The EMBO Journal vol.6 no.3 pp.767–777, 1987

Developmental and molecular analysis of *Deformed*; a homeotic gene controlling *Drosophila* head development

Michael Regulski, Nadine McGinnis, Robin Chadwick and William McGinnis

Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511, USA

Communicated by M.Noll

The characteristic morphology of many elements of the *Drosophila* body plan is crucially dependent upon the proper spatial expression of homeotic selector genes. The *Deformed* locus, which we isolated by virtue of its homology to the homeo box, is a candidate for a homeotic selector in the head region of the developing embryo. Here we show that null mutants of *Deformed* result in a loss of pattern elements derived from the maxillary and mandibular segments, and a duplication of a cuticular element of the larval head skeleton. Molecular analysis of the locus shows that *Dfd* transcripts are encoded in five exons distributed over 11 kb. The major transcript of 2.8 kb contains a 1758-bp open reading frame that would translate to yield a 63.5-kd protein containing a homeo domain and conspicuous regions of monotonic amino acid sequences. The *Dfd* protein exhibits extensive homology to a protein encoded by a *Xenopus* homeo box gene, *Xhox 1A*, suggesting that the *Xenopus* gene is the frog homologue of *Dfd*.

Key words: homeo box/homeotic gene/development/evolution/*Deformed*

Introduction

The homeotic genes of the Antennapedia (ANT-C) and Bithorax (BX-C) complexes of *Drosophila* appear to assign different fates to groups of cells on the anterior-posterior axis of the early embryo. This function is inferred principally from homeotic mutant phenotypes, which are pattern defects that often result in deletions of cuticular structures and replacement of missing structures with duplicated tissue normally developed at other positions on the anterior-posterior axis. For example, embryos with null mutations in the *Ultrabithorax* gene are missing cuticular markers from posterior second thoracic segment (pT2) through anterior first abdominal segment (aA1) and have structures normally derived from the posterior first thoracic through the anterior second thoracic segment (pT1-aT2) substituted in place of the missing structures (Sanchez-Herrero *et al.*, 1985). This homeotic transformation appears due to a default state, since *Ultrabithorax* transcripts and proteins normally present in pT2, pT3 and aA1 (Akam and Martinez-Arias, 1985; White and Wilcox, 1985; Beachy *et al.*, 1985) are replaced in the mutant by *Antennapedia* transcripts (Hafen *et al.*, 1984; Harding *et al.*, 1985). Thus it apears that some of these genes are part of a self-regulating system that controls morphological diversity in different body segments.

The individual genetic units of the Bithorax complex (BX-C) accomplish anterior-posterior diversification in the posterior thoracic and abdominal segments of the fly (Lewis, 1978; Sanchez-Herrero *et al.*, 1985; Karch *et al.*, 1985). Another separate cluster of genes, the Antennapedia complex (ANT-C), has individual units controlling diversity in the head and thoracic segments (Kaufman, 1983). At present, the two best defined homeotic genes in the ANT-C are *Antennapedia* and *Sex combs reduced*. *Antennapedia* lies near or at the right end (telomere side) of the ANT-C, and is necessary for the proper development of all three thoracic segments (Struhl, 1982). *Sex combs reduced* is important for the specification of cuticular structures in the first thoracic and labial segments (Kaufman, 1983; Struhl, 1983). Many of the other genetic units of the ANT-C can mutate to disrupt the development of the larval and adult head structures, *proboscipedia* being the only locus to yield an obvious homeotic phenotype in the head region (Kaufman, 1978). Of the remaining ANT-C loci, the best candidate for inclusion in the homeotic category is *Deformed* (*Dfd*). The original mutant allele at *Deformed* was dominant (Dfd^D) and caused a loss of ventral eye and orbital tissue. Vogt (1947) showed that a recessive allele of *Dfd*, when combined with *scute, echinus* and *cut* mutations, resulted in a high proportion of duplicated head structures (antennae and maxillary palps) in adult animals. Kaufman (1983) has also reported that somatic clones of *Dfd* mutant cells in adult heads show transformed phenotypes consistent with a homeotic transformation.

We have used homology to homeo box sequences, a family of protein-coding sequences conserved in *Antp* and other homeotic and segmentation genes, to isolate cloned regions that contain members of the *Drosophila* homeo box gene family. As reported previously, one of these regions appears to belong to the *Dfd* locus (Regulski *et al.*, 1985). Here we provide further evidence for that contention, and use molecularly characterized mutants of *Dfd* to define the terminal embryonic phenotype of animals lacking most or all *Dfd* gene products. The sequence of a full-length cDNA clone from *Dfd* indicates that a major product of the *Dfd* locus is a 63.5-kd protein.

Results

Lethal Dfd chromosomal breakpoint

In a screen for phenotypic revertants of a variety of dominant alleles of ANT-C genes, Hazelrigg and Kaufman (1983) recovered several revertants of Dfd^D that had lost the dominant adult phenotype (head deformations) but gained a recessive lethal phenotype when combined with alleles of the R3 complementation group of ANT-C (Lewis *et al.*, 1980). The complementation map and the clustering of reversion-induced breakpoints led them to propose that the R3 complementation group and Dfd^D were allelic and that the wild-type *Dfd* locus is located in chromosomal bands 84A4,5, in the middle of the ANT-C. The revertant breakpoint of one of the above, Dfd^{RX16}, was an entry point for the ANT-C DNA walk of Scott *et al.*, (1983), and provided a tentative definition of *Dfd* sequences in the ANT-C.

In a previous screen to recover homeo box copies from a *Drosophila* genomic library (McGinnis *et al.*, 1984), we isolated a clone, λ99, that contained DNA spanning the Dfd^{RX16} breakpoint. This clone also contained a transcribed region that was

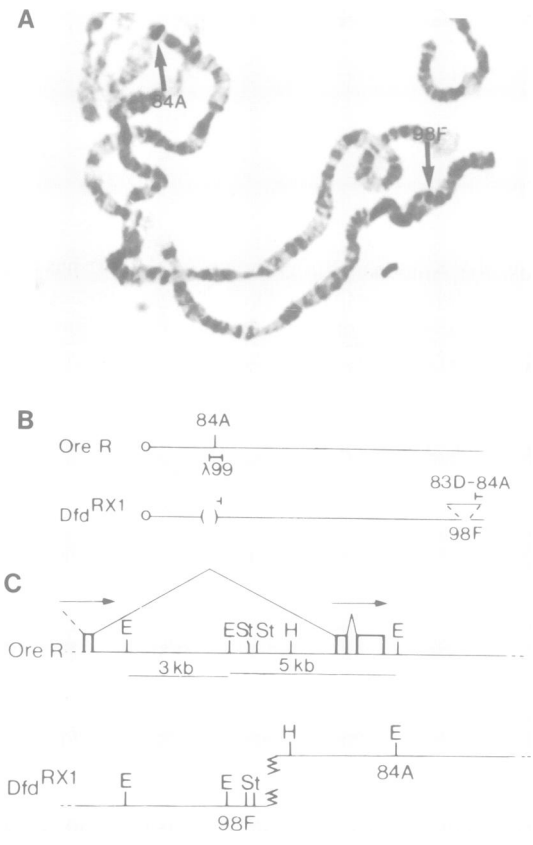

Fig. 1. Chromosomal breakpoint of Dfd^{RX1}. **A** shows a photomicrograph of polytene chromosomes from a larva of genotype $Dfd^{RX1}/OreR$. The chromosome squash was hybridized with a biotinylated probe from the *Dfd* locus (λ99, Regulski *et al.*, 1985) and the hybridized probe detected by an immunoperoxidase method (Langer-Safer *et al.*, 1982). The arrows indicate the sites of hybridization on the third chromosome at positions 84A and 98F. **B** is a schematic view of the right arm of the third chromosome of *OreR* and Dfd^{RX1} (Hazelrigg and Kaufman, 1983). The λ99 probe, which derives from in or near the wild-type *Dfd* locus in bands 84A4-5, labels only one site on the *OreR* chromosome. On the Dfd^{RX1} chromosome, two sites are labeled due to the translocation of part of 99 to an insertion site at the tip of 3R. **C** shows the location of the Dfd^{RX1} chromosomal breakpoint on a restriction map of the DNA from the *Dfd* locus. On the OreR DNA, the exons of the *Dfd* transcription unit (described more fully below) are shown as open boxes. The labeled lines below the OreR map denote the sizes of the two *Eco*RI restriction fragments in this interval. The arrows show the direction of transcription. On the Dfd^{RX1} DNA, the zigzag shows the position of the breakpoint as deduced from genomic blot analysis of $Dfd^{RX1}/OreR$ DNA probed with restriction fragments from the 99 insert. The DNA left of the breakpoint has been transposed to 98F, the DNA to the right remains at 84A. E = *Eco*RI, St = *Sst*I, H = *Hin*dIII.

interrupted by the Dfd^{RX16} breakpoint and was thus a candidate for the *Dfd* transcription unit. Another reversion-induced breakpoint with a lethal *Dfd* phenotype is carried by the Dfd^{RX1} chromosome (Hazelrigg and Kaufman, 1983), which has a transposition of bands 83D4,5-84A4,5 to an insertion point at 98F1,2 on the distal right arm of chromosome 3 (see Figure 1). Genomic sequences from λ99 are present on both sides of the breakpoint, as is shown by *in situ* hybridization to the Dfd^{RX1} chromosome (Figure 1). Southern blot analysis of Dfd^{RX1} DNA allows us to place the 84A4,5 breakpoint within a 5-kb genomic *Eco*RI fragment, very near the Dfd^{RX16} breakpoint (Scott *et al.*, 1983). Both breakpoints interrupt the homeo box-containing transcription unit shown schematically in Figure 1. We will henceforth refer to this as the *Dfd* transcription unit, since its integrity appears to be necessary for the Dfd^{D} function, its interruption results in a recessive lethal phenotype of the *Dfd* complementation group and molecular tests show it to be the sole transcription unit in this region (described below).

Dfd null mutations disrupt embryonic head development

The embryonic head of *Drosophila* can be roughly divided into two regions, the procephalon and the gnathocephalon. The number of segment primordia in the procephalon has long been a disputed issue, but two lobes, the procephalic and clypeolabral, are morphologically distinguishable (see Figure 2). The gnathocephalon is formed by three lobes, which correspond to the primordia for the mandibular, maxillary and labial segments. At early stages (Figure 2, 9 h) these lobes exhibit a metameric arrangement along the ventro-lateral aspect of the embryo. Throughout the remaining stages of embryonic head development, rearrangements and fusions of the gnathocephalic lobes obscure their segmental origins. In addition, *Drosophila* embryos involute the bulk of their head primordia during embryogenesis (Figure 2, 14 h). The first instar larva that emerges from this eggshell has only a few head structures visible at the anterior tip, the pseudocephalon, surrounded by the cuticle of the first thoracic segment.

Since the *Dfd* gene is a member of the homeo box gene family and is expressed in a discrete group of cells in the anlagen for the head (McGinnis *et al.*, 1984), we wished to examine the effect that loss of *Dfd* products had on embryonic head development. In particular, we wished to test which segmentally derived structures were missing, and which (if any) were duplicated in a homeotic transformation. Wakimoto *et al.* (1984) have previously described the abnormalities resulting from a variety of EMS-induced mutant alleles of *Dfd*. These mutants developed normal thoracic and abdominal segments, but had constricted anterior ends. Mouth hooks were reduced in size but cephalopharyngeal skeletons, maxillary sense organs and maxillary cirri had wild-type morphology.

We have examined individual larvae of the genotype $Dfd^{RX1}/Df(3R)Scr$, which we believe closely approximates a homozygous deficiency for all *Dfd* sequences while leaving all adjacent genes intact. The *Df(3R)Scr* is deleted for all *Dfd* sequences (Lewis *et al.*, 1980), and as shown in Figure 1, the Dfd^{RX1} chromosome breaks the *Dfd* transcription unit and transposes the 5′ end to the end of the third chromosome. Thus no intact *Dfd* gene is present in embryos of the genotype $Dfd^{RX1}/Df(3R)Scr$. We will henceforth refer to this as Dfd^{-}.

The Dfd^{-} genotype has a visible effect on all the gnathal lobes during embryonic development. From 9−11 h of normal development, the labial lobes move ventrally, take on an oblate form, partially rotate and join at the midline (Figure 2a). Soon afterwards (12−13 h) the labial lobes fuse, and eventually involute through the stomodeum, forming the ventral border of the atrium (Figure 2b). In Dfd^{-} embryos, the labial lobes do not migrate to the midline and they remain circular in form. Although the bilateral lobes eventually appear to fuse, the fusion is accomplished hours later in development than in Dfd^{+} (Figure 2b).

The maxillary lobe normally fuses with the mandibular and procephalic lobes between 9 and 10 h of development. It later elaborates many of the visible elements of the first instar larval head including the cirri, mouth hook and ventral elements of the maxillary sense organ (Turner and Mahowald, 1979; Jürgens *et al.*, 1986). The Dfd^{-} embryo exhibits severe disruptions in maxillary lobe development (Figure 2a and b). The lobe fails to fuse with either mandibular or procephalic lobes even though it occupies a roughly normal position throughout the movements of head involution. Neither cirri nor mouth hooks develop from

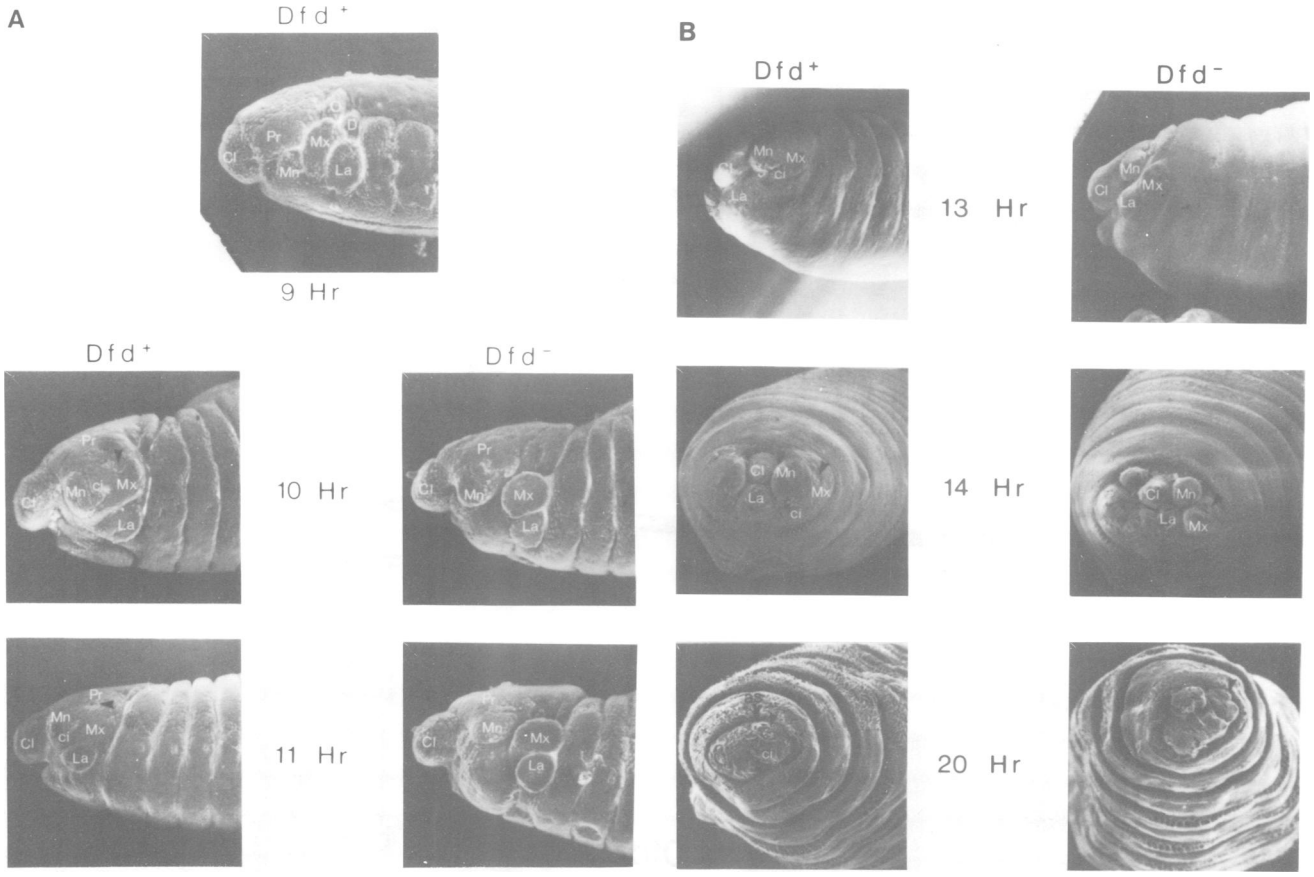

Fig. 2. *Dfd*$^-$ embryonic morphology. **(a)** A sequence of scanning electron micrographs of both wild-type and *Dfd*$^-$ embryos. Ventro-lateral views of the anterior of the embryos show the movements of the head segments and their respective structures from 9 to 11 h of development. *Dfd*$^+$ = *Dfd*^{RX1}/*Dfd*$^+$, *Dfd*$^-$ = *Dfd*^{RX1}/*Df(3R)Scr*, Cl = Clypeolabrum, Pr = Procephalic lobe, O = Optic lobe, D = Dorsal ridge, Mn = Mandibular lobe, Mx = Maxillary lobe, La = Labial lobe, ci = cirri, arrowhead = maxillary sense organ. **(b)** A sequence of scanning electron micrographs of both wild-type and *Dfd*$^-$ embryos from 13 to 20 h of development. 13 Hr − ventro-lateral view during head involution. 14 Hr − anterior view near the end of head involution. 20 Hr − anterior view. Cl = Clypeolabrum, Mn = Mandibular lobe, Mx = Maxillary lobe, La = Labial lobe, ci = cirri, as = antennal sense organ, arrowhead = maxillary sense organ.

the *Dfd*$^-$ maxillary lobe. A maxillary sense organ appears but is separated from its usual neighbor, the antennal sense organ. The maxillary organ that develops is disorganized and is missing the large dorso-medial and dorso-lateral papillae (DMP and DLP) that Frederick and Denell (1982) suggest are developmentally grouped with the antennal sense organ.

The mandibular lobe has few identifying features. In wild-type embryos, it fuses with the procephalic lobe between 9 and 10 h and in *Dfd*$^-$ embryos this developmental fusion is also accomplished. The mandibular lobe also moves dorsally from 10 to 13 h of normal development, and this movement is also present in *Dfd*$^-$ embryos (Figure 2a and b). The procephalic and clypeolabral lobes appear largely unaffected in the early *Dfd*$^-$ embryos.

Though the above phenotypes are somewhat variable, they are seen in >50% of the embryos assigned to the *Dfd*$^-$ class. As a class, these embryos are also slightly delayed in head involution, and have a characteristic ventral bulge at 11 h of development in the region normally occupied by the labial lobes. No developmental abnormalities are observed in thoracic and abdominal segments.

The Dfd$^-$ *phenotype involves a pattern duplication of part of the head skeleton*

Dfd$^-$ embryos die before emerging from the eggshell. The terminal phenotype of the cephalic cuticular structures can be observed in cleared preparations of the dead embryos. In Figure 3, we have compared the head skeleton of first instar larvae hemizygous for the *Dfd* locus (*Dfd*^{RX1}/*Balancer*) with fully developed *Dfd*$^-$ embryos, which are essentially unhatched first instar larvae. The cephalopharyngeal skeleton of the *Drosophila* first instar larve is composed of four major cuticular structures (Strasburger, 1932; Jürgens *et al.*, 1986). From anterior to posterior, these are the mouth hooks, median tooth, H-piece and cephalopharyngeal plates (Figure 3). These chitinous structures are secreted by cells of the atrium and pharynx of the developing embryo, and largely derive from involuted gnathal lobe cells.

In strong *Dfd*$^-$ phenotypes which comprise >50% of the *Dfd*$^-$ class, the mouth hooks and H-piece are missing, and the lateral processes of the cephalopharyngeal plates are duplicated in a more anterior position in the atrial cavity, replacing the lateralgraten and H-piece (Figure 3). A distorted version of the median tooth is attached to the duplicated cephalopharyngeal plates. We believe this structure is a version of the median tooth since weaker phenotypes always include the tooth in this position.

The antennal and maxillary sense organs of the head can also be observed in these cleared embryos. In *Dfd*$^-$ embryos, both the antennal and maxillary sense organs are present though in severely disrupted form. The antennal sense organ, which is normally situated on the pseudocephalon, now is found within the atrium or frontal sac. The maxillary sense organ is missing both dorso-lateral and dorso-medial papillae, and the remaining pap-

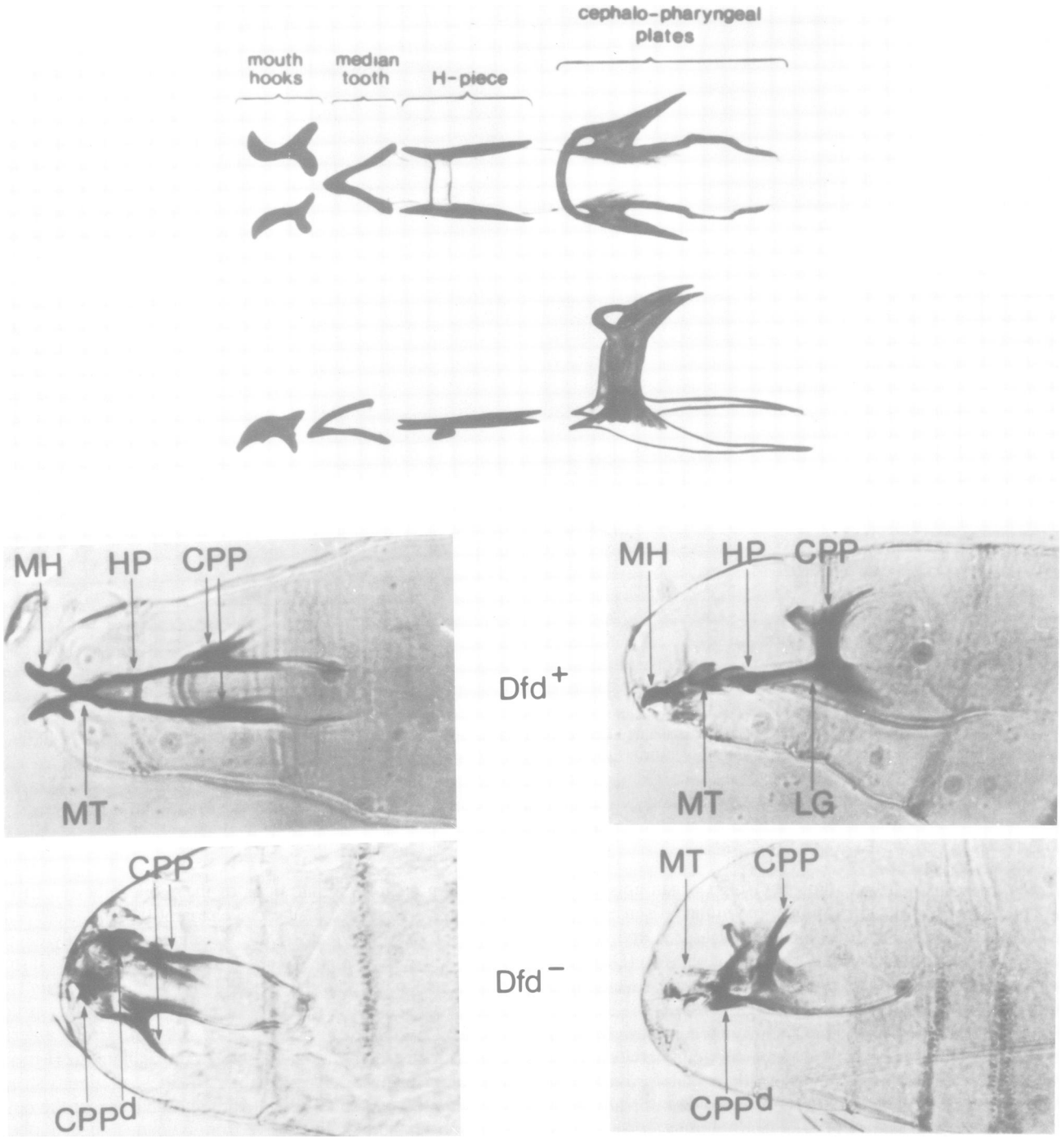

Fig. 3. Homeotic transformation. **(Top)** A schematic and exploded view (frontal and lateral) of the sclerotized mouth and head parts of a first instar larva. Discussed in the text are the recently named lateralgraten (Jürgens *et al.*, 1986) which are the ventral anterior extension of the cephalopharyngeal plates that are fused with the H-piece in first instar larvae. **(Bottom)** Cleared cuticle preparations of first instar larvae just before hatching. The left panels are frontal views of both Dfd^+ and Dfd^- embryos. Note the duplicated cephalopharyngeal plates in the Dfd^- embryos. $Dfd^+ = Dfd^{RX1}/TM3$ (a third chromosome balancer with a wild-type *Dfd* locus), $Dfd^- = Dfd^{RX1}/Df(3R)Scr$, MH = mouth hooks, MT = median tooth, HP = H-piece, LG = lateralgraten, CPP = cephalopharyngeal plates, CPP^d = duplicated cephalopharyngeal plates.

illae appear incompletely developed (not shown). The labial sense organ is sometimes obscured by the extra chitin secreted in the atrium due to the cephalopharyngeal plate duplication, but when visible appears to develop normally.

Molecular structure of the Dfd locus

Our molecular analysis of *Dfd* began with the genomic clone λ99, isolated previously using homeo box homology (McGinnis *et al.*, 1984), and a neighboring genomic clone, λ100. Together these span ~24 kb in the 84A4-5 region of the third chromosome. Using restriction fragments from the genomic intervals 240–255.8 (Figure 4) as probes, Northern blots containing RNA from the various stages of *Drosophila* development yield a single hybridization signal corresponding to an RNA species of 2.8 kb.

Two genomic *Eco*RI fragments, one of 5.0 kb spanning position 245–250, and another of 1.9 kb spanning 253.1–255.0

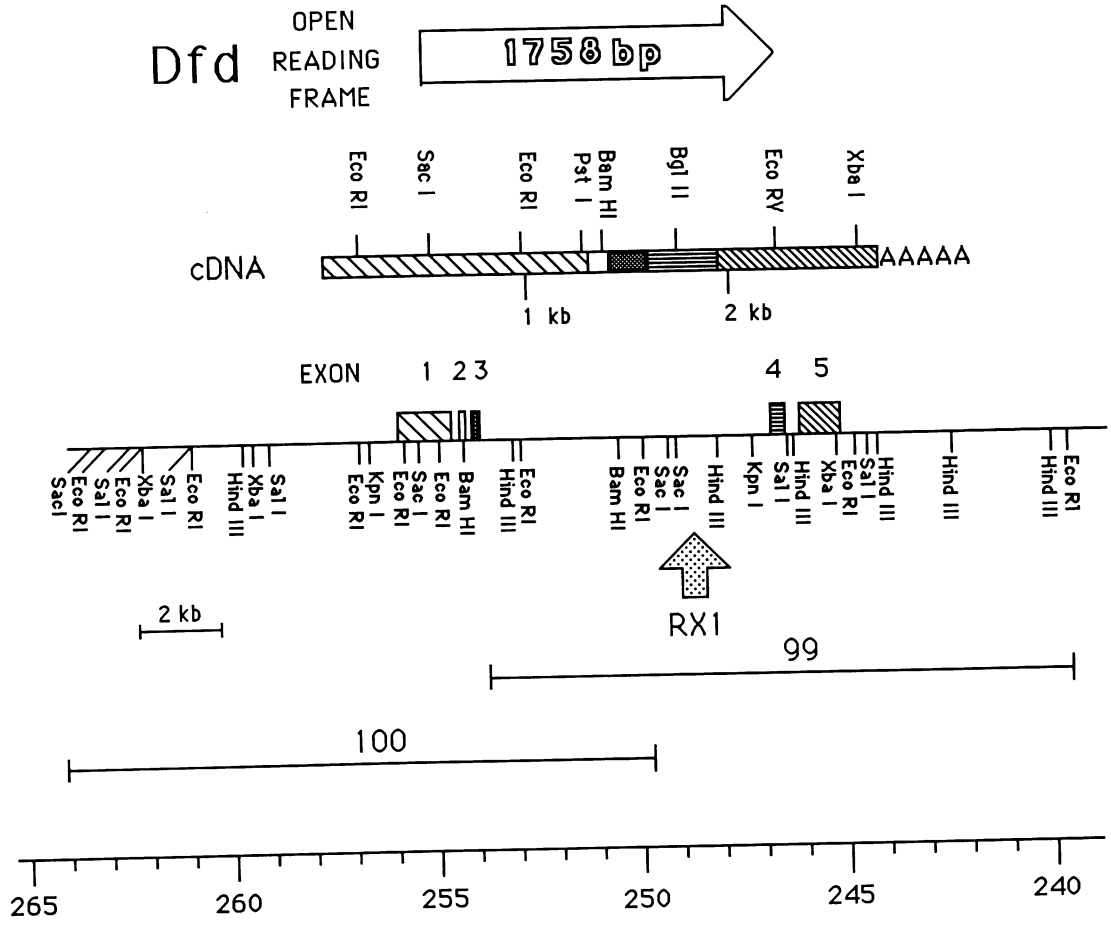

Fig. 4. *Dfd* genomic and cDNA maps. The middle line of this figure represents the genomic DNA in and around the *Dfd* locus. The restriction map of this DNA was constructed from cloned segments in λ100 and λ99. The extent of these cloned regions is indicated by the labeled brackets below the map. At the bottom is the distance in kb from the *Humeral* chromosomal breakpoint (Garber *et al.*, 1983). *Dfd* exons are denoted by open and filled boxes above the genomic DNA. The five exons are represented in the schematic cDNA (cDfd41) of 2.75 kb. The position and direction of the long open reading frame within the cDNA is shown by an open horizontal arrow. This cDNA contains a stretch of 20 A residues at its right end, preceding the cloning site. The dotted vertical arrow indicates the chromosomal breakpoint of the Dfd^{RX1} rearrangement (as shown in Figure 1). The direction of transcription is from left to right.

were used as hybridization probes on an OreR 3−12 h embryonic cDNA library of L.Kauvar (Poole *et al.*, 1985). Approximately 3×10^5 plaques were screened and a total of 12 purified cDNA clones were recovered. These were analyzed by restriction enzyme mapping and hybridization of labeled cDNA inserts to Southern blots of restriction digested genomic DNA. The cDNA inserts fell into three size classes of 2.75, 2.6 and 1.8 kb. All hybridized to the five exonic regions shown in Figure 4. However, the 2.6-kb class hybridized the 0.8-kb genomic *Eco*RI fragment from 255.0−255.8 which the 1.8-kb class did not hybridize. The 2.75 class, in addition, hybridized the 1.1-kb *Eco*RI fragment immediately to the left of the 0.8-kb fragment. Since the largest size class corresponded closely to the size expected for copies of the 2.8-kb *Dfd* transcript, we chose one (cDfd41) for nucleotide sequence determination. The 1.8- and 2.6-kb size classes presumably arose from *Eco*RI digestion of cDNAs during the construction of the library due to incomplete methylation of internal *Eco*RI sites.

The five exons contained in the cDfd41 are spread over 11 kb of genomic DNA. The 3rd and 4th exons are interrupted by a fairly large intron of 7 kb. The breakpoints for each of the lethal *Dfd* rearrangements, Dfd^{RX1} and Dfd^{RX16}, are found in this intron.

5′ limits of the Deformed transcription unit

To define the 5′ end of mature *Dfd* transcripts, we subjected poly(A)$^+$ RNA from 6−12 h embryos to both primer extension and S1 nuclease protection analyses. The primer for the extension reaction was a 150-bp *Taq*I fragment with a 3′ OH terminus 20 nucleotides from the left end of the cDfd41 insert (Figure 6), and extending past the *Eco*RI site into the polylinker region of an mp8 vector. This primer, after hybridization to *Dfd* transcripts in 6−12 h RNA, was extended for 22 nucleotides. The results are shown in Figure 5a. This places the 5′ end (which we arbitrarily designate as nucleotide # 1) two nucleotides upstream from the 5′ end of the cDfd41 insert. An S1 nuclease protection experiment using a 0.29-kb *Hae*III−*Eco*RI genomic fragment which overlaps the 5′ end of cDfd41 confirms this site as the 5′ terminus of *Dfd* transcripts (Figure 5b).

Sequence analysis of the Dfd genomic and cDNA

We have sequenced both strands of the cDNA, cDfd41 and the corresponding genomic DNA. As shown in Figure 6, the sequences differ at only a few sites. Most of these differences are single base pair substitutions which have no effect on the amino acid sequence encoded by the long open reading frame in this cDNA. In three positions, however, there are small deletions of 1, 9 and 3 bp (positions 119, 1919−1928 and 2122−2125, respectively) in the cDNA insert relative to the genomic sequence. The major difference is a stretch of 48 bp (position 2558−2606) where there is no similarity between cDNA and genomic sequences. Many of these differences could be due to inter-strain

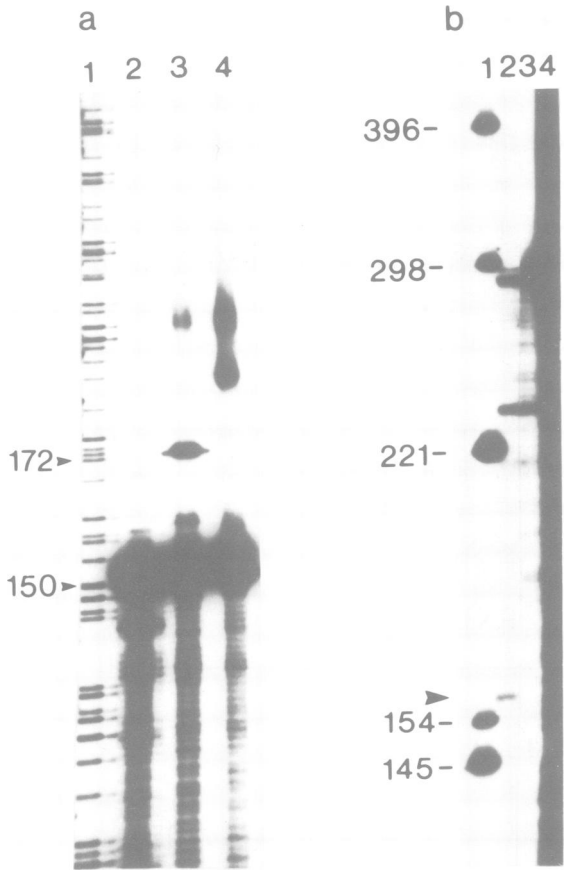

Fig. 5. 5′ end primer extension and nuclease protection. **(a)** Primer extension. **Lane 1** contains the products of a sequencing reaction as size markers. Numbers next to the arrowheads correspond to the size of unextended (150) and extended primer (172). **Lane 2** contains primer alone. **Lane 3** contains products of the extension reaction in which poly(A)$^+$ RNA was used as a template. **Lane 4** contains products of the reverse transcriptase reaction on carrier (yeast tRNA) only. **(b)** S1 nuclease protection. **Lane 1** contains size markers consisting of end-labeled *Hin*fI fragments of pAT153. **Lane 2** shows the signal arising from the end-labeled fragment protected from S1 nuclease digestion by poly(A)$^+$ embryonic RNA (see Results and Materials and methods for details). The protected fragment is 158 ± 1 bp long. Bands appearing above and below that fragment result from self-protection of the probe (see **lane 4**). **Lane 3** contains the same reaction mixture as **lane 2** except that carrier RNA (yeast tRNA) was added instead of poly(A)$^+$ RNA to control for non-specific protection. **Lane 4** contains only the probe to control for bands arising from self-protection.

polymorphism since the genomic and cDNA libraries were prepared from different fly strains (CantonS and OregonR, respectively). The length of the cDfd41 insert is 2736 bp, which is in reasonably good agreement with the size of the *Dfd* mRNA on Northern blots (2.8 kb). A stretch of 20 adenosine residues present on one of the ends of cDfd41 and not present in the corresponding genomic position provides a tentative definition of the 3′ end of the transcription unit to a G residue at position 2751. It is closely preceded by two consensus polyadenylation signals (positions 2698 and 2727) (Proudfoot and Brownlee, 1976).

Further comparison of *Dfd* genomic and cDNA sequences allowed us to allocate exon/intron boundaries. All of them closely matched consensus sequences for eukaryotic donor ($^{C}_{A}$AG/GT$^{A}_{G}$AGT) and acceptor [$(^{T}_{C})_{n}$N$^{C}_{T}$AG/G] splice sites (Mount, 1982). Thus the *Dfd* locus consists of five exons of the following sizes (based on the genomic sequence): exon 1, 1277 bp, exon 2, 101 bp, exon 3, 178 bp, exon 4, 389 bp and exon 5, 806 bp. They are separated by intervening sequences of lengths of: 385 bp (intron 1), 74 bp (intron 2), over 7 kb (intron 3) and 265 bp (intron 4). The homeo box homology and CAG repeat (McGinnis *et al.*, 1984; Wharton *et al.*, 1985; Regulski *et al.*, 1985) are found in exon 4.

There is only one long, ATG codon initiated, open reading frame (ORF) in cDfd41 which starts at the ATG at position 491 and is 1758 bp long. It is in-frame with the homeo box homology region. Although it is not known presently which ATG triplet in this frame serves as a translation start site, this ORF has a potential to code for a 586-amino acid protein with a mol. wt of 63.5 kd. The primary structure of such a protein is discussed below. Besides this ORF the *Dfd* transcript contains untranslated leader and trailer sequences of 490 and 491 bp, respectively.

Discussion

Dfd mutant phenotype

Our analysis of the embryonic head structures that are lost and duplicated in *Dfd* null mutants strongly suggests that *Dfd* is a homeotic selector gene (Garcia-Bellido, 1977), i.e. its persistent expression is necesary for the selective determination of structures deriving from two head segments. The structures found deleted from the head are all specified by the epidermis of the maxillary and mandibular segments (see Figure 7). The cephalopharyngeal plates, which are the most obvious duplicated structures, are proposed by Jürgens *et al.* (1986) to originate from the acron, the anterior-most, non-segmental anlagen of the insect head. Whether these plates derive from the acron or a procephalic head segment, the result is that more posterior structures are being replaced by a structure normally produced more anteriorly, resulting in a pattern duplication of the anterior element. In this case, due to the process of head involution, the positions of the anlagen have been reversed, and the duplicated structure is found in a more anterior position, the cephalopharyngeal plates replacing the lateralgraten and the lateral bars of the H-piece. This type of pattern replacement is seen in homeotic transformations caused by null mutations in many other genes of the selector class, in which a more posterior structure is replaced by a duplicated copy of a more anterior structure in the segmented body plan.

The loss of pattern elements clearly does not affect all derivatives of the maxillary segment as there is still a recognizable maxillary sense organ produced. Thus far we have looked carefully only for loss and/or duplication of cuticular or sensory structures derived from the epidermis of the head segments. It is not yet known whether the morphology of mesodermal or neural elements is affected.

Where does the *Dfd* gene fit in the homeotic hierarchy? In many ways it fits the profile of a homeotic gene or genes predicted by Struhl (1983) that would become activated in more posterior thoracic and abdominal body segments upon the elimination of ANT-C and BX-C homeotic genes, and specify the development of mouth hooks, cirri and other head structures in place of the missing thoracic and abdominal pattern elements. Thus *Dfd* might be directly or indirectly repressed in posterior regions by the single or combined effect of homeotic selector genes such as *Sex combs reduced, Antennapedia* and *Ultrabithorax*. In turn, *Dfd* itself might be required to repress the function of another unknown 'head' selector gene that normally specifies cephalopharyngeal plates.

Deformed structure

In comparison with other homeo box loci, e.g. *Antp* (~100 kb) and *Ubx* (~75 kb), *Dfd* appears as a medium-size transcription

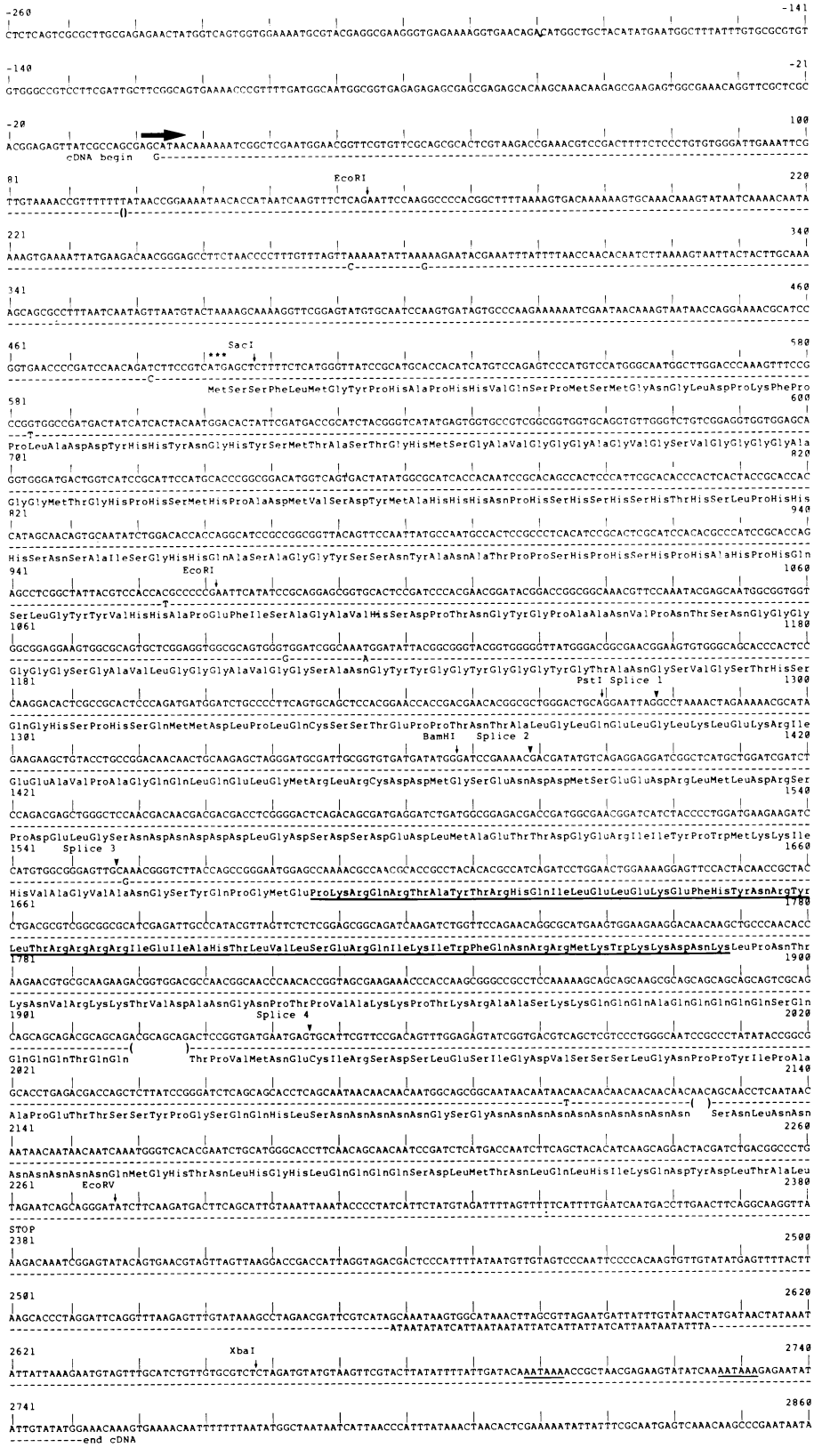

Fig. 6. Nucleotide sequence of the *Dfd* genomic and cDNA. The genomic sequence is given in letters. In addition to the regions corresponding to the cDNA, it includes 260 bp of upstream and 110 bp of downstream sequences flanking the *Dfd* transcription unit. The cDNA sequence is given beneath the dashes showing identities and letters indicating the differences from the genomic sequence. There are only a few changes from our sequence in the region matching the partial *Dfd* cDNA sequence reported by Laughon *et al.* (1985). The parentheses define the sequences deleted from the OregonR cDNA. The number of nucleotides starts at the transcription initiation site indicated by a horizontal arrow. The arrowheads point to the splice sites and the arrows to the restriction sites given in Figure 4. The polyadenylation signal consensus sequences are underlined. The predicted amino acid sequence of the *Dfd* protein is shown below the cDNA sequence. The first ATG triplet of the long ORF is indicated by asterisks and the stop codon is denoted by STOP. The amino acids of the homeo domain are underlined.

unit with five exons in 11 kb of genomic DNA (Schneuwly *et al.*, 1986; Bender *et al.*, 1983; Beachy *et al.*, 1985). We have thus far detected only one size class of transcripts from *Dfd* throughout the entire period of its expression. This distinguishes *Dfd* from other selector genes which encode a few classes of transcripts (Scott *et al.*, 1983; Akam and Martinez-Arias, 1985; Kuroiwa *et al.*, 1985).

The nucleotide sequence at the presumed transcription start site shows a considerable degre of homology (four matches of seven) to a consensus sequence for transcription initiation sites found for *Drosophila* genes ($ATCA^{G}_{T}T^{C}_{T}$) (Hultmark *et al.*, 1986). A search of upstream regions for the TATA box consensus sequence (Breathnach and Chambon, 1981) has not revealed a region of significant homology at the appropriate locations. A similar lack of TATA homology has also been reported for the second promoter region of the *Antp* locus (Schneuwly *et al.*, 1986). However, comparison of genomic sequences preceding the transcription start in both loci reveals significant sequence conservation. Three nucleotides immediately preceding the start site are identical in both cases and in total there are six matches between positions −9 and −1. (The *Dfd* sequence in this interval is TCGCCAGCG, *Antp* is TCACTGGCG, the consensus is TCPuCPyrPuGCG.) There is no significant homology between this region of *Dfd* and the corresponding part of the first promoter of *Antp* which contains the TATA box. It is possible that in the absence of the TATA box there may be other consensus sequences near the transcription start site which contribute to proper initiation of transcription.

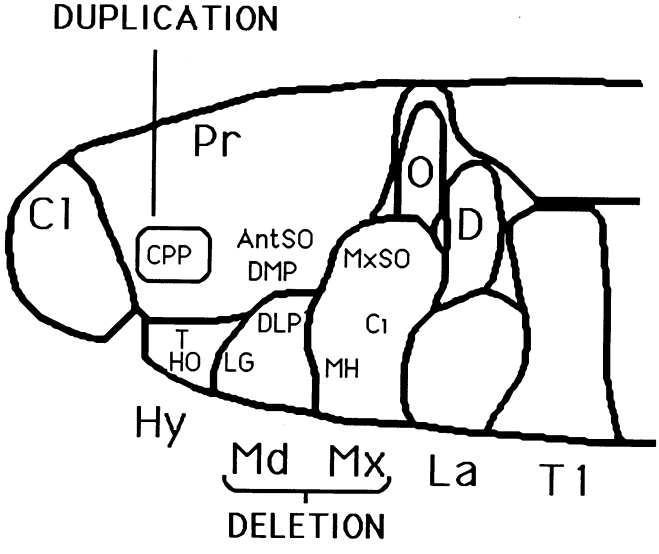

Fig. 7. Pattern deletions and duplications of *Dfd*⁻ larvae. A schematic lateral view of the anterior end of an embryo at the extended germ band stage. The metameric lobes of the head are most evident at this stage of embryonic development. The location of many of the cuticular and sensory structures fate mapped by Jürgens *et al.* (1986) are indicated. The bracket labeled deletion denotes that many (but not all) structures specified by mandibular and maxillary epidermis are missing from *Dfd*⁻ embryos. The site of the primordia for the duplicated cephalopharyngeal plates is also indicated. Cl = Clypeolabrum, Pr = Procephalic lobe, Hy = Hypopharyngeal lobe, Md = Mandibular lobe, Mx = Maxillary lobe, La = Labial lobe, T1 = 1st thoracic lobe, CPP = cephalopharyngeal plates, AntSo = Antennal sense organ, DMP = Dorso lateral papilla, LG = Lateralgraten, MxSO = Maxillary sense organ, Ci = Cirri, MH = Mouth hooks, O = Optic lobe, D = Dorsal ridge.

The 1758-bp ORF present in the cDNA contains 23 ATG codons. Of these, 18 are upstream and in-frame with homeo box codons. Though it is not known presently which of these triplets is used as a translation start signal, it seems likely that it is the first one at position 491. It has the best match to the consensus for translation start sequences found for *Drosophila* genes ($ANN^{C}_{A}A^{AA}_{CC}ATGNNN$) (D.Cavener, personal communication). Also, cell-free translation of *in vitro* synthesized *Dfd* RNA yields a protein of a size similar to the one which would result from translation of the 1758-bp ORF (M.Kuziora and M.Barad, personal communication).

The presumptive *Dfd* protein is the largest and most complex protein product of the homeo box genes characterized so far. Salient features of its primary structure are presented in Figure 8. Its amino acid composition compared with an average globular protein of a similar size shows large deviations in abundance of certain residues. Phenylalanine, cysteine, tryptophan and lysine are present at levels a few times lower than average (six times, three times, three times and twice, respectively). On the other hand, the number of histidine, methionine and asparagine residues is 2−4 times higher than average. In each case the value is 2

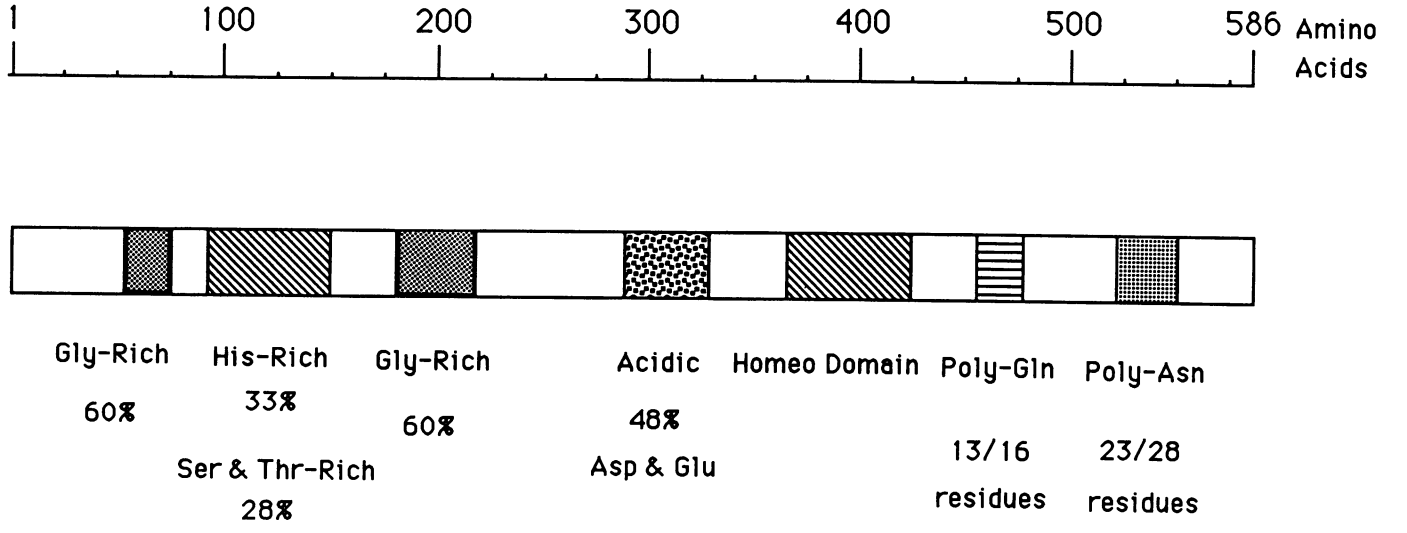

Fig. 8. Primary structure of the *Dfd* protein. The *Dfd* protein which would result from the translation of the long cDfd41 open reading frame is represented by an open bar. The scale in amino acid residues is given above. The filled boxes correspond to the conspicuous regions discussed in the text and the characteristic features of those regions are indicated below the bar.

standard deviations or more away from the expected for an average cysteine-poor globular protein (Sheridan *et al.*, 1985). The content of glycine, glutamine and serine is also higher than average. Some of these features are reflected in the presence of regions rich in certain amino acids in the *Dfd* protein. There are two glycine-rich regions in the N-terminal domain. A similar domain is present in the protein products of the *Ubx* locus (Beachy *et al.*, 1985). The two glycine-rich regions in *Dfd* are separated by a region rich in histidine, serine and threonine. All of these residues are hydrophilic and the whole region might be expected to form one of the outside surfaces of the *Dfd* protein. A serine-rich domain is also found in the *en* protein, though in *en* it consists of continuous strings of serine residues (Poole *et al.*, 1985). The middle part of the *Dfd* protein contains an acidic region and the homeo domain. Both residues are highly charged and are expected to be exposed to environment. A similar highly acidic region (~75%) is also believed to be present in the *engrailed* protein (Poole *et al.*, 1985).

The carboxyl part of the *Dfd* protein contains two regions of monotonic amino acids: polyglutamine and polyasparagine. The first was found to be present in a variety of proteins and is not an exclusive feature of homeo domain-containing proteins (Wharton *et al.*, 1985; Poole *et al.*, 1985; Regulski *et al.*, 1985; Schneuwly *et al.*, 1986). The second one is encoded by an $AA(^{T}_{C})$ repeat. It is present in many copies in the fly genome (M.Mlodzik, personal communication). Both polyGln and polyAsn are hydrophilic and could give rise to random coil regions at the carboxyl terminus of the protein, and might be important factors affecting the stability of the protein.

It has previously been shown that *engrailed* class genes of *Drosophila* have apparent homologues in the mouse genome, one of which is named Mo-en.1 (Joyner *et al.*, 1985). Though it is still unclear to what extent common functions are conserved, the mouse and *engrailed* class genes share extensive homology both within their respective homeo box regions as well as downstream of the homeo box. At the least, this homology implies a common evolutionary history separate from other homeo box genes prior to the divergence between the vertebrate and arthropod lineages.

The comparison of *Dfd* amino acid sequences with those from other homeo box genes indicates that it also has an apparent vertebrate homologue, a *Xenopus* gene designated *Xhox-1A* (Harvey *et al.*, 1986). This *Xenopus* gene was isolated with homeo box probes and shows the highest level of homology (89%, 54/61 residues) to *Dfd* within the amino acids of the homeo domain. The frog gene also shares extensive homology with *Dfd* outside the homeo domain (45 of the remaining 169 residues). Overall, the *Xhox-1A* protein is 43% homologous to *Deformed* (99/230 residues). The homologous regions of the respective

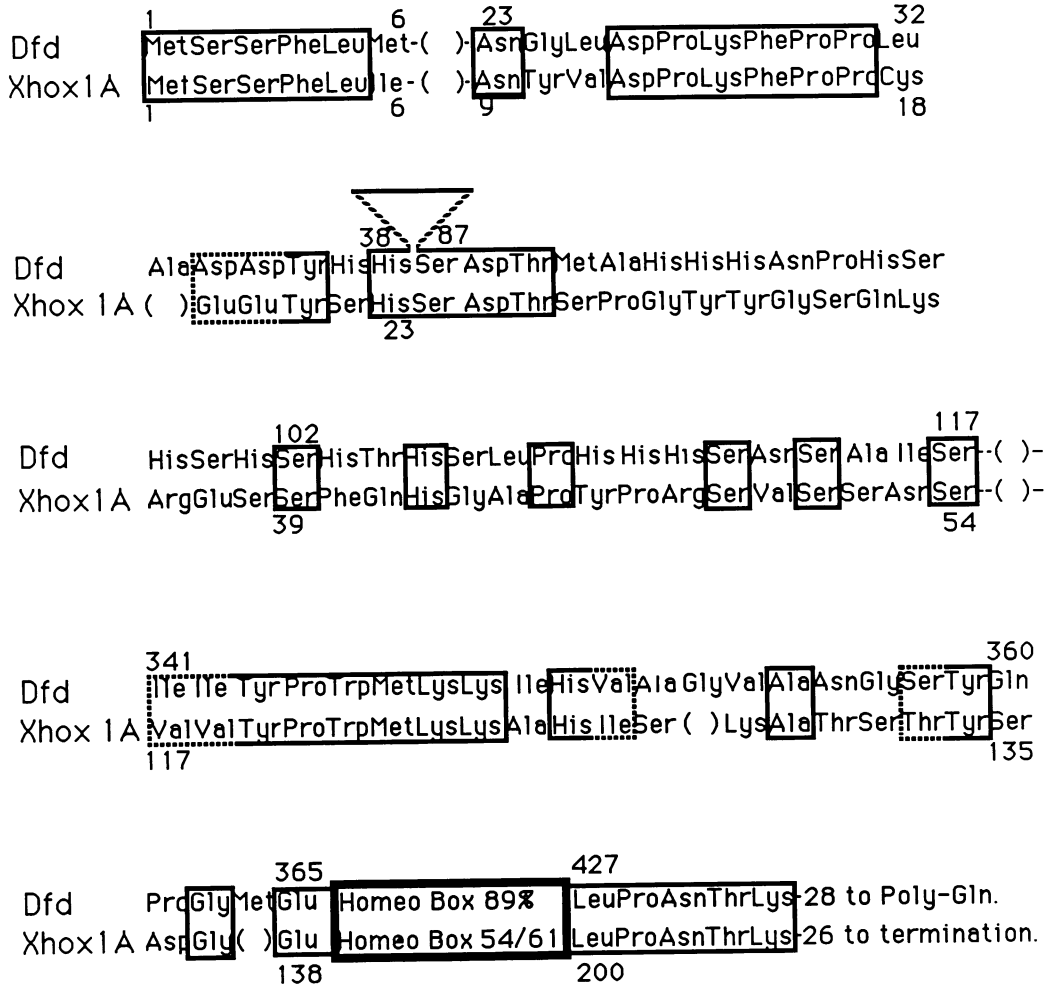

Fig. 9. A comparison of the amino acid sequence of putative protein products of the *Drosophila Deformed* gene and the *Xenopus Xhox-1A* gene. The numbering of the *Dfd* protein sequence starts with the first methionine residue in the long cDfd41 ORF. The numbering of the *Xhox-1A* protein sequence starts with the second methionine in the long ORF (Harvey *et al.*, 1986) which is more closely related to the consensus sequence for eukaryotic translation start sites (Kozak, 1985) and coincides with the first methionine of the *Dfd* protein. Brackets indicate gaps in both sequences introduced to obtain the best alignment. The inverted triangle denotes an insertion in the *Dfd* protein between positions 38 and 87.

proteins are shown aligned in Figure 9. The *Xhox-1A* and *Dfd* proteins differ greatly in their expected sizes, with *Xhox-1A* at 230 amino acids and *Dfd* at 586. However, as can be seen in Figure 8, much of the *Dfd* protein consists of very monotonous stretches of repetitive amino acid sequence, which are largely missing from homologous positions of the *Xhox-1A* protein. A comparison of *Deformed* and *Xhox-1A* with *Antennapedia* (Schneuwly *et al.*, 1986, the only other full length sequence extant from the Antennapedia class of the homeo box family) reveals much lower levels of similarity.

The extensive cross-homology between *Xhox-1A* and *Dfd* indicates that a *Dfd*-like gene has evolved and diverged from other homeo box genes before the evolutionary divergence between arthropod and vertebrate lineages. If the function of this gene has been similarly conserved on both lineages, then one would expect the *Xhox-1A* gene to specify and maintain a rostro−caudal positional identity, perhaps, as does *Dfd*, near the rostral end of the developing embryo. In the case of *Deformed*, as shown here, this positional determination is necessary for the proper development of structures from the mandibular and maxillary segments of the larval head. Interestingly, Harvey *et al.* (1986) have reported that *Xhox-1A* transcripts first appear during gastrulation of the frog embryo, a stage at which positional identities along the rostro-caudal axis are determined (reviewed in Slack, 1983).

Materials and methods

Embryonic phenotypes

Embryos of $Dfd^{RX1}/Df(3R)Scr$ for scanning electron microscopic study were collected and aged at 25°C on agar plates until they reached the desired stages of development. They were dechorionated in 3% chlorox for 2 min and rinsed well with distilled water. The vitelline membranes were removed and the embryos were rehydrated (Mitchison and Sedat, 1983). The embryos were fixed for 90 min in 4% glutaraldehyde in 0.1 M sodium cacodylate buffer then rinsed with 0.1 M sodium cacodylate. Post-fixing was in 1% OsO_4 in water overnight. After a thorough rinsing with distilled water, the fixed embryos were dehydrated through a graded series of ethanols. The embryos were dried by the critical point technique (Horridge and Tamm, 1969) using 100% ethanol as the transition fluid. The dried embryos were mounted on stubs with double-sided sticky tape, coated with a thin layer of 60:40 gold:palladium, and photographed with an ISI SS-40 operated at 5 or 10 kV. Embryos were scored for phenotype based on their head development. The following results were obtained:

Stage	Dfd^+	Dfd^-	% Dfd^-
6−9 h	129	37	29
9−12 h	113	33	23
12−15 h	132	34	26
18−21 h	149	41	27

Of the Dfd^- embryos in each stage, 50−60% exhibited strong phenotypes as seen in Figure 2a and b.

The embryos of $Dfd^{RX1}/Df(3R)Scr$ for cuticle preparations were collected and aged on agar plates at 25°C until they reached the desired developmental stage. Then they were fixed and mounted by the method of Van der Meer (1977) and viewed by phase contrast. Embryos were scored for phenotype based on the appearance of the cuticular structures in the head. The following results were obtained:

Stage	Dfd^+	Dfd^-	% Dfd^-
31−34 h	194	49	25
Dfd^- total	Cephalopharyngeal plate duplications		% Duplications
42	22		52

In situ hybridization to polytene chromosomes

Clone λ99 was nick-translated with biotinylated nucleotide and hybridized to polytene chromosomes of $Dfd^{RX1}/OreR$. The hybridizing sequence was detected by an immunoperoxidase method (Langer-Safer *et al.*, 1982) and the chromosomes were stained with Giemsa and photographed.

Primer extension

Labeled primer for the reverse transcriptase (RT) reaction was synthesized as described by Burke (1984) with minor modifications. 1 μg of ssDNA of a M13mp8 clone containing the 0.17-kb *Eco*RI fragment from cDfd41 was used as a template. The resulting double-stranded DNA was digested with *Taq*I, separated on a 2% agarose gel and a fragment of 150 bp was isolated. The fragment (2.5×10^4 c.p.m.) was heated for 5 min at 80°C and added to the annealing mix containing 10 μg of poly(A)$^+$ RNA from 6−12 h OregonR embryos, 10 μg of carrier tRNA and 20 units RNasin (Promega Biotec) in 20 μl of water. Annealing was performed at 56°C for 30 min. After the incubation, the following components were added: 5 μl of 10 × M-MLV RT buffer (BRL), 50 units of RNasin, nucleotides to final concentrations of 0.5 mM, 400 units of M-MLV RT, and water to 50 μl. Following the incubation, the DNA was ethanol precipitated, resuspended in 5 μl of water and treated with 2 μl of 10 ng/μl RNase for 30 min at 37°C. 2 μl of formamide plus dye was added, samples were boiled for 5 min in open tubes, and loaded on a 5% polyacrylamide gel.

S1 nuclease protection

A 1.1-kb *Eco*RI genomic fragment from λ100 was phosphatased with calf intestinal phosphatase (Boehringer Mannheim) and kinased with T4 polynucleotide kinase (BRL) according to Maxam and Gilbert (1980). The fragment was digested with *Hae*III (BRL), separated on 2% agarose and a 0.29-kb *Hae*III−*Eco*RI fragment known to overlap the 5′ end of cDfd41 was isolated. Approximately 3×10^5 c.p.m. of labeled fragment was ethanol precipitated with 10 μg of poly(A)$^+$ RNA from 6−12 h OregonR embryos, and 5 μg of carrier tRNA. The remainder of the hybridization and S1 nuclease protection protocol was as in Maniatis *et al.* (1982). The reaction products were separated on a 5% polyacrylyamide gel for 3 h at 70 W constant power. The gel was dried and autoradiographed at −80°C for 3 days.

Sequencing

The nucleotide sequence of genomic and cDNA clones was determined by the dideoxynucleotide sequencing procedure (Sanger *et al.*, 1977) using [^{35}S]dATP. Reaction products were separated on a 5% polyacrylamide gel containing 8 M urea.

Acknowledgements

We are grateful to Tom Jack, Mike Kuziora and Mark Barad for helpful comments on the manuscript, and Sue Pepin for help with sequencing. We also thank Barry Piekos for his knowledge of scanning electron microscopy and his patience. Tulle Hazelrigg kindly provided the Dfd^{RX1} chromosome, and Mike Levine catalyzed many lively discussions. This research was supported by grants to W.M. from the National Science foundation (DCB8501822) and the Searle Scholar Fund (86-B-112).

References

Akam,M.E. (1983) *EMBO J.*, **2**, 2075−2084.
Akam,M.E. and Martinez-Arias,A. (1985) *EMBO J.*, **4**, 1689−1700.
Beachy,P.A., Helfand,S.L. and Hogness,D.S. (1985) *Nature*, **313**, 545−551.
Bender,W., Akam,M., Karch,F., Beachy,P.A., Pfeifer,M., Spierer,P., Lewis, E.B. and Hogness,D.S. (1983) *Science*, **221**, 23−29.
Breathnach,R. and Chambon,P. (1981) *Annu. Rev. Biochem.*, **50**, 349−383.
Burke,J.F. (1984) *Gene*, **30**, 63−68.
Frederik,R.D. and Denell,R.E. (1982) *Int. J. Insect Morphol. Embryol.*, **11**, 227−233.
Garber,R.L., Kuroiwa,A. and Gehring,W.J. (1983) *EMBO J.*, **2**, 2027−2036.
Garcia-Bellido,A. (1977) *Am. Zool.*, **17**, 613−629.
Hafen,E., Levine,M. and Gehring,W.J. (1984) *Nature*, **307**, 287−289.
Harding,K., Wedeen,C., McGinnis,W. and Levine,M. (1985) *Science*, **229**, 1236−1242.
Harvey,R.P., Tabin,C.J. and Melton,D.A. (1986) *EMBO J.*, **5**, 1237−1244.
Hazelrigg,T. and Kaufman,T.C. (1983) *Genetics*, **105**, 581−600.
Horridge,G.A. and Tamm,S.L. (1969) *Science*, **163**, 817−818.
Hultmark,D., Klemenz,R. and Gehring,W. (1986) *Cell*, **44**, 429−438.
Joyner,A.L., Kornberg,T., Coleman,K.G., Cox,D.R. and Martin,G.R. (1985) *Cell*, **43**, 28−37.
Jürgens,G., Lehman,R., Schardin,M. and Nusslein-Volhard,C. (1986) *Wilhelm Roux's Arch. Dev. Biol.*, **195**, 359−377.
Karch,F., Weiffenbach,B., Bender,W., Pfeifer,M., Duncan,I., Celneken,S., Crosby,M. and Lewis,E.B. (1985) *Cell*, **43**, 81−96.
Kaufman,T.C. (1978) *Genetics*, **90**, 579−596.

Kaufman,T.C. (1983) In *Time, Space and Pattern in Embryonic Development.* Alan R.Liss, NY, pp. 365–383.
Kozak,M. (1986) *Cell*, **44**, 283–292.
Kuroiwa,A., Kloter,U., Baumgartner,P. and Gehring,W.J. (1985) *EMBO J.*, **4**, 3757–3764.
Langer-Safer,P.R., Levine,M. and Ward,D.C. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 4381–4385.
Laughon,A., Carroll,S.B., Storfer,E.A., Riley,P.D. and Scott,M.P. (1985) *Cold Spring Harbor Symp. Quant. Biol.*, **50**, 253–262.
Lewis,E.B. (1978) *Nature*, **276**, 565–570.
Lewis,R.A., Wakimoto,B.T., Denell,R.E. and Kaufman,T.C. (1980) *Genetics*, **95**, 383–397.
Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) *Molecular Cloning. A Laboratory Manual.* Cold Spring Harbor Laboratory Press, NY.
Maxam,A.M. and Gilbert,W. (1980) *Methods Enzymol.*, **65**, 499–560.
McGinnis,W., Levine,M., Hafen,E., Kuroiwa,A. and Gehring,W.J. (1984) *Nature*, **308**, 428–433.
Mitchison,T.J. and Sedat,J. (1983) *Dev. Biol.*, **99**, 261–264.
Mount,S. (1982) *Nucleic Acids Res.*, **10**, 459–471.
Poole,S.J., Kauvar,L.M., Brees,B. and Kornberg,T. (1985) *Cell*, **40**, 37–43.
Proudfoot,N.J. and Brownlee,G.G. (1976) *Nature*, **262**, 211–214.
Regulski,M., Harding,K., Kostriken,R., Karch,F., Levine,M. and McGinnis,W. (1985) *Cell*, **43**, 71–80.
Sanchez-Herrero,E., Vernos,I., Marco,R. and Morata,G. (1985) *Nature*, **313**, 108–113.
Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463–5467.
Schneuwly,S., Kuroiwa,A., Baumgartner,P. and Gehring,W.J. (1986) *EMBO J.*, **5**, 733–739.
Scott,M.P., Weiner,A.J., Hazelrigg,T.I., Polisky,B.A., Pirotta,V., Scalenghe,F. and Kaufman,T.C. (1983) *Cell*, **35**, 763–776.
Sheridan,P.R., Scott,J., Dixon,R., Venkataragharan,R., Kuntz,I.D. and Scott, K.P. (1985) *Biopolymers*, **24**, 1995–2023.
Slack,J.M.W. (1983) *Determinative Events in Early Development.* Cambridge University Press, Cambridge.
Strasburger,M. (1932) *Z. Wiss. Zool.*, **140**, 539–649.
Struhl,G. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 7380–7384.
Struhl,G. (1983) *J. Embryol. Exp. Morphol.*, **76**, 297–331.
Turner,R.F. and Mahowald,A.P. (1979) *Dev. Biol.*, **68**, 96–109.
Van der Meer,J.M. (1977) *Drosophila Inf. Serv.*, **52**, 160.
Vogt,M. (1947) *Biol. Zentralbl.*, **66**, 81–105.
Wakimoto,B.T., Turner,F.R. and Kaufman,T.C. (1984) *Dev. Biol.*, **102**, 147–172.
Wharton,K.A., Yedvobnick,B., Finnerty,V.G. and Artavanis-Tsakonas,S. (1985) *Cell*, **40**, 55–62.
White,R.A.H. and Wilcox,M. (1985) *EMBO J.*, **4**, 2035–2043.

Received on October 31, 1986; revised on December 23, 1986

Note added in proof

The conceptual Deformed protein also has extensive homology to the HHo.c13 gene product of humans, recently cloned with homeo box probes by Mavilio,F. *et al.* (1986) *Nature*, **324**, 664–668. The figure below shows in schematic form the regions of homology between the *Drosophila* (*Dfd*), *Xenopus* (*Xhox-1A*), and human proteins. The black lines in the schematic Xhox1A and c13 protein sequences represent identities with indicated regions of the *Dfd* protein. The numbers above the *Xenopus* sequence indicate amino acid residues from the amino terminus.