The EMBO Journal vol.8 no.4 pp.1183 - 1190, 1989

Conservation of the paired domain in metazoans and its structure in three isolated human genes

Maya Burri[1], Yvonne Tromvoukis[2], Daniel Bopp[3], Gabriella Frigerio[1] and Markus Noll[1]

Department of Cell Biology, Biocenter of the University, Klingelbergstr. 70, CH-4056 Basel, Switzerland

[1]Present address: Institute for Molecular Biology II, University of Zürich, CH-8093, Switzerland
[2]Present address: IFREC, CH-1066 Epalinges, Switzerland
[3]Present address: Department of Biology, Princeton University, Princeton, NJ 08540, USA

Communicated by M.Noll

Sequences homologous to the paired domain of *Drosophila melanogaster* have been conserved in species as distantly related as nematodes, sea urchins, or man. In particular, paired domains of three human genes, HuP1, HuP2 and HuP48, have been isolated and sequenced. Together with four *Drosophila* paired domains, they fall into two separate paired domain classes named according to their *Drosophila* members, *paired*–*gooseberry* and P29 class. The P29 class includes the mouse Pax 1 and the human HuP48 gene which are nearly identical in their sequenced portions and hence might be true homologues. In addition to the paired domain, the two human genes HuP1 and HuP2 share the highly conserved octapeptide HSIAGILG with the two *gooseberry* genes of *Drosophila*. Possible functions of the paired domain are discussed in the light of a predicted helix-turn-helix structure in its carboxy-terminal portion.

Key words: evolution/gene network concept/human and metazoan paired domains

Introduction

Genes (defined as transcription units) regulating complex integrated functions, such as the programming of early development, often encode proteins with multiple conserved domains. We have suggested that these genes are integrated into functional networks which evolved by duplication and recombination of a small number of primordial genes corresponding to the conserved protein domains (Bopp *et al.*, 1986; Frigerio *et al.*, 1986). The resulting superfamilies of genes form structural and functional networks because all genes belonging to the network are linked to each other by members that combine two or more domains in one transcription unit; for example the set of genes containing a paired box is linked to the homeo box gene set by all those genes that contain both a paired and a homeo box, and so on.

The existence of such structurally defined gene networks is of biological interest because conservation of structure implies preservation of essential functions. Our gene network concept makes two major predictions of biological interest that may be tested experimentally. First, starting from a single, isolated gene that is part of the network in a particular organism, it should be possible to determine experimentally all members of the network. Second, the same set of conserved domains is expected to define analogous networks in all organisms linked by evolution.

In our previous work, we have tested the first prediction by screening *Drosophila* libraries at low stringency, starting with a *paired* (*prd*) gene probe containing the PRD-repeat (Frigerio *et al.*, 1986). So far, this approach has resulted in the isolation of 15 genes, three of which could be identified with known phenotypes [*bicoid* (Frohnhöfer and Nüsslein-Volhard, 1986) and the two *gooseberry* (*gsb*) genes (Nüsslein-Volhard and Wieschaus, 1980)]. Characterization of some of these genes led to the discovery of a new domain, the paired box (Bopp *et al.*, 1986), as well as two new types of homeo boxes (Frigerio *et al.*, 1986). As expected from the predicted independent assortment of domains (Frigerio *et al.*, 1986), the three domains examined have been found to occur in five out of six possible combinations with each other, different arrangements within the transcription units and embedded in non-conserved flanking sequences. Moreover, all of the genes isolated by this approach appear to be involved in early development (Bopp *et al.*, 1986; Frigerio *et al.*, 1986; Bopp *et al.*, in preparation).

In the present study, we have begun to examine the evolutionary aspect of the network concept. We show that the paired domain has been conserved in a variety of organisms widely separated on the evolutionary scale. In particular, we have isolated three human genes containing a highly conserved paired domain. These three human paired domains fall into two separate classes: in two of the human genes, the paired domains are more closely related to those of the *Drosophila* genes *prd* and *gsb* (BSH9 and BSH4), while in the third human gene, the paired domain is similar to that of the *Drosophila* gene P29 (Bopp *et al.*, in preparation) and nearly identical with that of the recently described Pax 1 gene of the mouse (Deutsch *et al.*, 1988).

Results

Conservation of the paired domain throughout living nature

To test the prediction (Bopp *et al.*, 1986) that the *Drosophila* paired domain has been conserved in other organisms, a Southern blot of *Eco*RI-digested DNA from a variety of species ranging from yeast to man was hybridized at reduced stringency with a labelled DNA fragment containing only the paired box of the *gsb* gene BSH9, P_{BSH9}. As evident from Figure 1, all genomes analysed exhibit several hybridization signals above background. Even yeast DNA shows two or three weak bands. However, their significance is debatable as the amount of yeast DNA loaded relative to its complexity is ten times that of *Drosophila melanogaster*. The same argument questions the significance of the bands displayed by the *Arabidopsis* genome. On the other hand, the bands visible in the genomes of the nematode *Parascaris equorum* (germ cells and intestinal cells), the sea urchin

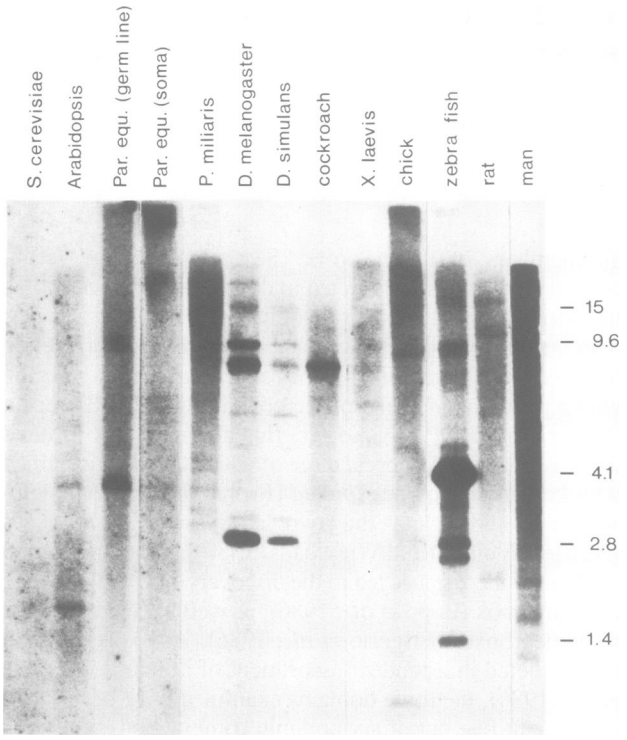

Fig. 1. Paired box homologies in the genomes of a variety of organisms. Genomic DNAs of the following organisms were digested with *Eco*RI, run on a 0.6% agarose gel, transferred to a Zetabind matrix (Cuno Lab. Prod.), and hybridized at reduced stringency with a ^{32}P-labelled paired box probe of BSH9 (amount of DNA loaded and approximate genome size of organism are indicated in parentheses): *Saccharomyces cerevisiae* (1 μg; 0.16 $\times$ 10^5 kb), the dicot. *A.thaliana* (5 μg; 0.8 $\times$ 10^5 kb), oocytes (germ line: 2.3 $\times$ 10^6 kb) and intestine (soma: 0.4 $\times$ 10^6 kb) of the nematode *P.equorum* (7.5 μg each), the sea urchin *P.miliaris* (10 μg; 0.75 $\times$ 10^6 kb), the fruit flies *D.melanogaster* and *D.simulans* (1 μg each; 0.17 $\times$ 10^6 kb), the cockroach (2 μg; $\sim 10^6$ kb), the frog *X.laevis* (10 μg; 3 $\times$ 10^6 kb), chicken (10 μg; 1.2 $\times$ 10^6 kb), zebra fish (4 μg; 2 $\times$ 10^6 kb), rat (10 μg; 3 $\times$ 10^6 kb) and man (5 μg; 3 $\times$ 10^6 kb). Autoradiography occurred for 1, 2 or 5 days at -70°C with an intensifying screen. On the right, size markers are indicated in kb.

Psammechinus miliaris, *D.simulans*, the cockroach, *Xenopus laevis*, chick, zebra fish, rat and man appear to reflect a true homology to the paired box probe since similar numbers of genomic copies have been analysed as for *D.melanogaster*. Particularly strong bands are observed for zebra fish DNA. Hence, we consider it likely that the paired domain is present in vertebrates, echinoderms and arthropods as well as in nematodes and thus has spread throughout the animal kingdom since the separation of the aschelminthes (nematodes) from the coeloid type of animals. If the hybridization observed with *Arabidopsis thaliana* DNA is significant, the paired domain would even date back to the separation of animals and plants.

The 9.6, 2.8, 8.3, 8.1 and 15 kb *Eco*RI bands of *D.melanogaster* have been identified with the paired box genes *prd*, *gsb* (BSH9 and BSH4) (Bopp *et al.*, 1986), P4 and P29 (Bopp *et al.*, in preparation), respectively. The weaker bands at 3.2, 4.8 and 5.8 kb probably reflect polymorphisms at *Eco*RI sites of the BSH9 locus (neighbouring *Eco*RI fragments of 1.6, 0.4, 2.8 and 3.0 kb) because no additional genes could be isolated by extensive screens of a genomic library with the P_{BSH9} probe. *Eco*RI bands of identical size are observed for the closely related

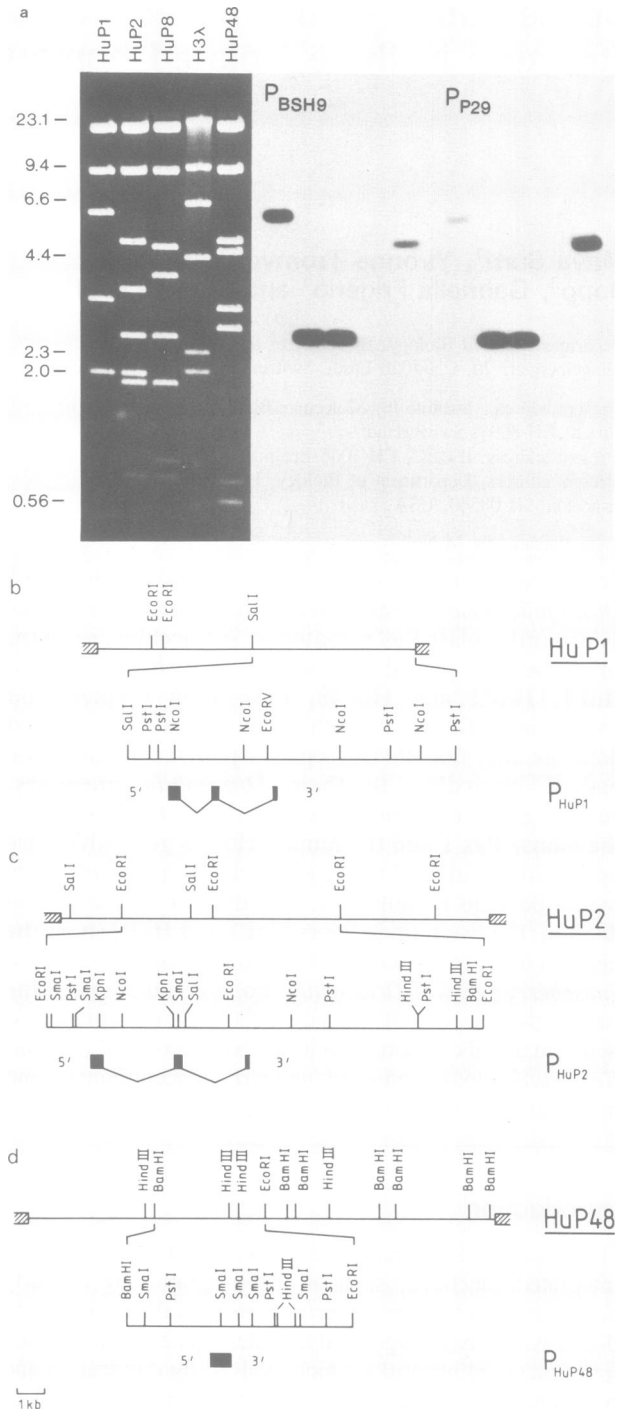

Fig. 2. Restriction maps of three human genomic clones containing paired domains. (**a**) Two Southern blots of four human genomic clones isolated from an EMBL3 library (kindly provided by Anne-Maria Frischauf) and digested with *Eco*RI and *Sal*I (HuP1, HuP2, HuP8) or *Bam*HI and *Sal*I (HuP48) were hybridized at reduced stringency with ^{32}P-labelled paired domain probes of the *Drosophila* genes BSH9 (Bopp *et al.*, 1986; Baumgartner *et al.*, 1987) or P29 (Bopp *et al.*, in preparation) and exposed for 4 h at -70°C with an intensifying screen (middle and right panel). The left panel shows the DNA and size markers (*Hin*dIII-digest of λ-DNA) before transfer after staining the 0.6% agarose gel with ethidium bromide [fragment lengths (kb) are indicated on the left]. (**b**)–(**d**) Restriction maps of the three isolated human genomic regions containing paired domains are shown. Above and below each map, the extent of the genomic clone in EMBL3 and the region coding for the paired domain (filled bar) and its transcriptional organization are depicted. The scale of the three phage inserts is indicated in (d).

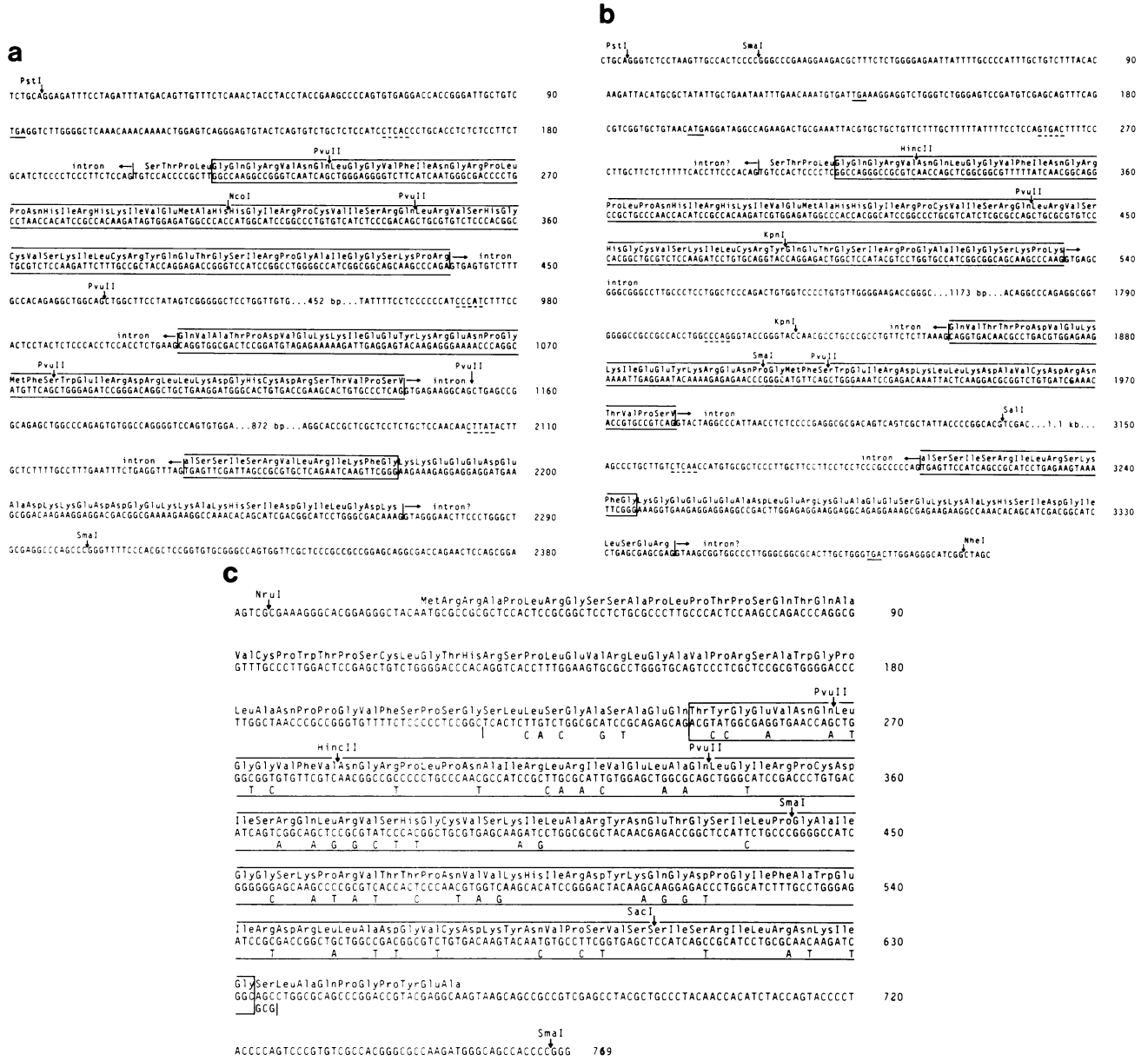

Fig. 3. (**a**) Genomic DNA sequence of a region comprising the paired domain of the human gene HuP1. The DNA sequence coding for the paired domain of HuP1, which is interrupted by two introns, is boxed. Intron boundaries are indicated by vertical bars and their branch site consensus sequences for lariat formation are underlined by a broken line. A stop codon (TGA) preceding the paired domain in the same frame is underlined. (**b**) Genomic DNA sequence of a region comprising the paired domain of the human gene HuP2. For explanations see text and legend to (a). (**c**) Genomic sequence of a region comprising the paired domain of the human gene HuP48. Below the DNA sequence coding for the paired domain (boxed), deviations of the paired domain sequence of the mouse Pax 1 gene (Deutsch *et al.*, 1988) are indicated between its published limits (indicated by vertical lines). In the first codon of the paired domain of the published Pax 1 sequence (Deutsch *et al.*, 1988) a C has been introduced, producing a frameshift in the sequence preceding it and resulting in a stretch of 11 additional identical amino acids that Pax 1 shares with HuP48. The published Pax 1 sequence is likely to be erroneous in this portion because its corrected sequence reveals four third base wobbles and one first base difference that do not alter its amino acid sequence with respect to that of HuP48. Amino acids beyond the 3′-end of the paired box are indicated only up to a potential 5′-splice site.

species *D.simulans*, suggesting that these fragments carry the same genes as in *D.melanogaster*.

Isolation and sequence of three human paired domains

To verify the results of the whole genome Southern analysis shown in Figure 1, we screened a human genomic library in EMBL3 (Frischauf *et al.*, 1983) for sequences homologous to P_{BSH9} or P_{P29}, a paired domain of the *Drosophila* gene, P29, which shows a number of deviations from the *prd−gsb* paired domain and is thought to belong to a separate class of paired domains (Bopp *et al.*, in preparation). Three phages, HuP1, HuP2 and HuP8, were isolated by the screen with P_{BSH9} and one, HuP48, by screening with P_{P29}.

The two phages HuP2 and HuP8 overlap, as is evident from cross-hybridization (not shown), and contain the same paired domain gene, HuP2. HuP1 and HuP48 derived from separate loci, each harbouring one paired domain gene, called HuP1 and HuP48, respectively (Figure 2a). Restriction maps of the three isolated human paired domain genes,

HuP1, HuP2 and HuP48, are shown in Figure 2b–d. The DNA regions containing the paired domains of all three genes were sequenced on both strands, and relevant sequences are shown in Figure 3a–c.

Whereas the paired domains of HuP1 and HuP2 are interrupted by two introns at identical positions, the paired domain of HuP48 has no introns. The intron boundaries have been assigned to the positions indicated in Figure 3 so as to obtain maximum homology of the resulting exons with known paired domains and to be consistent with the known 5′- and 3′-splice site consensus sequences (Padgett *et al.*, 1986). In addition, this choice of introns would close the open reading frame (ORF) of the interrupted exons after a short stretch of the intron sequence and reveal sequences homologous to branch sites (Keller and Noon, 1985) 35–41 nucleotides from the 3′-splice site. The only ambiguity concerns the 3′-end of the first intron interrupting the paired domains of HuP1 and HuP2. In both cases splicing could occur three nucleotides downstream from the positions indicated in Figure 3a and b. This would eliminate a Gln in position 75 of the paired domain and improve the homology to the *Drosophila* paired domains (Figure 4). By contrast, in favour of the splicing shown in Figure 3a and b is the observation that 3′-splice sites avoid AGs because these may be inhibitory to splicing (Wieringa *et al.*, 1984) and our recent finding of a *Drosophila* gene (P4) in which a Gln has been inserted at position 75 of the paired domain (Bopp *et al.*, in preparation).

It is probable that two of the three isolated human paired domains are preceded by introns. The ORF of the HuP1 paired box is in frame with an upstream stop codon (underlined in Figure 3a) and lacks an ATG initiation codon in between. The 3′-end of its preceding intron has been assigned on the basis of three criteria: (i) agreement with the 3′-splice site consensus sequence following a pyrimidine-rich region, (ii) absence of AGs between positions −3 and −19 from the 3′-splice site (Keller and Noon, 1985) and (iii) the occurrence of a 3′-splice signal for lariat formation within 20–50 nucleotides from the 3′-splice site (Keller and Noon, 1985). The ORF of the paired box of HuP2 is also in frame with upstream stop codons, the first occurring at positions 136–138 (underlined in Figure 3b). Since this stop codon is followed by an ATG in frame at position 196–198 (underlined in Figure 3b), it is possible that the beginning of the paired domain is encoded in the first exon of the HuP2 gene. However, two observations argue against this possibility and suggest that the paired domain of HuP2 is preceded by an intron as well. First, the potential 3′-splice site indicated in Figure 3b fulfills the three criteria mentioned above, and second, the homology between HuP1 and HuP2 extends upstream of the paired box and apparently ends at the indicated 3′-splice sites. Therefore, we suggest that the paired domains of HuP1 and HuP2 are preceded by introns. They would share this property with all five cloned *Drosophila* genes containing a paired domain (Figure 4; Bopp *et al.*, in preparation). Similarly, the homology between HuP1 and HuP2 also extends downstream from the paired domain (Figure 5) and ends at potential 5′-splice sites indicated in Figure 3a and b. Since these sites exhibit a striking homology to the consensus sequence AGGUAAGU of 5′-splice sites (Padgett *et al.*, 1986), it seems probable that they are true intron boundaries. A similar argument cannot be made for HuP48 as, in the sequenced portion, its homology with HuP1 and HuP2 does not extend on either side of the paired domain (Figure 3).

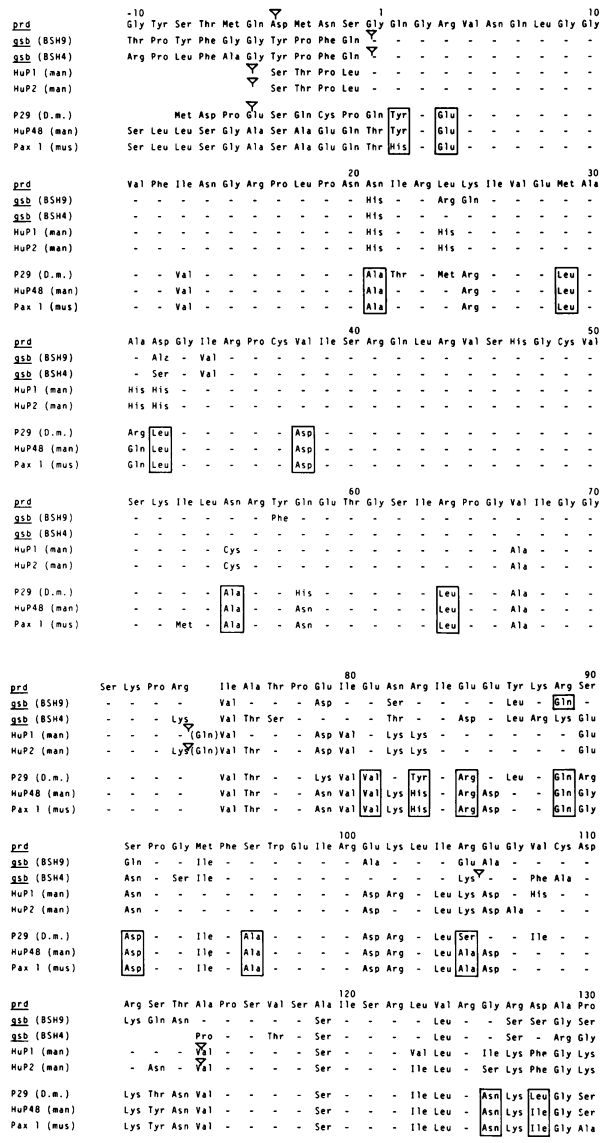

Fig. 4. Paired domains of the *prd–gsb* and P29 class in *Drosophila*, mouse and man. The amino acid sequences of the paired domains of three human genes, HuP1, HuP2 and HuP48, four *Drosophila* genes, *prd*, *gsb* (BSH9 and BSH4) (Bopp *et al.*, 1986; Baumgartner *et al.*, 1987), and P29 (Bopp *et al.*, in preparation), and of the mouse gene Pax 1 (Deutsch *et al.*, 1988) are shown. A probable sequencing error in the first codon of the paired domain of the mouse Pax 1 gene has been corrected (cf. legend to Figure 3c). Positions at which amino acids in the P29 class consistently differ from amino acids conserved in the *prd–gsb* class characterize the two paired domain classes. The corresponding amino acids of the P29 class paired domains are boxed. Amino acids identical to those at corresponding positions of the *prd* sequence are represented by a dash. The locations of introns are indicated by triangles.

Different classes of paired domains in Drosophila, mouse and man

The primary structures of the paired domains of the four *Drosophila* genes *prd*, *gsb* (BSH9 and BSH4) (Bopp *et al.*, 1986) and P29 (Bopp *et al.*, in preparation), of the three human genes HuP1, HuP2 and HuP48, and of the recently described mouse gene Pax 1 (Deutsch *et al.*, 1988) are shown in Figure 4. It is evident that these paired domains

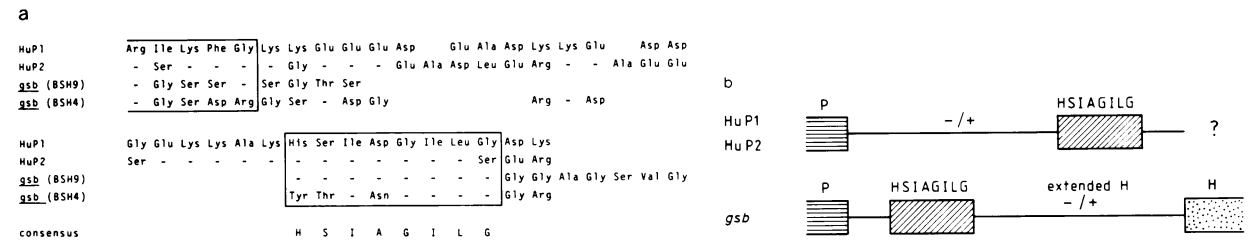

Fig. 5. HSIAGILG, an octapeptide homology between the two *gsb* genes of *Drosophila* and the two human paired domain genes HuP1 and HuP2. **(a)** Amino acid sequences following the end of the paired domains of HuP1, HuP2 and the two *gsb* genes, BSH9 and BSH4. The *gsb* sequences shown are immediately followed by their extended *prd*-homeo domains (Bopp *et al.*, 1986). No homology was detected with the sequenced portion downstream of the HuP48 paired domain. The last five amino acids of the paired domains and the shared homologous octapeptide consensus sequence, HSIAGILG, are boxed. Amino acids identical to those at corresponding positions of the HuP1 sequence are represented by a dash. **(b)** The reverse order of the HSIAGILG octapeptide and a sequence of highly charged amino acids (−/+) following the paired domain P in HuP1 and HuP2 as compared to that between the paired domain P and homeo domain H in the two *gsb* genes. In the two *gsb* genes, the highly charged amino acid sequence corresponds to an extended homologous sequence of 18 amino acids preceding their homeo domains and that of the *prd* gene (Bopp *et al.*, 1986). It is not known whether HuP1 and HuP2 have a homeo domain as well.

Table I. Matrix of amino acid homologies between paired domains of the *prd*−*gsb* and P29 class

		prd	HuP1	HuP2	P29	HuP48
prd, *gsb* class	*prd*					
	HuP1	0.79 . 0.92 (0.89) . (0.92)				
	HuP2	0.78 . 0.91 (0.88) . (0.92)	0.94 . 0.99 (0.97) . (1.00)			
P29 class	P29	0.65 . 0.78 (0.77) . (0.81)	0.72 . 0.80 (0.78) . (0.84)	0.71 . 0.78 (0.79) . (0.84)		
	HuP48	0.66 . 0.81 (0.79) . (0.86)	0.74 . 0.81 (0.81) . (0.86)	0.74 . 0.80 (0.82) . (0.86)	0.88 . 0.93 (0.92) . (0.95)	
	Pax 1	0.65 . 0.80 (0.79) . (0.86)	0.74 . 0.80 (0.81) . (0.86)	0.74 . 0.78 (0.82) . (0.86)	0.86 . 0.91 (0.92) . (0.95)	0.98 . 0.97 (0.99) . (0.99)

The first value of the upper lines indicates for the entire paired domain (amino acids 1−129) the fraction of identical amino acids between any two paired domains of the genes shown, the second value shows the corresponding fraction for the more highly conserved first 74 amino acids of the paired domain. The values in parentheses underneath represent corresponding fractions of amino acid homologies if conservative changes are neglected. The values comparing paired domains of the same class (enclosed by triangles) are consistently higher than those when paired domains belonging to different classes are compared. Nucleotide homologies between entire paired domains range from 0.71 to 0.86 for domains of the same class, or between 0.63 and 0.71 for domains of different classes.

can be grouped into two different classes, one of the *prd*−*gsb* type, which also includes those of HuP1 and HuP2, and the other of the P29 type, which comprises the paired domains of P29, HuP48 and Pax 1. The P29-type paired domain deviates at a number of positions from the *prd*−*gsb* type by non-conservative amino acid changes (boxed in Figure 4). This division of the eight paired domains into two different classes is also apparent from a quantitative analysis of their mutual amino acid homologies shown in the matrix of Table I. Within each class the homologies are clearly higher (78−94% identity for the *prd*−*gsb* class, 86−98% identity for the P29 class) than between any two members belonging to different classes (65−74% identity). Comparing only the paired domains of the three human genes, the homology between HuP1 and HuP2 is extremely high (94% identity), yet considerably lower with respect to HuP48 (74% identity).

This distinction between two classes of the three human genes is also reflected in the sequence organization of their paired domains. Both HuP1 and HuP2 contain two introns within their paired domains at identical positions whereas the paired domain of HuP48 is not interrupted by intervening sequences. The extremely high structural homology between the paired domains of HuP1 and HuP2 is only surpassed by that between the paired domains of the mouse gene Pax 1 (Deutsch *et al.*, 1988) and the human gene HuP48 (Table I). These two genes are identical over the entire length of amino acids 3−129 of their paired domains with the only exception of amino acid 53 which is Met in Pax 1 but Ile in all other known paired domains (Figure 4). Moreover, if we insert a C in front of the third nucleotide of the published paired domain of Pax 1, HuP48 and Pax 1 code for at least 11 additional identical amino acids from the first amino acid of their paired domains towards the amino terminal (Figures 3c and 4). It is thus possible that HuP48 and Pax 1 are homologous genes in mouse and man.

As noted previously (Bopp *et al.*, 1986; Deutsch *et al.*, 1988), the first 74 amino acids exhibit a considerably higher degree of conservation (second values indicated in Table I) than amino acids 75–129. However, this difference between these two portions of the paired domain is drastically reduced if conservative amino acid changes are neglected (values in parentheses in Table I), indicating that the last 40% of the paired domain tolerates more conservative changes than the first 60%. Therefore, the significance of the finding that the last 40% of the P29-type paired domain are more closely related to the homologous region of HuP1 and HuP2 than of *prd* remains unclear. A separation of the first 74 amino acids from the remainder of the paired domain may also be emphasized by an intron at this position in HuP1 and HuP2.

The HSIAGILG sequence, another homology between gsb and HuP1/HuP2

The homology between HuP1 and HuP2 is not restricted to the paired domain but extends up to the probable exon boundaries on either side of it (Figures 3a,b and 5). The amino acid sequence following the paired domain shows particularly interesting features. It consists of a negatively charged region of alternating clusters of acidic and basic amino acids followed by the octapeptide HSIAGILG (or S). Interestingly, a similarly charged sequence has been observed earlier in the extended portion of the *prd*-homeo domains of *prd* and *gsb* that are immediately preceding their homeo domains (Bopp *et al.*, 1986). It was, therefore, of interest to examine the question of whether an additional homology could be detected between HuP1/HuP2 and *prd* or *gsb*. Surprisingly, the octapeptide HSIAGILG was found either unaltered or with minor changes in the two *gsb* genes BSH9 and BSH4 (Figure 5a) but not in *prd*. Thus, the remaining sequenced portions of HuP1 and HuP2 following their paired domains exhibit a striking similarity to the two *gsb* genes. However, whereas in the two *gsb* genes the HSIAGILG sequence follows the paired domain and precedes the highly charged sequence of the extended *prd*–type homeo domain, in the human genes HuP1 and HuP2 the highly charged sequence links the paired domain with the HSIAGILG sequence (Figure 5b).

Is it possible that the homology between *gsb* and HuP1 or HuP2 extends even further such that HuP1 or HuP2 also contain a homeo domain? Although such a finding would bear out an additional prediction of the gene network concept, we have not been able to locate a homeo domain downstream of the paired domains of HuP1 or HuP2. However, we cannot exclude that such a domain would be found on a more distantly located exon of these genes. It will be interesting to find out whether the HSIAGILG sequence represents a new domain that is associated with different domains in other genes of the same network.

Discussion

In agreement with our concept of gene networks characterized by network-specific domains, we have shown previously that the *prd* gene of *Drosophila* consists of at least three different domains shared with other genes belonging to the same network as *prd* (Bopp *et al.*, 1986; Frigerio *et al.*, 1986). One of these domains is the paired domain common to at least five *Drosophila* genes (Bopp *et al.*, in preparation). We have tested our prediction that the paired domain has been conserved in all organisms that have evolved more recently than arthropods. Analysis of the genomes from a variety of organisms by hybridization to a paired domain probe strongly supports this hypothesis and suggests that the paired domain is perhaps not limited to metazoans but may well have evolved before the division of animals and plants about one billion years ago.

Furthermore, we have confirmed by DNA sequencing that the paired domain has been highly conserved in man. The three sequenced human paired domains belong to two different classes that correspond to the *prd–gsb* and the P29 class of *Drosophila*. This conservation not only of the paired domain proper but also of various paired domain classes in different organisms may reflect a high degree of homology among the gene networks to which these domains belong. Another example of a P29 class paired domain has recently been found in the mouse Pax 1 gene (Deutsch *et al.*, 1988). Interestingly, Pax 1 is expressed, like its most closely related *Drosophila* paired domain homologue P29 (Bopp *et al.*, in preparation), exclusively in mesodermal tissues. An extremely high degree of conservation has been observed between the human (HuP48) and mouse P29-type paired domains (Pax 1) which are identical in 127 out of 129 amino acids and are presumably preceded by a stretch of at least ten additional identical amino acids. Hence, it would be interesting to know whether the two genes are homologous in their remaining sequences as well.

Truly homologous genes would be genes whose *cis*-regulatory and protein coding domains are all homologous. It is not clear at present whether such homologues exist frequently or only rarely in different organisms and whether they reflect isomorphic gene networks. Nevertheless, it is important to know to what extent genes of different organisms sharing a homologous domain are homologous in their remaining domains. In this respect, it was interesting to note that the two human genes containing a *prd–gsb* type paired domain, HuP1 and HuP2, share at least one additional homologous octapeptide, HSIAGILG, with the two *gsb* genes of *Drosophila*. The question of whether HuP1 and HuP2 are true *gsb* homologues has not been answered definitively because, although no homeo domains were found within the cloned regions of HuP1 and HuP2, the possibility remains that such domains are located on more distant exons. In fact, our concept predicts that genes harbouring both a homeo and a paired domain should also be found in man for both domains are characteristic for genes belonging to the same network and hence are likely to be combined in genes of all organisms that have conserved them. Since none of the three isolated human paired domain clones appears to encode a homeo domain, this prediction of the gene network concept still awaits confirmation.

The function of the paired domain is not known. Its association with a homeo domain, in *prd* and *gsb*, suggests a gene regulatory role. Two regions stand out particularly when predictions of secondary structure are examined for all nine known paired domains (Bopp *et al.*, 1986 and in preparation; Deutsch *et al.*, 1988). One region, amino acids 23–31, consists of a highly amphipathic α-helix (Figure 4), the other, amino acids 80–105, exhibits a helix-turn-helix motif (Figure 6). The first helix of the helix-turn-helix structure is again highly amphipathic whereas the region between the two helices is characterized by a high flexibility and very poor helical or β-pleated sheet structures indicative

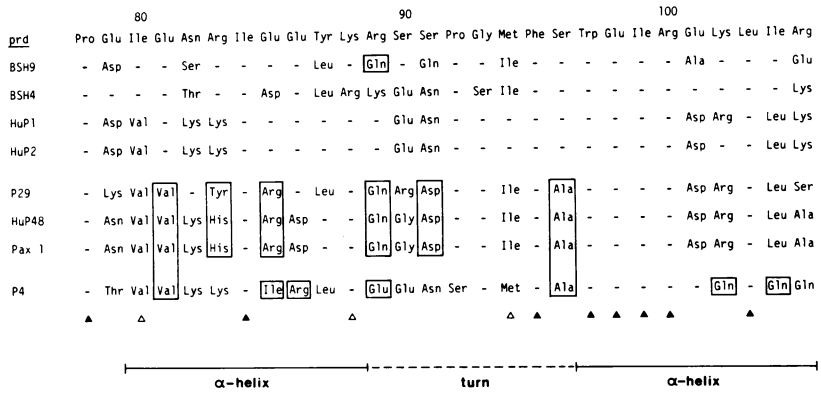

Fig. 6. Prediction of a helix-turn-helix structure within the paired domain. The regions of all known paired domains [of the *Drosophila* genes *prd*, *gsb* (BSH9 and BSH4) (Bopp *et al.*, 1986), P29 and P4 (Bopp *et al.*, in preparation); of the human genes HuP1, HuP2 and HuP48; and of the mouse Pax 1 gene (Deutsch *et al.*, 1988)] between amino acids 78 and 105, for which a helix-turn-helix motif is predicted, are shown. The accuracy of the secondary structure prediction of the Garnier–Osguthorpe–Robson method has been improved by a program of Thomas Niermann which is based on averaging the predicted values of all homologues at corresponding positions (Crawford *et al.*, 1987; T.Niermann, personal communication). The regions comprising the highest predicted α-helical values are indicated by a solid line whereas the region exhibiting highest flexibility indicative of a turn is represented by a broken line. It is possible, however, that the α-helices are slightly longer at the expense of the turn region. Amino acids identical to those shown at the top for the *prd* gene are represented by a dash. Amino acids that deviate at positions highly conserved in the *prd–gsb* class of paired domains are boxed. Positions exhibiting strictly conserved amino acids or those only tolerating conservative changes are indicated by filled and open triangles.

of a turn. It may be significant that the helix-turn-helix region is frequently flanked but never interrupted by introns (Figure 4). Moreover, the insertion of a Gln at position 75 in the paired domain of P4 (Bopp *et al.*, in preparation) and possibly of HuP1 and HuP2 and the deletion of three amino acids at positions 111–113 in the BSH4 paired domain (Figure 4) while being close to each end are still outside the predicted helix-turn-helix structure, possibly reflecting a requirement to preserve the helix-turn-helix region as an entity.

A helix-turn-helix motif has been proposed for the carboxy-terminal half of the homeo domain to function as a DNA binding domain (Laughon and Scott, 1984) based on its similarity to known prokaryotic helix-turn-helix structures that bind DNA (for a recent review see Schleif, 1988). The predicted helix-turn-helix region of the paired domain exhibits α-helices of similar lengths as that of the homeo domain but has a considerably longer turn region (eight versus three amino acids). Even though the helix-turn-helix region of the paired domain is not homologous to that of the homeo domain, it consists, like the homeo domain, of one highly conserved α-helix, the WEIRD helix (which is the second α-helix of the helix-turn-helix region), and another that contains several class-specific differences in its amino acid sequence (Figure 6). Hence it is attractive to speculate that the class-specific region binds DNA as well. If true, the first helix might serve as the helix recognizing specific DNA sequences while the WEIRD helix might be required for proper positioning, the reverse of the order seen in the homeo domain.

If the paired like the homeo domain binds to specific DNA sequences, the products of genes like *prd* and *gsb* that contain both domains would have two DNA binding domains. Such an arrangement would offer interesting regulatory possibilities, such as switching between two DNA binding sites depending on the protein concentration or on the presence of additional interacting protein factors. However, in the absence of direct evidence for the DNA binding function of the paired domain, other possible functions must be considered, including a role in protein–protein interaction. It may be possible to distinguish between these two roles for the paired domain—DNA binding gene regulator or gene regulator in combination with DNA binding domain—by DNA binding studies of the products of the two *Drosophila* genes, P29 and P4, whose paired domains are not associated with a homeo domain, and of the paired domain alone.

Materials and methods

Hybridization of paired domain to genomic DNA blots at reduced stringency

*Eco*RI digests of the genomes from various organisms were analysed on 0.6% agarose gels in TBE buffer according to standard procedures (Maniatis *et al.*, 1982). The DNA was transferred under denaturing conditions to a Zetabind matrix (G.McMaster, to be published) and hybridized with 2 × 10^6 d.p.m./ml of a heat denatured paired domain probe of the BSH9 gene (*Eco*RV–*Bam*HI fragment of BSH9c2; Baumgartner *et al.*, 1987) 'oligo-labelled' (Feinberg and Vogelstein, 1984) with [α-^{32}P]dATP (3000 Ci/mmol). Hybridization occurred in 5 × SSC, 3.5 × Denhardt's, 0.1% SDS, 0.05% $Na_4P_2O_7 \cdot 10H_2O$, 0.2 mg/ml yeast RNA, 50 U/ml heparin for 18 h at 62°C. The filter was washed twice in 5× SSC, 0.1% SDS, twice in 2 × SSC, 0.1% SDS, and finally twice in 1 × SSC, 0.1% SDS for 15 min each time at room temperature.

Isolation of paired domain clones from human genomic library

A human genomic library in EMBL3 (a generous gift of Anne-Maria Frischauf) was blotted to nitrocellulose filters according to Benton and Davis (1977) and screened for paired domain homologues by hybridization with two paired domains of *Drosophila* belonging to different classes, P_{BSH9} (*Eco*RV–*Bam*HI fragment of BSH9c2; Baumgartner *et al.*, 1987) or P_{P29} (*Mbo*II–*Sac*I fragment of P29c1; Bopp *et al.*, in preparation) as described above except that the hybridization buffer contained 43% formamide and the hybridization temperature was reduced to 37°C. After hybridization, the filters were washed twice for 20 min in 2 × SSC, 0.1% SDS at 37°C.

Localization of paired domains on isolated genomic clones and DNA sequencing

To localize the sequences containing the paired domains on the isolated genomic clones more precisely, fragments exhibiting hybridization with P_{BSH9} or P_{P29} at reduced stringency were determined by Southern blot analysis using the detailed restriction maps shown in Figure 2b–d. The 3′-end exon of the paired domain of HuP2 was localized by hybridization at reduced stringency with a subclone of the previously sequenced 3′-end

exon of HuP1. All DNA sequences were read on both strands by the dideoxynucleotide method of Sanger *et al.* (1977), using the M13 vectors mWB3296 and mWB3226 (Frigerio *et al.*, 1986; Baumgartner *et al.*, 1987) derived from M13 vectors described by Barnes *et al.* (1983). For some sequences, the sequencing strategy described by Henikoff (1984) was used.

Acknowledgements

We are indebted to Gary McMaster for his generous help and patience in the preparation and hybridization of the genomic blot, using Zetabind as a matrix, and for providing human lymphocyte DNA. We thank Anne-Maria Frischauf for the human genomic library and Thomas Niermann for his improved analysis of the paired domain with respect to secondary structure predictions. We would also like to thank Thomas Gutjahr for the preparation of cockroach and *D.simulans* DNA, Elliot Meyerowitz, Fritz Müller, Alcide Barberis, Joseph Schwager, Robert Schwartz, Monte Westerfield and Doriano Fabbro for providing DNA samples of *Arabidopsis, Parascaris, P.miliaris, X.laevis*, chicken, zebra fish and man. We are grateful to Hans Noll, Leslie Pick and Elisabeth Jamet for criticism and a careful reading of the manuscript. This work was supported by the Swiss National Science Foundation grant 3.348-0.86 and by a fellowship of the Geigy-Jubiläumsstiftung (to D.B.)

References

Barnes,W.M., Bevan,M. and Son,P.H. (1983) *Methods Enzymol.*, **101**, 98–122.

Baumgartner,S., Bopp,D., Burri,M. and Noll,M. (1987) *Genes Dev.*, **1**, 1247–1267.

Benton,W.D. and Davis,R.W. (1977) *Science*, **196**, 180–183.

Bopp,D., Burri,M., Baumgartner,S., Frigerio,G. and Noll,M. (1986) *Cell*, **47**, 1033–1040.

Crawford,I.P., Niermann,T. and Kirschner,K. (1987) *Prot. Struct. Funct. Genet.*, **2**, 118–129.

Deutsch,U., Dressler,G.R. and Gruss,P. (1988) *Cell*, **53**, 617–625.

Feinberg,A.P. and Vogelstein,B. (1984) *Anal. Biochem.*, **137**, 266–267.

Frigerio,G., Burri,M., Bopp,D., Baumgartner,S. and Noll,M. (1986) *Cell*, **47**, 735–746.

Frischauf,A.-M., Lehrach,H., Poustka,A. and Murray,N. (1983) *J. Mol. Biol.*, **70**, 827–842.

Frohnhöfer,H.G. and Nüsslein-Volhard,C. (1986) *Nature*, **324**, 120–125.

Henikoff,S. (1984) *Gene*, **28**, 351–359.

Keller,E.B. and Noon,W.A. (1985) *Nucleic Acids Res.*, **13**, 4971–4981.

Laughon,A. and Scott,M.P. (1984) *Nature*, **310**, 25–31.

Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) *Molecular Cloning. A Laboratory Manual.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Nüsslein-Volhard,C. and Wieschaus,E. (1980) *Nature*, **287**, 795–801.

Padgett,R.A., Grabowski,P.J., Konarska,M.M., Seiler,S. and Sharp,P.A. (1986) *Annu. Rev. Biochem.*, **55**, 1119–1150.

Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA.*, **74**, 5463–5467.

Schleif,R. (1988) *Science*, **241**, 1182–1187.

Wieringa,B., Hofer,E. and Weissmann,C. (1984) *Cell*, **37**, 915–925.

Received on October 10, 1988; revised on January 13, 1989