998510112257899999999610 0
        prE      0000000000001210010000000000000000000036651000000000000000
        prL      9872000011468867765310000000000000000048941113210000000038 9
>82% SUBSET      LLLHHHHHHH..LLLLLL..HHHHHHHHHHHHHHHHHH.LL.E...HHHHHHHHHH.LLL


                 ....,....7....,....8...
        Dri      PSSITSAAFTLRTQYMKYLYPYE
                  HHHHHHHHHHHHHHHHHH   H
        Rel      52378999999999897725624

detail:
        prH      2558899999998888741135
        prE      0000000000000000121111
        prL      7431100000000101137742
>82% SUBSET      L..HHHHHHHHHHHHHH.LL..
```

Figure 3.7 The PHD predicted structure for the widely conserved motif identified in Dri. This method uses the compiled amino acid sequences of Bright, RBP1, MRF1, SWI1 and XE169 to give reliability estimates for the prediction at each residue in Dri. The prediction is given as H: helix, E: sheet, blank: loop or other. The reliability for this three state prediction is given under each residue (0-9). The section under 'detail:' gives the probability for assigning each state. A subset of the prediction with an expected average accuracy of >82% is shown as the final line, with L: loop and '.': no reliable prediction. This prediction was made using the algorithm of Rost and Sandler (1994).

of other species. The few known functions of these proteins suggested that this motif had identified a novel class of DNA-binding domains.