Cell, Vol. 47, 735–746, December 5, 1986, Copyright © 1986 by Cell Press

Structure of the Segmentation Gene *paired* and the Drosophila PRD Gene Set as Part of a Gene Network

Gabriella Frigerio, Maya Burri, Daniel Bopp, Stefan Baumgartner, and Markus Noll
Department of Cell Biology
Biocenter of the University
Klingelbergstrasse 70
CH–4056 Basel, Switzerland

Summary

The sequence of *paired*, a pair-rule gene required for segmentation in Drosophila, is presented. A search for genes with domains homologous to the *paired* gene was initiated and three homologues from a set of 12 were characterized with respect to temporal or spatial expression and sequence homologies. All four are transcribed in early development, one in the oocyte and during cleavage stages in the form of a gradient. In addition to the *prd*-specific his-pro repeat, some of the 12 genes contain M-repeats and two new types of homeo boxes not detectable by hybridization with the two known classes of homeo boxes. The observed linking of gene sets through combinations of homologies coding for protein domains is consistent with a general network concept of gene action.

Introduction

The genetic analysis of the bithorax complex in Drosophila led Lewis to the conclusion that complex genes with related functions in development evolved from a smaller number of ancestral genes by gene duplication and subsequent specialization (Lewis, 1978). It has become increasingly clear during recent years that proteins and their genes may be further subdivided into functional domains that are subject to recombination and independent assortment (Gilbert, 1978; Gilbert et al., 1986). The first concept implies that functions that are retained or modified are expected to appear as partly conserved DNA sequences of either a *cis*-acting regulatory nature or coding for protein domains. Genes sharing such a homologous DNA or protein domain we define as a gene set. Independent assortment of domains has the interesting consequence that a particular gene may belong to more than one set of genes. This is a novel feature not encountered in gene families defined by the conservation of the entire gene and has been used by Doolittle to characterize superfamilies (Doolittle, 1981). An illustration of this concept whose general significance has not been recognized is the occurrence of a homeo box associated with some genes of the M-repeat set (McGinnis et al., 1984a; Regulski et al., 1985; Kuroiwa et al., 1985; Laughon et al., 1985). In its generalized form, the concept predicts that any multidomain gene shares sequences with several gene sets. Systematic searches may therefore be initiated to find these sets. Because of our interest in development, we applied this approach to the recently isolated segmentation gene *paired* (*prd*) of Drosophila (Kilchherr et al., 1986) and found the PRD gene set. In addition, we indeed discovered that some of its members contain other homologies, such as M-repeats and two new types of homeo boxes.

Results

Sequence of the *prd* Gene

The *prd* gene was sequenced using the strategy described in Experimental Procedures (Figure 10). The following features of the *prd* gene sequence are of interest (Figure 1). The longest open reading frame starts with the ATG at position 497 and codes for a possible protein of 613 amino acids terminated by an amber codon at position 2692. All three reading frames upstream of the ATG are closed. The termination signal is followed by 330 or 331 bases of untranslated trailer sequence comprising the canonical poly (A) addition sequence AATAAA, 28 or 29 nucleotides from its end. The open reading frame is interrupted by a 356 bp intron after the first 22 amino acids. The 5′ end of the *prd* transcript has not yet been determined. The following points suggest, however, that it is located between the PvuII site, at the beginning of the sequence shown in Figure 1, and the 5′ end of the c7340.4 cDNA at position 252. The length of the *prd* transcript as determined by Northern analysis is about 2.5 kb (Kilchherr et al., 1986) and is similar to the cDNA distance of 2418 nucleotides between the 5′ end of c7340.4 and the poly (A) tail. Furthermore, no transcript is detected upstream of the PvuII site by Northern analysis.

Several cDNA sequences exhibit various polymorphisms. Most of them represent third-base changes and do not alter the amino acid sequence. Two of them, however, result in an amino acid change, one at position 1342 changing an ala to a thr, the other at 1509 altering a phe into an ile (Figure 1).

Molecular Lesion of the Mutant *prd*$^{2.45.17}$

We have shown that the X-ray-induced *prd* mutant *prd*$^{2.45.17}$ (which was a kind gift from C. Nüsslein-Volhard) exhibited a 1.1 kb insertion into the *prd* gene. Northern analysis revealed a *prd* transcript that was 1.1 kb longer than the wild-type *prd* mRNA and hence suggested that the insertion had occurred into an exon or into an intron in such a way that its splicing became abnormal and produced a transcript that was 1.1 kb longer than the wild-type *prd* mRNA (Kilchherr et al., 1986). We have sequenced part of a 1.5 kb genomic SmaI fragment of *prd*$^{2.45.17}$ which contained the entire 1.1 kb insertion (see Figure 3 of Kilchherr et al., 1986). As illustrated in Figure 2, in *prd*$^{2.45.17}$, five base pairs of the second exon are deleted between 1064 and 1070 and replaced with a 1.1 kb insertion of which the sequences at the two ends are shown. The inserted DNA occurs well over a hundred times in the entire genome and is scattered over all four chromosomes (Kilchherr et al., 1986). As we have only se-

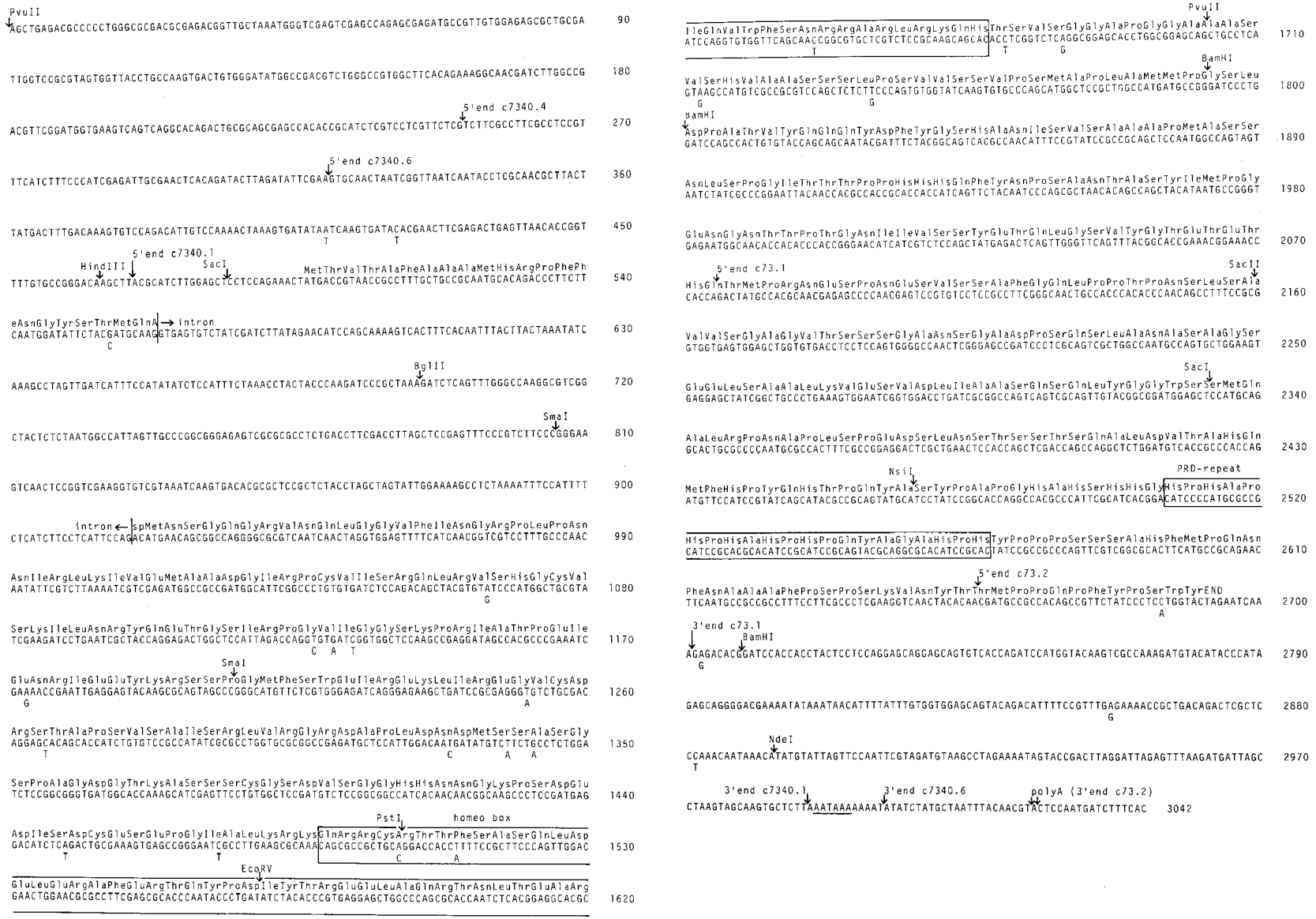

Figure 1. DNA Sequence of the *prd* Gene and Corresponding Amino Acid Sequence of the Putative *prd* Protein

The DNA sequence between the PvuII site at the 5′ end (Figure 10) and the 3′ end of the *prd* transcript is numbered from 1 to 3042. Variations in the nucleotide sequence found in one of the sequenced cDNAs are indicated below; the amino acid sequence corresponding to the longest open reading frame is shown above the main sequence. The boundaries of a 356 bp intron (564–919) are indicated by vertical lines. The position of the 5′ and 3′ ends of the various cDNAs, including the site of polyA addition, and of the cleavage sites of some restriction endonucleases are indicated by vertical arrows. The poly (A) addition signal is underlined. The sequences corresponding to the homeo box (1489–1668) (Figures 8 and 9) and the PRD-repeat (2506–2568) (Figure 6) are boxed.

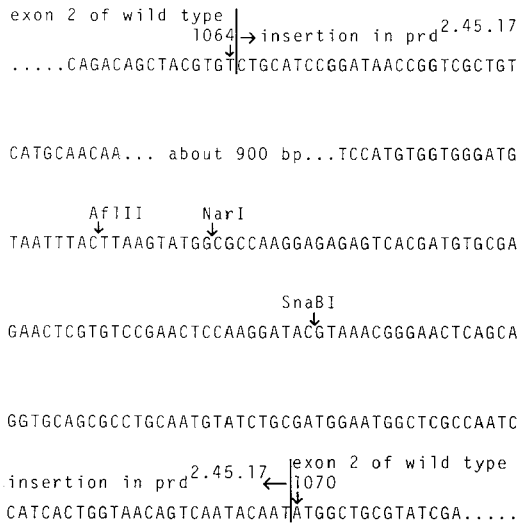

Figure 2. DNA Sequence Surrounding the Site of Insertion into exon 2 in *prd*$^{2.45.17}$

In the mutant *prd*$^{2.45.17}$, five base pairs are deleted between position 1064 and 1070 of the wild-type sequence (Figure 1) and replaced with a 1.1 kb insertion. The mutant DNA sequence surrounding the site of insertion into exon 2 of the *prd* gene is shown.

quenced a small portion of the inserted DNA element, we do not know whether it is a member of one of the characterized middle repetitive sequence families (a restriction map published by Kilchherr et al., 1986, has shown that it contains four PvuII, three PstI, and one SalI restriction site). Nor could we detect in the sequenced portions of the two ends any evidence for a direct or an inverted repeat.

Isolation of 11 Loci Containing DNA Sequences Homologous to the 3′ End of the *prd* Coding Region

We searched the Drosophila genome for sequences homologous to the *prd* transcript with two cDNA probes. One cDNA, c73.2, corresponded to the 3′ end of the *prd* mRNA coding for the last 11 amino acids and the entire 3′ trailer sequence; the other cDNA, c73.1, consisted of about 600 bp and coded for the last 205 amino acids of the long open reading frame of 613 amino acids of *prd*, as indicated in Figure 1 and Figure 3a. Both cDNAs were hybridized at reduced stringency to whole-genome Southern blots of two different D. melanogaster strains. While the 3′ trailer cDNA did not hybridize any sequences other than those originating from *prd* (not shown), the 600 bp cDNA revealed five fairly strong and six faint bands (Figure 3b) in addition to the 9.6 kb EcoRI fragment of *prd*. Since nearly all these fragments were shorter than the 9.6 kb *prd* fragment and showed up in both Drosophila strains, they could not be due to partial EcoRI-digests nor was it likely that they originated from polymorphisms in the *prd* gene region. Therefore, they probably represented sequences homologous to *prd* elsewhere in the genome.

To examine these *prd*-related loci more closely, a genomic EMBL4 library was screened with the 600 bp *prd* cDNA at reduced stringency of hybridization. To avoid reisolation of the *prd* locus, duplicate filters of the library were hybridized at normal stringency with probes flanking the 9.6 kb EcoRI fragment of *prd* on either side. Rescreening showed that from 125 phages picked (PRD 1–PRD 125), about 5% still contained the *prd* gene. The remaining phages could be divided into 11 groups from their cross-hybridization patterns. These correspond to 11 different loci containing the sequences homologous to *prd*, tentatively called PRD genes 1–11. EcoRI maps of these loci, established by whole-genome Southern analysis and, in some cases, by mapping of the phage DNAs containing the cloned regions, are depicted in Figure 3c.

To determine which EcoRI fragment contained the sequence homologous to *prd*, EcoRI fragments of representative phages of each of the 11 loci were resolved in an agarose gel (Figure 4a), blotted to nitrocellulose, and hybridized at reduced stringency with the 600 bp *prd* cDNA. As evident from Figure 4b, the signal intensities of the 11 homologies exhibit considerable variation (about 15-fold) but are all rather strong considering the short exposure time used (Figure 4b). This implies either that the homologies are very good or that the homologous sequence consists of a repetitive element. Upon hybridization to polytene chromosomes, subclones containing the homologous sequences of the 11 cloned regions revealed only one signal at the chromosomal bands indicated in Figure 3c.

Regions Containing Homologous Sequences Are Transcribed during Early Development

Differential transcript mapping (Frei et al., 1985; Kilchherr et al., 1986) was used to gain an idea of the temporal expression pattern of these genes during early development. When blots from gels (Figure 4a) were hybridized with ^{32}P-labeled cDNA of poly(A)$^+$ RNA isolated from various developmental stages (follicles; 0–4 hr-, 4–8 hr-, 8–12 hr-old embryos; and first instar larvae), it was found that all except one region of the 11 loci produced transcripts detectable in oocytes or embryos up to about the 8th–12th hr of development at 25°C. No transcripts have been observed by this method in PRD gene 2 and only very weak transcription in the EcoRI fragments containing the homologous sequences of genes 5, 9, and 10 during early development.

In order to assess the functional significance of the observed homologies, we wanted to know first, whether they coded for conserved amino acid sequences, second, whether they were transcribed selectively during early development, and third, whether their transcripts exhibited a specific distribution in oocytes or young embryos. All of these questions are discussed below in experiments that examine PRD genes 3, 4, and 7 in greater detail.

Northern analysis using genomic subclones or the corresponding cDNA that contained the *prd* homology was performed since differential transcript mapping does not resolve different transcripts that are coded on the same restriction fragment. The results confirmed that transcripts of the PRD genes 3, 4, and 7 predominate before 8–12 hr of embryonic development (not shown).

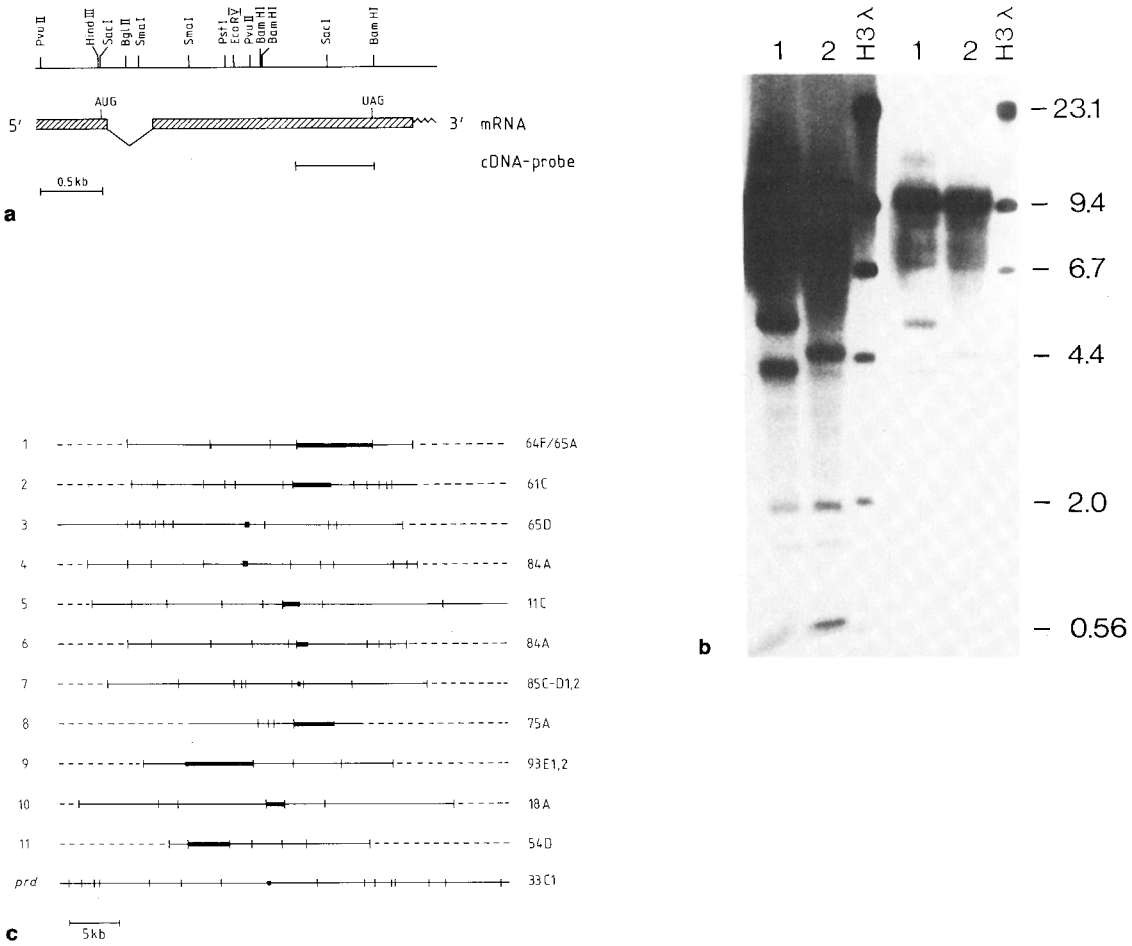

Figure 3. Isolation of Genomic Clones Containing Sequences Homologous to the 3′ End of the *prd* Gene

(a) Restriction map of the *prd* gene indicating position of a 0.6 kb *prd* cDNA probe. The beginning and end of the longest open reading frame and the positions of an intron and of the 0.6 kb *prd* cDNA used as probe in (b) are indicated.

(b) Genomic Southern analysis at reduced stringency of hybridization with 0.6 kb *prd* cDNA. Genomic DNA (10 μg each) of two Drosophila strains (1: *odd*$^{1.36.18}$ *b pr cn sca l CyO*; 2: *b pr l b pr*) was digested with EcoRI; the fragments were separated by electrophoresis in a 0.6% agarose gel in TBE-buffer (Maniatis et al., 1982) and transferred bidirectionally to two nitrocellulose filters (Smith and Summers, 1980), one of which was hybridized at reduced stringency (McGinnis et al., 1984a) for two days with a 0.6 kb *prd* cDNA probe labeled by nick-translation (Rigby et al., 1977) with α-^{32}P-dATP (3000 Ci/mmole). Two autoradiograms after exposures for one day (left panel) and four hr (right panel) to reveal both weakly and strongly hybridizing EcoRI fragments are shown. For size calibration, marker DNA fragments (H3λ) of known lengths (indicated in kb at the right) are shown in the right lane of each panel.

(c) EcoRI restriction maps of 11 genomic regions containing sequences homologous to 0.6 kb *prd* cDNA. The sequences homologous to the 0.6 kb *prd* cDNA have been mapped to the regions represented by solid bars. The cytological locations of each cloned region on polytene chromosomes, indicated at the right, were determined by in situ hybridizations to salivary gland chromosomes of biotinylated probes according to Langer-Safer et al. (1982). At the bottom, the *prd* gene region at 33C1 (Kilchherr et al., 1986) is shown. EcoRI restriction maps of the 11 isolated regions containing sequences homologous to the 0.6 kb *prd* cDNA were established from genomic clones and whole-genome Southern analysis with the genomic clones.

A Gradient of PRD Gene 4 Transcripts

Particularly interesting was PRD gene 4, whose transcripts had already appeared in oocytes. In situ hybridization to tissue sections with a cDNA probe of PRD gene 4 revealed a striking distribution of transcripts (Figure 5). In ovaries, transcripts are found in the nurse cells and accumulate at particularly high concentrations at the anterior end of the oocytes (Figures 5a and 5b). During cleavage stages of the embryo, the concentration of transcripts remains highest in the most anterior portion and falls off in a concentration gradient along the anteroposterior axis, reaching background behind the anterior third of the embryo (Figures 5c and 5d). While transcripts are reduced during syncytial blastoderm, they still predominate in the cortical cytoplasm (Figures 5e and 5f). This phenomenon of transcript guidance to the cortical cytoplasm has been observed for the zygotically expressed segmentation genes *fushi tarazu* (Hafen et al.,

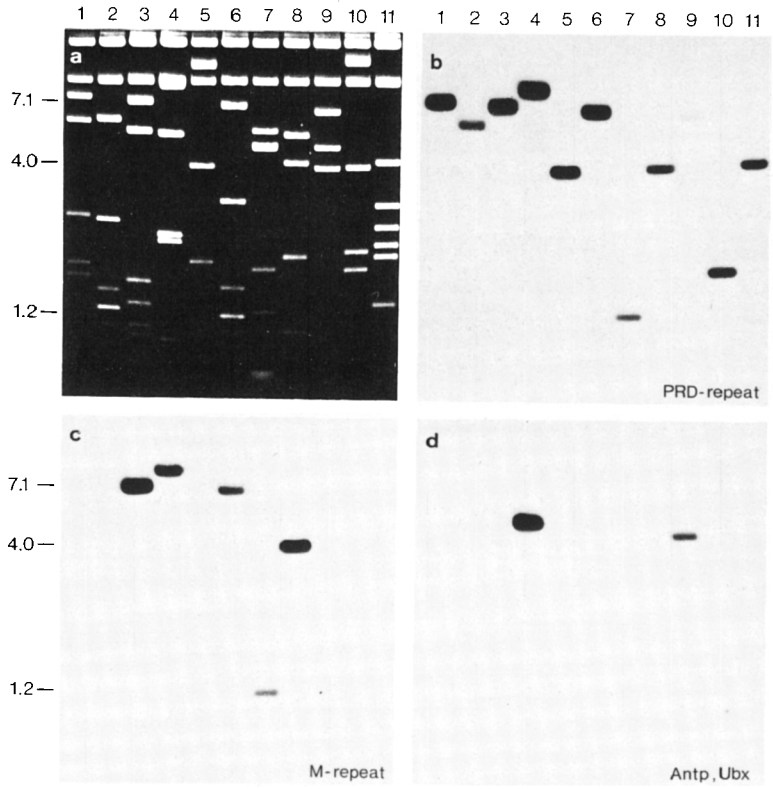

Figure 4. Association of PRD-Repeats with M-Repeats and Homeo Boxes as Examined by Phage Southerns

EcoRI digests of the following phage DNAs are shown in lanes 1–11 (containing the PRD genes 1–11): PRD 22, PRD 39, PRD 17, PRD 31, PRD 54, PRD 24, PRD 21, PRD 68, PRD 76, PRD 89, and PRD 108, after staining with ethidium bromide (a), and after autoradiography of Southern transfers hybridized at reduced stringency with 0.6 kb *prd* cDNA (b), an M-repeat probe (c), or an *Antp*- and *Ubx*-homeo box probe (d). For size calibration, at the left margin the lengths of a few DNA fragments are indicated. DNA from one phage of each of the 11 cloned regions (0.66 μg) was digested with EcoRI; the fragments were separated on a 0.6% agarose gel, stained with ethidium bromide, and transferred to nitrocellulose filters. The filters were hybridized for 17 hr at reduced stringency with nick-translated DNA fragments of the 0.6 kb *prd* cDNA, of the M-repeat (SmaI–PvuII fragment of p903 [McGinnis et al., 1984a]), or of the *Antp*- and *Ubx*-homeo boxes (BamHI–PvuII fragment of p903 [McGinnis et al., 1984a] and the ClaI–BglI fragment of p96 [McGinnis et al., 1984a]), as described for Figure 3b. The specific activity of all probes was 1–2 × 10^8 dpm/μg DNA, and hybridization occurred at 10^6 dpm/ml. The filters were washed in 2× SSC, 0.1% SDS for 15 min at room temperature and for 30 min at 50°C, and autoradiographed for 1 hr (b), 2 hr (c), or 15 hr (d). The signal of the 0.6 kb *prd*-cDNA after hybridization to its *prd* cognate DNA was nearly ten times that of the strongest band in (b).

1984), *hairy* (Ingham et al., 1985), and *prd* (Kilchherr et al., 1986). After formation of the cellular blastoderm, PRD gene 4 RNA was not detected on tissue sections (not shown).

DNA Homology Codes for a Protein Domain Consisting of a Histidine-Proline Repeat

cDNAs of genes 4 (c53.46.1) and 7 (c21.12.1) and a genomic subclone of gene 3 (PRD 25), all containing the *prd* homologies, were sequenced and compared to the sequence of the 600 bp *prd* cDNA. The aligned homologous sequences resulting in a maximum correspondence are shown in Figure 6. The homology consists, with minor deviations, of the hexanucleotide CA$\frac{T}{C}$CCG repeated many times over (PRD-repeat), which explains the strong signals of hybridization in Figure 4b. The reading frame of PRD gene 4 was derived from cDNA sequences comprising the longest open reading frame and was shown to be in phase with that of *prd* as indicated in Figure 6. We sequenced only the homologous regions of genes 3 and 7 without determining their reading frames, assuming that reading frames are used that result in amino acid sequences homologous to *prd* and gene 4 (Figure 6). The results in Figure 6 show a conserved pattern of alternating histidine and proline amino acids. This histidine–proline repeat is imperfect and occasionally interrupted by the deletion of a histidine or proline or by replacement with another amino acid. Allowing for such changes, the his–pro repeats extend over 20 to 30 amino acids.

Are PRD-Repeats Associated with Homeo Boxes or M-Repeats?

For the reasons outlined in the introduction, a search was initiated for the presence in the PRD-gene set of the homeo box (McGinnis et al., 1984a; Scott and Weiner, 1984) and the M- (or opa-) repeat (McGinnis et al., 1984a; Laughon et al., 1985; Wharton et al., 1985), the two other domains known to be associated with genes active in development (McGinnis et al., 1984a; Scott and Weiner, 1984; Regulski et al., 1985; Kuroiwa et al., 1985; Laughon et al., 1985; Wharton et al., 1985). Blots of the gel shown in Figure 4a ("phage Southerns") were hybridized with ^{32}P-labeled M-repeat (Figure 4c) or with *Antp* and *Ubx* homeo box probes (Figure 4d). While five of the fragments containing the PRD-repeat also hybridize with the M-repeat (see Figures 4b and 4c), only two fragments were detected with the homeo box probe (Figure 4d). The presence of an M-repeat in PRD genes 3, 4, and 7 was further confirmed by DNA sequencing (Figure 7). Thus at least some of the PRD genes belong to the set of M-repeat genes as well.

The strong homeo box signal in Figure 4d originated from a 5.2 kb EcoRI fragment located next to the fragment containing the PRD-repeat (compare Figures 4b and 4d with the map in Figure 3c). To test whether both the homeo box domain and the PRD-repeat were on the same transcript, cDNAs were isolated with both the 5.2 kb and 8.7 kb EcoRI fragment of the region containing PRD gene 4 (Figure 3c). Analysis of the cDNAs, and Northern analysis

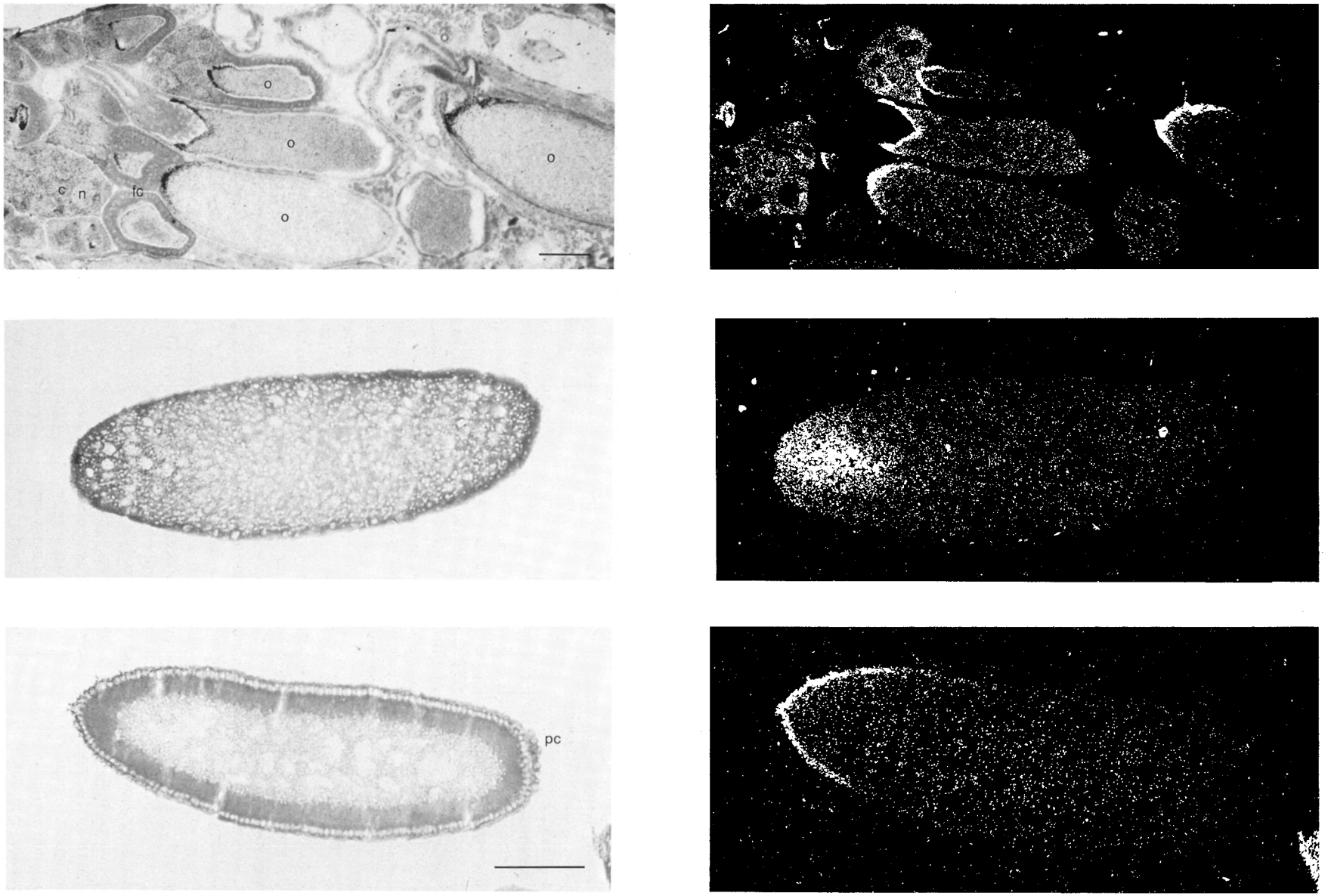

Figure 5. Localization of PRD Gene 4 Transcripts in Ovaries and in Cleavage Stage and Syncytial Blastoderm Embryos

Photomicrographs of tissue sections through a female abdomen (a, b), a cleavage stage embryo (c, d), and an early stage 14 syncytial blastoderm embryo (e, f) (Foe and Alberts, 1983) were taken under phase-contrast (left) or dark-field illumination (right). Tissue sections are oriented with their anterior end to the left and their dorsal side up. In (a), some oocytes (o), follicle cells (fc), nurse cell nuclei (n), and cytoplasm (c) and in (e) the pole cells (pc) are marked. Note that in ovaries nearly all hybridization is in the cytoplasm of the nurse cells and at the anterior pole of the oocytes. Magnifications, represented by horizontal bars indicating a length of 0.1 mm in (a) and (e), are the same in (a) and (b) and in (c)–(f). Embryonic tissue sections were prepared, hybridized in situ with ^{3}H-labeled 2.6 kb PRD gene 4 cDNA, c53.46.6, and exposed for 28 days according to Hafen et al. (1983). Female abdomens were sectioned in the same way as the pretreated embryos.

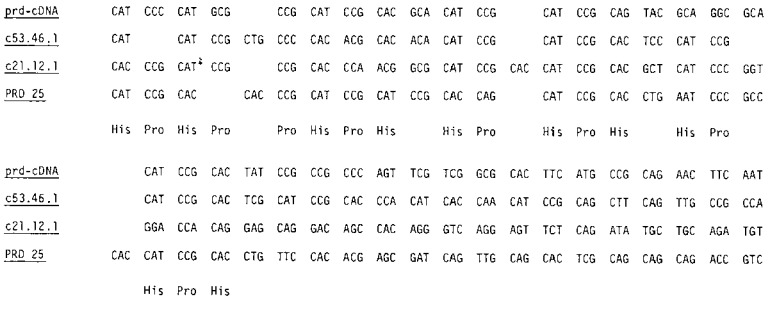

Figure 6. Homologous DNA Sequences of the PRD Gene Set

The DNA sequences of PRD gene 4 (c53.46.1), PRD gene 7 (c21.12.1), and PRD gene 3 (PRD 25) that are homologous to the 0.6 kb *prd* cDNA are compared with the corresponding sequence of the *prd* gene (*prd*-cDNA). Below the DNA sequences, His and Pro are indicated at positions where they correspond to the most frequent amino acid. The reading frames have been determined as indicated in the text. Note that the His–Pro repeat of c53.46.1 is longer than in the three other genes.

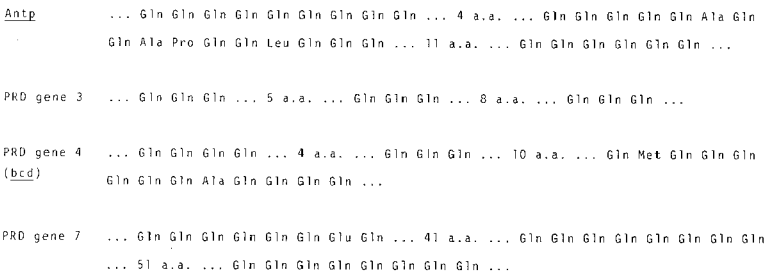

Figure 7. M-Repeats of PRD Genes 3, 4, and 7

The clusters of oligo-Gln (M-repeat) in the products of PRD genes 3, 4, and 7 are shown. The amino acid sequences were derived from the corresponding DNA sequences. For PRD gene 3, only part of its genomic sequence has been determined, hence it is possible that its M-repeat is longer than indicated here. For comparison, the M-repeat, first discovered in the *Antp* gene (McGinnis et al., 1984a), is shown (Schneuwly et al., 1986).

with cDNA probes, showed that two different transcripts are encoded on the two genomic fragments. Thus the 8.7 kb EcoRI fragment gave rise to a 2.7 kb transcript corresponding to the PRD gene 4 mRNA, while the 5.2 kb fragment produced a 1.7 kb transcript that hybridized with the homeo box. The weak homeo box signal in the region of PRD gene 9 (Figure 4d) is not located on the same EcoRI fragment as the PRD-repeat but rather separated from it by at least 4.0 kb of a genomic EcoRI fragment (see Figure 3c). Since differential transcript mapping indicated that this region is transcribed only at a very low level during 4–12 hr of embryonic development, no attempts were made to clarify whether the PRD-repeat and the homeo box are linked in one transcript. Hence, although at least two PRD genes were located close to homeo box genes, in no case was it possible by hybridization to show that a PRD gene was a member of the homeo box gene set as well.

PRD Gene 4 Is Located between *zen* and *Dfd*

In situ hybridization to salivary gland chromosomes of the phage DNA containing PRD gene 4 revealed a signal at 84A. Hence, PRD gene 4 was probably located to the left of *Antp* (Lewis et al., 1980). Comparison of the EcoRI map of the region containing gene 4 (Figure 3c) with that of the *Antp* complex at 84A showed that it was identical to that of a region between 20 and 50 kb to the left of *Dfd* that had been cloned previously (Scott et al., 1983). Judging from the genetic map (Scott et al., 1983; Wakimoto et al., 1984), the gene containing the homeo box could be *zerknüllt* (*zen*). This assumption is supported by a recent report correlating the temporal and spatial patterns of transcripts originating from the fragment containing the homeo box (Doyle et al., 1986) with the gastrulation-defective phenotype of *zen* (Wakimoto et al., 1984). PRD gene 4, however, is not identical with *zen* because *zen* is only expressed zygotically (Wakimoto et al., 1984). A much more likely candidate for gene 4 is *bicoid* (*bcd*), which has been characterized and mapped to this region (Frohnhöfer and Nüsslein-Volhard, 1986). The phenotype of *bcd* exhibits a maternal effect and shows abnormal head development (Frohnhöfer and Nüsslein-Volhard, 1986), which is also in agreement with the temporal and spatial distribution of PRD gene 4 transcripts (Figure 5).

prd and PRD Gene 4 Each Contain a PRD-Repeat and a Homeo Box

When the DNA sequences coding for full-length transcripts of *prd* (Figure 1) and PRD gene 4 were compared, another region of homology that differed from the PRD-repeat was discovered. Closer inspection of size (180 bp) and sequence revealed two new types of homeo boxes (Figure 8). As already mentioned, the extent of homology with *Antp*-, *Ubx*-like homeo boxes (Scott and Weiner, 1984; McGinnis et al., 1984b; Regulski et al., 1985; Kuroiwa et al., 1985; Laughon et al., 1985) is insufficient (42%–50%) for detection by low stringency hybridization (Figure 4d). Neither would we expect the *prd* and PRD gene 4 homeo boxes to hybridize with the other known class of homeo boxes, *en* (Poole et al., 1985; Fjose et al., 1985), because of an equally low degree of DNA homology (46%–50%). Moreover, the DNA homology between the *prd* and PRD gene 4 homeo boxes is not very close (48%), hence neither would be detected by hybridization with the other in a routine library screen, yet hybridization between the two is detectable in the more sensitive assay of a phage

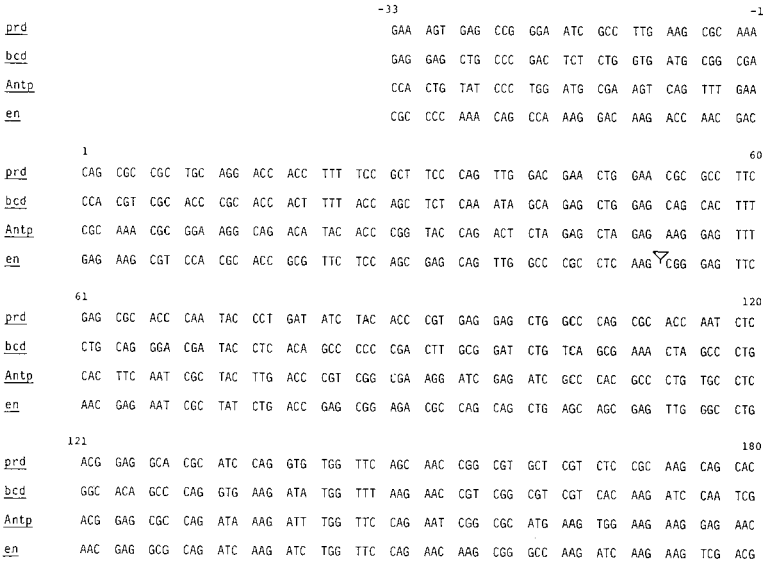

Figure 8. DNA Sequences of the *prd*, *bcd*, *Antp*, and *en* Homeo Boxes

The DNA sequences (5'-3') from the homeo box regions of the *prd* and *bcd* (PRD gene 4) genes are compared to those of the *Antp* (Kuroiwa et al., 1985) and *en* genes (Poole et al., 1985). The position of an intron in the *en* homeo box (Poole et al., 1985; Fjose et al., 1985) is indicated by a triangle. The DNA sequence of the *bcd* gene has been determined from nearly full-length cDNAs as well as from the genomic sequence.

Southern (not shown). In the more meaningful comparison of the conservation of the amino acid sequence (Figure 9), we find homologies of 37% between the *prd* and *Antp*, 42% between the PRD gene 4 and *Antp*, and 38% between the *prd* and PRD gene 4 homeo boxes. It is apparent from Figure 9 that the *prd* and PRD gene 4 homeo boxes have undergone nonconservative amino acid changes at many positions that are strongly conserved in all *Antp*-, *Ubx*-like homeo boxes. Conversely, about an equal number of positions (17) appear to be strictly conserved or to tolerate only conservative amino acid changes in all presently known homeo boxes of Drosophila (Figure 9). Most importantly, the results depicted in Figure 9 further support our hypothesis that members of the PRD gene set belong to other sets of genes, particularly to the set of homeo box genes. We note that unlike the *en*-type (Poole et al., 1985; Fjose et al., 1985), the two new types of homeo boxes are each contained in a single exon.

It is interesting to discuss the *prd* and PRD gene 4 homeo boxes in terms of the helix-turn-helix model proposed for DNA-binding proteins (Anderson et al., 1981; McKay and Steitz, 1981; Pabo and Lewis, 1982; Matthews et al., 1982; Sauer et al., 1982; Weber et al., 1982; Pabo and Sauer, 1984). The model is based on crystallographic, genetic, and biochemical analyses of the interaction of the λDNA-binding proteins *cI* and *cro*, and of the E. coli catabolite-activating protein with their operator DNAs (reviewed in Pabo and Sauer, 1984). It proposes that the two α-helices linked by a tight turn specifically bind to DNA sequences. The second α-helix provides the binding specificity by recognizing the bases in the major groove and is held in position by the first α-helix interacting with the DNA backbone. The model has been applied to a sequence of 20 amino acids within the homeo domain suggested to form a helix-turn-helix structure as indicated in Figure 9 (Laughon and Scott, 1984; Laughon et al., 1985). A comparison of the corresponding regions of the *prd* and PRD gene 4 homeo domains with the salient features of the helix-turn-helix model shows that several amino acids of the second α-helix at positions responsible for DNA-binding specificity differ from those in *Antp*-, *Ubx*-like homeo domains (Figure 9). Thus the lys-ile amino acid pair, which is conserved in all other homeo domains, is changed to gln-val in *prd*, and PRD gene 4 carries a thr where normally glu is found. In addition, both *prd* and PRD gene 4 have an ala in common with *en* and *inv* at a position of helix 2 where all other homeo domains exhibit an arg. Hence, it appears that the *prd* and PRD gene 4 proteins might bind to different DNA sequences that are not recognized by the other homeo domains or that the two proteins exhibit altered affinities for the same DNA sequence. In contrast to these differences, the *prd* and PRD gene 4 proteins are conserved at positions thought to be important for the overall geometry of the domain comprising the two α-helices. The only exception represents the ser at position 5 of the first helix of PRD gene 4 where normally an ala resides that interacts with the ile or val at position 4 of helix 2 (Pabo and Sauer, 1984). The same deviation has been noted for the *en*- and *inv*-domains (Laughon et al., 1985).

Whole-genome Southern analysis with the *prd* and PRD gene 4 homeo boxes as probes uncovered at least one DNA sequence with a good homology to each homeo box elsewhere in the genome. Isolation of these sequences and preliminary sequence analysis indicate that the *prd* and PRD gene 4 homeo boxes are highly conserved at other genomic loci. Thus the *prd* and PRD gene 4 homeo boxes belong to two new classes of homeo boxes. Most interestingly, one of the regions homologous to the PRD gene 4 homeo box is located on PRD gene 6. Hence, at least three genes belong to the PRD-repeat as well as to the homeo box gene set.

Discussion

Here we describe a new set of genes defined by a conserved putative $(his\text{–}pro)_n$ domain, the PRD gene set. We are proposing two hypotheses. First, the PRD gene set is

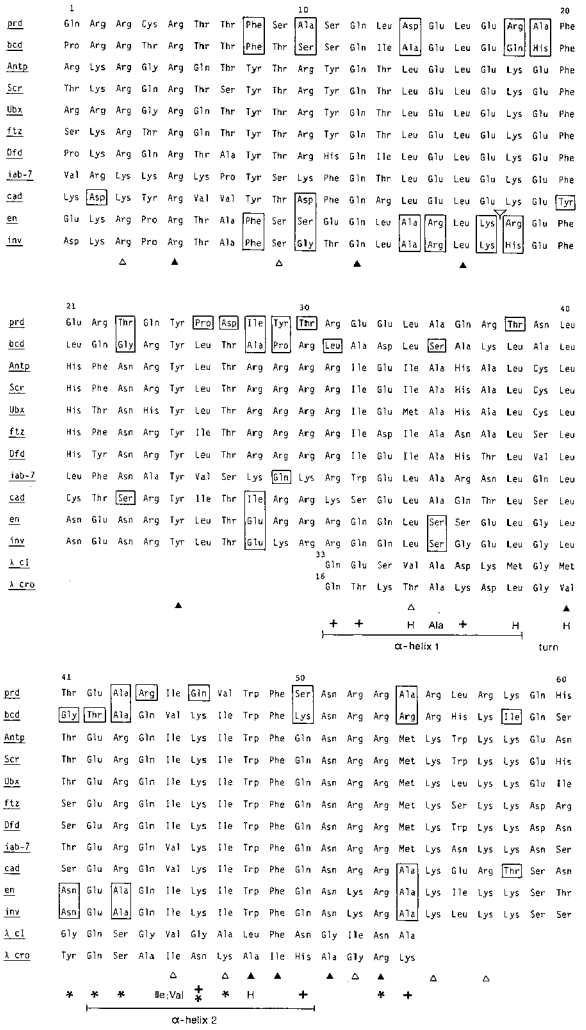

Figure 9. Amino Acid Sequences of Drosophila Homeo Boxes Interpreted in the Helix-Turn-Helix Model

The amino acid sequences of the *prd* and PRD gene 4 (*bcd*) homeo boxes are compared to those of *Antp*, *Scr*, *Ubx*, *ftz* (all from Kuroiwa et al., 1985; Laughon et al., 1985), *Dfd* (Regulski et al., 1985; Laughon et al., 1985), *iab-7* (Regulski et al., 1985), *cad* (Mlodzik et al., 1985), *en* (Poole et al., 1985; Fjose et al., 1985), and *inv* (Poole et al., 1985). At position 17 of the *Dfd* homeo box, Asp instead of Glu has been reported by Laughon et al. (1985), and at position 33 of the *en* homeo box, Glu instead of Gln by Fjose et al. (1985). Positions at which the amino acid has been strictly conserved (▲) or where only conservative changes are observed (△) are marked. Nonconservative amino acid changes at positions that are highly conserved in *Antp*-, *Ubx*-like homeo boxes are boxed. The positions of the turn and the two α-helices in the helix-turn-helix model (Pabo and Sauer, 1984; Laughon and Scott, 1984; Laughon et al., 1985) are indicated below amino acids 31 to 50. In this region the homeo box sequences are compared to those of the λ repressor proteins λ*cI* and λ*cro* (Hsiang et al., 1977; Roberts et al., 1977; Sauer and Anderegg, 1978; Pabo and Sauer, 1984; Laughon and Scott, 1984). Positions in these proteins known to interact with the DNA backbone (+) or with the bases in the major groove (*), or positions occupied by conserved Ala, Ile/Val, or hydrophobic amino acids (H) thought to be important for the overall geometry of the two α-helices, are indicated (Pabo and Sauer, 1984). The triangle between the *en* and *inv* sequence marks the position of an intron in these two related genes (Poole et al., 1985; Fjose et al., 1985).

one of many gene sets, each defined by its cognate domain (for example, the homeo domain or M-repeat). Second, all of these gene sets are linked through genes containing more than one of these domains. That this might be a general principle was suggested, though not recognized, by the presence of homeo boxes in genes of the M-repeat set (McGinnis et al., 1984a; Regulski et al., 1985; Kuroiwa et al., 1985; Laughon et al., 1985). To test this proposal, we searched the recently isolated *prd* gene (Kilchherr et al., 1986) for domains occurring elsewhere in the genome. We found a group of 11 additional genes linked by the PRD-repeat. Three of those that showed a peak of activity before germ band retraction were characterized in greater detail.

The heuristic value of this concept is illustrated by the fact that this approach has led so far to the isolation of one known developmental gene (*bicoid*) as well as to the discovery of two new types of homeo boxes that could only be recognized by sequencing. More recent work has shown that the *prd* gene contains, in addition to the PRD-repeat and the homeo box, a further domain extending over a length of 128 amino acids (positions 931–1314 in Figure 1) that enlarges the network to four interconnected sets (Bopp et al., 1986). We may thus extend the network by jumping from one set to the next as long as new domains are discovered.

In contrast to most previously analyzed genes in which the homeo box is located near the carboxy-terminal (McGinnis et al., 1984a, 1984b; Scott and Weiner, 1984; Regulski et al., 1985; Kuroiwa et al., 1985; Poole et al., 1985), the newly found homeo boxes are either near the amino terminal (PRD gene 4) or in the middle of the protein (*prd*, Figure 1). The same variability of position is found in the case of PRD- and M-repeats (Regulski et al., 1985; Laughon et al., 1985; our unpublished results). This is what would be expected if domains assort independently during evolution. Independent assortment of domains rather than evolutionary drift might also account for the high variability of the regions flanking a domain. This explanation is in better accord with the slow accumulation of single base changes estimated by Ohno (1985).

In apparent contradiction to independent assortment, it is striking that a relatively large fraction of the PRD domains are either linked to or in close proximity to homeo domains. At least two explanations come to mind. Biological factors favor the association of homeo boxes and PRD-repeats either within the same gene or in neighboring genes. Combination of a homeo and PRD domain within the same gene might be explained if an originally random association proved functionally advantageous and hence stabilized their linkage during evolution. Occurrence of PRD and homeo domains in neighboring genes, on the other hand, is expected if homeo and PRD domains occur preferentially in developmental genes that frequently have been found to be clustered (e.g., genes in the bithorax or Antennapedia complex). Because the number of such clustered genes might be relatively large and the number of different domains present in these genes might be small, a nonexclusive second possibility is that there exist many more copies of homeo or PRD domains that have

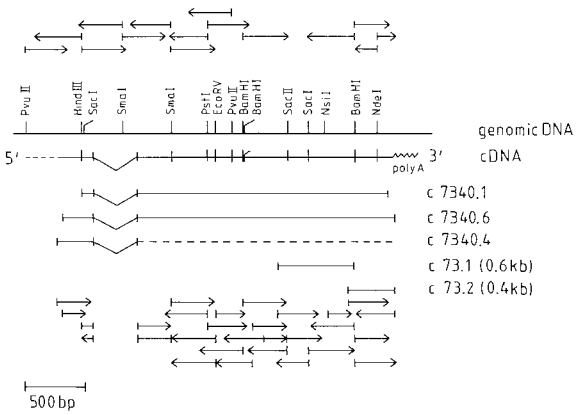

Figure 10. Sequencing Strategy of the *prd* Gene

A restriction map of the region containing the *prd* gene (genomic DNA) and the location of the 3′ end and of the intron boundaries of the *prd* transcript (cDNA) are shown. The precise position of its 5′ end has not been determined. Below, the regions covered by several isolated cDNAs are indicated (Kilchherr et al., 1986). The portion of the cDNA c7340.4 shown as a broken line has not been sequenced. Position, length, and orientation (DNA strand) of the sequenced regions are represented by corresponding arrows at the top (genomic sequences) and bottom (cDNA sequences). A small part of the genomic sequence in the 0.9 kb BamHI fragment has not been determined. Its equivalence with the corresponding cDNA sequence is assumed from their comparison by restriction analysis, which revealed no intron.

yet to be discovered. It is improbable that many more PRD genes will be discovered because the sensitivity of hybridization with repeat structures is intrinsically high. In contrast, homeo boxes easily escape detection, as shown here for the *prd* gene and PRD gene 4.

Networks are a general form of organization used to integrate complex functions (Kauffman, 1971). Although the recognition of their existence in biology may not have provided insights of immediate predictive value, their characterization and study are nevertheless important. Present methods of finding functional links between genes that are forming an integrated system are tedious and the results difficult to interpret (e.g., modification of expression patterns of one gene by another as studied by in situ hybridization to tissue sections). A more efficient approach to elucidating functional networks may be the systematic method of exploiting structural links by jumping from one gene set to another, as we have proposed. Is it a valid assumption that we preferentially find genes that belong to an integrated system? In favor of this hypothesis are the following observations and arguments. First, homeo boxes are associated with genes functioning in development (McGinnis et al., 1984a; Scott and Weiner, 1984; Regulski et al., 1985; Kuroiwa et al., 1985; Laughon et al., 1985; Poole et al., 1985; Fjose et al., 1985). Second, PRD genes are either known to be important in early determinative events (*bcd*: Frohnhöfer and Nüsslein-Volhard, 1986; *prd*) or are suspected of having a developmental function on the basis of their temporal expression pattern. Third, some genes are found to possess both the homeo and PRD domain.

Our structural analysis suggests that genes are linked in networks of functional significance. The functional importance of such networks has been discussed in a general theory of the behavior of gene regulation networks by Kauffman (1971) and in the special case of the immune system by Jerne (1974). If our structural approach to discovering gene networks is general, it must be applicable to genes of other integrated systems (nervous system, cytoskeleton, metabolic pathways, cell–cell recognition, immune system, DNA replication, and so on). All these systems must be in turn interconnected within a network of a still higher order to make up the phenotype of an organism.

Experimental Procedures

Isolation of cDNA Clones

A cDNA-library of poly $(A)^+$ RNA from 0–4 hr-old embryos was prepared in λgt10, and nearly full-length cDNAs of *prd*, PRD gene 4 (c53.46.1 and c53.46.6), and PRD gene 7 (c21.12.1) were isolated according to standard procedures (Maniatis et al., 1982). The 0.6 kb *prd* cDNA was isolated from a cDNA-library kindly provided by Tom Kornberg and prepared from poly $(A)^+$ RNA of 0–3 hr-old embryos.

Isolation of Genomic DNA Clones with Sequences Homologous to *prd*

Genomic clones were isolated from an *odd*$^{1.36.18}$ *b pr cn sca l CyO* library in EMBL4 (Frischauf et al., 1983) by hybridizing Benton-Davis transfers (Benton and Davis, 1977) of the library with the radioactively labeled (Rigby et al., 1977) 0.6 kb *prd* cDNA probe at reduced stringency (McGinnis et al., 1984a). Subclones of the isolated regions were prepared according to standard procedures (Maniatis et al., 1982).

DNA Sequencing

The isolation of genomic and cDNA clones of the *prd* gene has been described (Kilchherr et al., 1986). Figure 10 shows a restriction map of the *prd* gene (see Figure 3 of Kilchherr et al., 1986) and indicates the region covered by several cDNAs that have been isolated by screening a λgt10 library containing DNA complementary to poly $(A)^+$ RNA from 0–4 hr-old (c7340.1, c7340.4, c7340.6) or 0–3 hr-old (c73.1, c73.2) Drosophila embryos. As evident from the sequencing strategy illustrated in Figure 10, many regions of both the genomic and various cDNAs were sequenced in several clones, and all regions were sequenced on both strands. Only one strand of the 5′ ends of the two cDNA clones, c7340.4 and c7340.6, was sequenced, whereas the corresponding genomic DNA was sequenced on both strands. Conversely, the 0.9 kb region between two BamHI sites was sequenced on both strands of several cDNAs, yet only on one strand or not at all on the genomic DNA. This omission was permitted because comparison of the genomic and cDNA by restriction analysis within this region demonstrated no intron.

All other DNA sequences were analyzed on both strands, reading each sequence at least twice on independent clones. The DNAs were sequenced in the M13 vector mWB3296 by the dideoxynucleotide sequencing procedure of Sanger et al. (1977). The mWB3296 vector was constructed from mWB2344 (Barnes et al., 1983) by replacing the ClaI fragment with that of M13mp8 and thus eliminating the BamHI site at 2220. In a second step, the polylinker was replaced by the M13mp18 polylinker.

Acknowledgments

This paper is dedicated to Niels Kaj Jerne in honor of his 75th birthday.

We thank Hans Noll for extensive discussions and for his contribution to the formulation of the gene network concept, John Shepherd for his help in the search for DNA homologies between the *prd* and *bcd* gene, and Erich Frei and John Shepherd for many stimulating discussions. We thank Christiane Nüsslein-Volhard for the *prd*$^{2.45.17}$ strain, Tom Kornberg for a cDNA-library, Bill McGinnis for the M-repeat subclone p903P2, the *Antp*-homeo box subclone p903G2, and the *Ubx*-

homeo box subclone p96 (all described by McGinnis et al., 1984a). This work was supported by Swiss National Science Foundation grant No. 3.600–0.84 and by the Kanton Basel.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *"advertisement"* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received July 28, 1986; revised October 21, 1986.

References

Anderson, W. F., Ohlendorf, D. H., Takeda, Y., and Matthews, B. W. (1981). Structure of the cro repressor from bacteriophage λ and its interaction with DNA. Nature *290*, 754–758.

Barnes, W. M., Bevan, M., and Son, P. H. (1983). Kilo-sequencing: creation of an ordered nest of asymmetric deletions across a large target sequence carried on phage M13. Meth. Enzymol. *101*, 98–122.

Benton, W. D., and Davis, R. W. (1977). Screening λgt recombinant clones by hybridization to single plaques in situ. Science *196*, 180–182.

Bopp, D., Burri, M., Baumgartner, S., Frigerio, G., and Noll, M. (1986). Conservation of a large protein domain in the segmentation gene *paired* and in functionally related genes of Drosophila. Cell, in press.

Doolittle, R. F. (1981). Similar amino acid sequences: chance or common ancestry? Science *214*, 149–159.

Doyle, H. J., Harding, K., Hoey, T., and Levine, M. (1986). Transcripts encoded by a homeo box gene are restricted to dorsal tissues of *Drosophila* embryos. Nature *323*, 76–79.

Fjose, A., McGinnis, W. J., and Gehring, W. J. (1985). Isolation of a homeo box–containing gene from the *engrailed* region of *Drosophila* and the spatial distribution of its transcripts. Nature *313*, 284–289.

Foe, V. E., and Alberts, B. M. (1983). Studies of nuclear and cytoplasmic behaviour during the five mitotic cycles that precede gastrulation in *Drosophila* embryogenesis. J. Cell Sci. *61*, 31–70.

Frei, E., Baumgartner, S., Edström, J.-E., and Noll, M. (1985). Cloning of the *extra sex combs* gene of *Drosophila* and its identification by P-element-mediated gene transfer. EMBO J. *4*, 979–987.

Frischauf, A.-M., Lehrach, H., Poustka, A., and Murray, N. (1983). Lambda replacement vector carrying polylinker sequences. J. Mol. Biol. *170*, 827–842.

Frohnhöfer, H. G., and Nüsslein-Volhard, C. (1986). Organization of anterior pattern in the *Drosophila* embryo by the maternal gene *bicoid*. Nature, in press.

Gilbert, W. (1978). Why genes in pieces? Nature *271*, 501.

Gilbert, W., Marchionni, M., and McKnight, G. (1986). On the antiquity of introns. Cell *46*, 151–154.

Hafen, E., Levine, M., Garber, R. L., and Gehring, W. J. (1983). An improved *in situ* hybridization method for the detection of cellular RNAs in *Drosophila* tissue sections and its application for localizing transcripts of the homeotic Antennapedia gene complex. EMBO J. *2*, 617–623.

Hafen, E., Kuroiwa, A., and Gehring, W. J. (1984). Spatial distribution of transcripts from the segmentation gene *fushi tarazu* during Drosophila embryonic development. Cell *37*, 833–841.

Hsiang, M. W., Cole, R. D., Takeda, Y., and Echols, H. (1977). Amino acid sequence of Cro regulatory protein of bacteriophage lambda. Nature *270*, 275–277.

Ingham, P. W., Howard, K. R., and Ish-Horowicz, D. (1985). Transcription pattern of the *Drosophila* segmentation gene *hairy*. Nature *318*, 439–445.

Jerne, N. K. (1974). Towards a network theory of the immune system. Ann. Immunol. Inst. Pasteur *125C*, 373–389.

Kauffman, S. (1971). Gene regulation networks: a theory for their global structure and behaviors. Curr. Topics Dev. Biol. *6*, 145–182.

Kilchherr, F., Baumgartner, S., Bopp, D., Frei, E., and Noll, M. (1986). Isolation of the *paired* gene of *Drosophila* and its spatial expression during early embryogenesis. Nature *321*, 493–499.

Kuroiwa, A., Kloter, U., Baumgartner, P., and Gehring, W. J. (1985). Cloning of the homeotic *Sex combs reduced* gene in *Drosophila* and *in situ* localization of its transcripts. EMBO J. *4*, 3757–3764.

Langer-Safer, P. R., Levine, M., and Ward, D. C. (1982). Immunological method for mapping genes on *Drosophila* polytene chromosomes. Proc. Natl. Acad. Sci. USA *79*, 4381–4385.

Laughon, A., and Scott, M. P. (1984). Sequence of a *Drosophila* segmentation gene: protein structure homology with DNA-binding proteins. Nature *310*, 25–31.

Laughon, A., Carroll, S. B., Storfer, F. A., Riley, P. D., and Scott, M. P. (1985). Common properties of proteins encoded by the Antennapedia complex genes of *Drosophila melanogaster*. Cold Spring Harbor Symp. Quant. Biol. *50*, 253–262.

Lewis, E. B. (1978). A gene complex controlling segmentation in *Drosophila*. Nature *276*, 565–570.

Lewis, R. A., Kaufman, T. C., Denell, R. E., and Tallerico, P. (1980). Genetic analysis of the Antennapedia gene complex (Ant-C) and adjacent chromosomal regions of *Drosophila melanogaster*. I. Polytene chromosome segments 84B–D. Genetics *95*, 367–381.

Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982). Molecular Cloning: A Laboratory Manual (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory).

Matthews, B. W., Ohlendorf, D. H., Anderson, W. F., and Takeda, Y. (1982). Structure of the DNA-binding region of *lac* repressor inferred from its homology with *cro* repressor. Proc. Natl. Acad. Sci. USA *79*, 1428–1432.

McGinnis, W., Levine, M. S., Hafen, E., Kuroiwa, A., and Gehring, W. J. (1984a). A conserved DNA sequence in homeotic genes of the *Drosophila* Antennapedia and bithorax complexes. Nature *308*, 428–433.

McGinnis, W., Garber, R. L., Wirz, J., Kuroiwa, A., and Gehring, W. J. (1984b). A homologous protein-coding sequence in *Drosophila* homeotic genes and its conservation in other metazoans. Cell *37*, 403–408.

McKay, D. B., and Steitz, T. A. (1981). Structure of catabolite gene activator protein at 2.9A resolution suggests binding to left-handed B-DNA. Nature *290*, 744–749.

Mlodzik, M., Fjose, A., and Gehring, W. J. (1985). Isolation of *caudal*, a *Drosophila* homeo box–containing gene with maternal expression, whose transcripts form a concentration gradient at the pre-blastoderm stage. EMBO J. *4*, 2961–2969.

Ohno, S. (1985). Dispensable genes. Trends in Gen. *1*, 160–164.

Pabo, C. O., and Lewis, M. (1982). The operator-binding domain of repressor: structure and DNA recognition. Nature *298*, 443–447.

Pabo, C. O., and Sauer, R. T. (1984). Protein-DNA recognition. Ann. Rev. Biochem. *53*, 293–321.

Poole, S. J., Kauvar, L. M., Drees, B., and Kornberg, T. (1985). The *engrailed* locus of *Drosophila*: structural analysis of an embryonic transcript. Cell *40*, 37–43.

Regulski, M., Harding, K., Kostriken, R., Karch, F., Levine, M., and McGinnis, W. (1985). Homeo box genes of the Antennapedia and bithorax complexes of *Drosophila*. Cell *43*, 71–80.

Rigby, P. W. J., Dieckman, M., Rhodes, C., and Berg, P. (1977). Labeling deoxyribonucleic acid to high specific activity *in vitro* by nick translation with DNA polymerase I. J. Mol. Biol. *113*, 237–251.

Roberts, T. M., Shimatake, M., Brady, C., and Rosenberg, M. (1977). Sequence of *cro* gene of bacteriophage lambda. Nature *270*, 274–275.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA. *74*, 5463–5467.

Sauer, R. T., and Anderegg, R. (1978). Primary structure of the λ repressor. Biochemistry *17*, 1092–1100.

Sauer, R. T., Yocum, R. R., Doolittle, R. F., Lewis, M., and Pabo, C. O. (1982). Homology among DNA-binding proteins suggests use of a conserved super-secondary structure. Nature *298*, 447–451.

Schneuwly, S., Kuroiwa, A., Baumgartner, P., and Gehring, W. J. (1986). Structural organization and sequence of the homeotic gene *Antennapedia* of *Drosophila melanogaster*. EMBO J. *5*, 733–739.

Scott, M. P., and Weiner, A. J. (1984). Structural relationships among genes that control development: sequence homology between the An-

tennapedia, Ultrabithorax, and *fushi tarazu* loci of *Drosophila*. Proc. Natl. Acad. Sci. USA *81*, 4115–4119.

Scott, M. P., Weiner, A. J., Hazelrigg, T. I., Polisky, B. A., Pirrotta, V., Scalenghe, F., and Kaufman, T. C. (1983). The molecular organization of the *Antennapedia* locus of *Drosophila*. Cell *35*, 763–776.

Smith, G. E., and Summers, M. D. (1980). The bidirectional transfer of DNA and RNA to nitrocellulose or diazobenzyloxymethyl-paper. Anal. Biochem. *109*, 123–129.

Southern, E. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. *98*, 503–517.

Wakimoto, B. T., Turner, F. R., and Kaufman, T. C. (1984). Defects in embryogenesis in mutants associated with the Antennapedia gene complex of *Drosophila melanogaster.* Dev. Biol. *102*, 147–172.

Weber, I. T., McKay, D. B., and Steitz, T. A. (1982). Two helix DNA binding motif of CAP found in *lac* repressor and *gal* repressor. Nucl. Acids Res. *10*, 5085–5102.

Wharton, K. A., Yedvobnick, B., Finnerty, V. G., and Artavanis-Tsakonas, S. (1985). *opa*: a novel family of transcribed repeats shared by the *Notch* locus and other developmentally regulated loci in D. melanogaster. Cell *40*, 55–62.