Proc. Natl. Acad. Sci. USA
Vol. 88, pp. 263–267, January 1991
Developmental Biology

The Polycomb protein shares a homologous domain with a heterochromatin-associated protein of *Drosophila*

(homeotic gene/repressor/chromatin)

RENATO PARO* AND DAVID S. HOGNESS†

*Zentrum für Molekulare Biologie, Universität Heidelberg, Im Neuenheimer Feld 282, D-6900 Heidelberg, Federal Republic of Germany; and †Department of Developmental Biology, Stanford University School of Medicine, Stanford, CA 94305-5427

Contributed by David S. Hogness, October 16, 1990

ABSTRACT **The Polycomb (*Pc*) gene of *Drosophila melanogaster* is a member of a large class of genes (Pc group) required for the segment-specific repression of homeotic selector genes. Mutations in *Pc*-group genes show strong posterior transformations in homozygous embryos resulting from an ectopic expression of homeotic genes in segments where they are not supposed to be active. Genetic evidence suggests that *Pc* is part of a cellular memory mechanism responsible for the transmission of the homeotic expression pattern through developmental time. We have determined the nucleotide sequence for the genomic DNA of the *Pc* gene and for cDNAs corresponding to the 2.5-kilobase *Pc* mRNA. The deduced sequence of the *Pc* protein exhibits a homology to the heterochromatin-associated protein HP1 encoded by the suppressor of position effect variegation gene *Su(var) 205*. The homology is confined to a 37-amino acid domain in the N-terminal part of the two proteins. Our findings extend to the molecule level the genetically identified parallels between the *Pc*-group genes and the modifiers of position effect variegation. This suggests that *Pc* could use analogous mechanisms at the level of the higher order chromatin structure for the stable transmission of a determined state, as has been proposed for the clonal propagation of heterochromatin domains.**

The metameric organization of the *Drosophila* embryo is dependent on the correct spatial expression of the homeotic genes of the Antennapedia (ANT-C) and Bithorax (BX-C) complexes (1–3). The Polycomb (*Pc*) gene together with other members of the *Pc*-group genes act formally as repressors of the homeotic genes (1, 4–6). Homozygous mutants of *Pc* die at the end of embryogenesis, showing generally posterior transformations of head and body structures. This phenotype has been interpreted as resulting from ectopic expression in anterior segments of homeotic genes normally expressed only more posteriorly (1, 4). The negative regulatory function of *Pc* on homeotic selector genes is required throughout most of development: *Pc* mutants show a maternal effect, and, zygotically, the *Pc* gene product is required during the entire larval period for normal adult development (7–11).

In situ localization of homeotic gene RNAs and proteins has confirmed that in Pc^- embryos, these genes are ectopically expressed in segments anterior to their normal domain of expression (12, 13). However, molecular analyses revealed that the initial pattern of homeotic gene expression was normal up to the end of germ band extension; only after this embryonic stage does the normal pattern degenerate to yield anterior expression (ref. 14; R.P., unpublished data). Similar results were obtained for the gene extra sex combs (*esc*), another member of the *Pc*-group genes, where in addition it could be shown that the delayed effect was not due to the persistence of the maternal component (15). These results suggest that *Pc* and other *Pc*-group genes are responsible for the transmission through the rest of development of the early homeotic gene pattern determined by the transient expression of the maternal and segmentation genes.

In *Drosophila* the phenomenon of position effect variegation points to a mechanism that can stably turn off genes by heritable features of the chromatin (16). Euchromatic genes placed next to heterochromatin by chromosomal rearrangements can become inactivated by a spreading effect of the heterochromatin. The influence of heterochromatin on neighboring euchromatin can occur to a different extent in different cells. Once established, however, the pattern is stably and clonally inherited, giving rise to a variegated pattern. Similar mechanisms seem to effect the inactivation of the X chromosome in mammals (17). Several dominant *Drosophila* mutants that enhance or suppress variegation have been identified (16). Their gene products are thought to be structural components of heterochromatin. Based on the gene dose sensitivity of this class of genes, Tartof and colleagues (18) suggested that their protein products are constituents of multimeric complexes that spread along the chromosome by forming concatenated structures. The *Pc* group of genes and the family of modifiers of variegation share the property that their effect on gene inactivation is dose dependent.

In this work we have characterized a gene product of the Polycomb (*Pc*) gene.‡ We show that the major 2.5-kilobase (kb) mRNA encodes a 390-amino acid (aa) protein and that this *Pc* protein contains a domain homologous to one within a heterochromatin-associated protein of *Drosophila* (19). This finding extends the similarities between the *Pc*-group genes and the modifiers of variegation to the molecular level, suggesting that *Pc* may be involved in the stable transmission of a determined state by its effects on chromatin structure.

MATERIALS AND METHODS

Isolation of cDNAs and Sequencing. cDNAs were isolated from a 3- to 12-hr embryonic library (E6–8) kindly provided by L. Kauvar (20). Plaques (2×10^5) were screened with the *Eco*RI–*Hin*dIII fragment encoding most of the *Pc* gene (Fig. 1). Twelve Pc-specific cDNAs were identified and subcloned into the pEMBL8 vector (21). Subclones of the genomic fragments in pEMBL8, as well as full-length cDNAs, were prepared for sequencing by generating overlapping, directed *Exo* III deletions (22). Sequencing from the single-stranded DNA clones was performed using the dideoxy method of Sanger *et al.* (23). Protein homologies were identified using the FASTP program on the Swiss-prot (EMBL Nucleotide Sequence Data Library) and National Biomedical Research Foundation data bases (24).

***In Situ* Hybridization.** Isolated DNA fragments were labeled with digoxigenin-dUTP using the kit and the accom-

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

Abbreviations: aa, amino acid; ORF, open reading frame.

‡The sequence reported in this paper has been deposited in the GenBank data base (accession no. X55702).

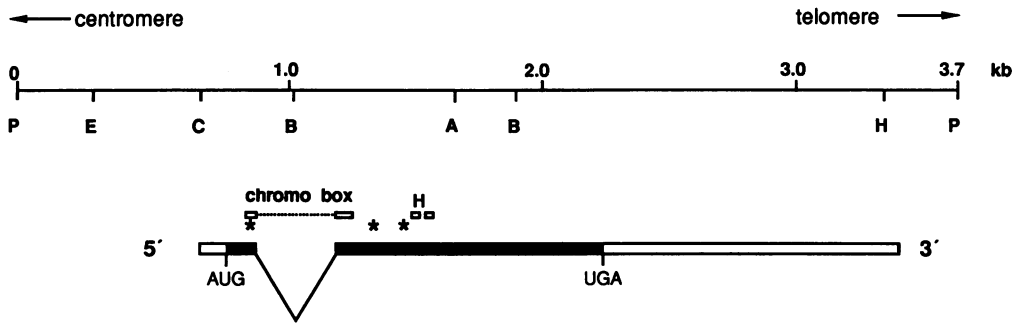

FIG. 1. Structure of the *Pc* locus. The *Pc* locus was cytologically mapped to 78D 7,8. The upper line shows the orientation and the extent of the genomic DNA with the location of some characteristic restriction enzymes (A, *Ava* I; B, *Bal* I; C, *Cla* I; E, *Eco*RI; H, *Hin*dIII; P, *Pst* I). The lower part shows the composite structure of the 2.5-kb *Pc* mRNA, as deduced from the complete sequence of the Pc-12c cDNA, the partial sequence of the Pc-6c cDNA, the sequence of the genomic DNA (see Fig. 2), and S1 nuclease digestion experiments. The open reading frame (ORF) is in black and its AUG start and its UGA stop codons are indicated. Some characteristic features of the *Pc* protein are shown above. The sequence encoding the chromo box is broken by the 313-base-pair (bp)-long intron. The two homopolymeric stretches of 10 and 8 histidines are indicated by H, and the potential nuclear targeting signals deduced from the protein sequences are marked by asterisks.

panying protocol distributed by Boehringer Mannheim. Nucleotides were incorporated for 3 hr at room temperature using the "random primer" technique. Fixation, *in situ* hybridization, and detection of the nonradioactively labeled probes followed the method described by Tautz and Pfeifle (25), with some slight modifications. Young embryos (0–8 hr) were treated with 100 μg of proteinase K per ml (Boehringer Mannheim) for 5 min at room temperature. For older embryos (8–12 hr) the same treatment was prolonged to 13 min. A 1:1200 dilution of the antibody provided by the Boehringer Mannheim kit was used for the detection of the digoxigenin-dUTP. Stained embryos were immediately embedded in the JB-4 medium (Polysciences).

RESULTS

Sequence of the *Pc* Gene. The *Pst* I fragment shown in Fig. 1 encompasses the entire *Pc* gene. We will show elsewhere that this fragment can rescue lethal *Pc* alleles when introduced into the genome by P-element transformation (R.P., J. Lauer, and D.S.H., unpublished data). The *Pc* gene encodes a major 2.5-kb mRNA that is transcribed throughout development. A minor 2.0-kb mRNA is found in poly(A)$^+$ RNA preparations of unfertilized eggs and early (0–4 hr) embryos. Another minor 1.0-kb mRNA was found to be expressed in larval salivary glands (26). We have used the *Eco*RI–*Hin*dIII fragment in Fig. 1 to screen an embryonic cDNA library (20). Two cDNAs (Pc-12c and Pc-6c) that represented a full-length copy of the 2.5-kb mRNA were selected for sequencing. No cDNAs of the minor mRNA classes were isolated, although different cDNA libraries from appropriate stages were screened. Fig. 1 shows the structure of the 2.5-kb mRNA as determined by comparing the genomic sequence of the 3.7-kb *Pst* I fragment and the sequence of the full-length cDNAs. Only one intron separates the smaller 5′ exon from the 3′ exon. This arrangement was confirmed by S1 nuclease protection experiments using various genomic fragments to protect embryonic poly(A)$^+$ RNAs (data not shown). The cDNA sequence revealed an ORF for a 390-aa (44 kDa) protein (Fig. 2).

Characteristics of the *Pc* Protein. The 390-aa *Pc* protein shows a high content of charged amino acids [basic (H, K, R) 20%, acidic (D, E) 15%] scattered over the entire protein. An overall dense distribution of polar amino acids results in a strong tendency toward large hydrophilic domains in hydrophobicity plots (28). Two striking homopolymeric stretches of 10 and 8 histidines are found in the middle of the protein. Considering that histidine can change its charge via small differences in physiological pH, these domains could have dramatic effects on the tertiary structure of the protein. The *Pc* protein has been shown to be a nuclear protein (26). Three potential nuclear targeting sequences underlined in Fig. 2 may be responsible for this localization.

A comparison of the *Pc* protein sequence with different protein data bases did not reveal a homology to functionally characterized protein domains. The only significant homology, excluding those due to the histidine repeats, was to the HP1 protein (formerly C1A9) encoded by the suppressor of variegation gene *Su(var) 205* (19, 29). This homology is restricted to a stretch of 37 aa in the N-terminal portion of both proteins (Figs. 1–3). Since the two similar domains occur in two proteins that are involved in the organization of the chromatin we have named this the "chromo domain" (chromatin organization modifier). The 111 bp encoding the chromo domain is called the "chromo box." Fig. 3 shows that the two chromo domains are to 65% identical over their 37 aa, with conservative replacements accounting for another 19%. No other significant similarities between the two proteins were found in the remaining sequences.

Distribution of the *Pc* Transcripts. Two genomic fragments *Eco*RI–*Ava* I and *Ava* I–*Hin*dIII (Fig. 1) were used separately to detect the distribution of the *Pc* transcripts in embryos. Both fragments showed the same pattern in all stages tested. Northern analysis as well as *in situ* hybridizations showed that *Pc* transcripts are most abundant in unfertilized eggs and early cleavage stage embryos, when the *Pc* mRNA is homogeneously distributed (data not shown). At blastoderm stages *Pc* transcripts are still essentially uniformly distributed, though less staining is evident at the anterior and posterior poles (Fig. 4). *Pc* transcripts are found in almost all cells and tissues throughout gastrulation and organogenesis though at a much lower level than in early syncytial stages. In late stages *Pc* transcripts are predominantly expressed in the central nervous system (data not shown).

DISCUSSION

Mutations in the *Pc* gene show a very complex phenotype with multiple transformations seen in all segments affecting all homeotic genes of the two complexes. Transformations in embryonic and imaginal tissue can be directed posteriorly as well as anteriorly (4, 9, 30). This complexity of the *Pc* phenotype does not seem to be encoded in the organization of the gene. Its structure is relatively simple and its regulatory domain cannot be very extensive, considering that the 3.7-kb *Pst* I fragment (Fig. 1) is sufficient to rescue the lethality associated with the *Pc* mutation and that this fragment has only 750 bp of upstream sequences. The simple ubiquitous expression pattern of the *Pc* transcripts found during embryogenesis might be the result of this small regulatory region. The highest levels of transcripts were detected in very early, nuclear cleavage stage embryos and appear to have

```
GTCGGCCACTCCTTCACTACCTACCTACCTGAGCTTTCGGCGCTCTTGTGCTTTCTCACA   60
CTCTCTCCCGCACTTTCTCTCCCTCTTGAGGATGTGGCATTTGAAGAGTTCGAAGAGGGT  120
AACTAGTTCAGTTGTAATCAAGTTAATTATCAAACTTAATTGAATAGGAATATGGGGCGA  180
AGTATATGTGGCAATTTGTGATACAATAAGTGAAGCAGGACAAAGGAAAATCAAAACTCA  240
ATATGGGATGAAAGCAACACAACGATAAACGTTAACGAAAGTATAATATTTCGTACCTGA  300
ATTCCCAATTCAATTCAACATGCAGTTTAAAAAGAATACTGTACAAAGGATTCATTAAAA  360
ATGCGCTTATTAAAATTTGATTAAAAGCTATTTTATATTTAAGCTAATTGGTTTAACTTC  420
TCGCCGGTTTAGTAATTTGTAGAAACGATAAAAGAAAATAGTGTATAAGGCAACAGCGCC  480
AAAAAAATGTTTTCTAAAACATTTATATACCCCATTTATAAATAATAGAAAAATGTCAAG  540
AAATTTTGAAAATATTTTAACAAAACCATTTACATTTTAAAAACGCACAATTAGGTGTGT  600
CTAGTCATCAAAGTTTTGAGTCTTTGCTAATTTTGCCAGGAAAGGGTGTGCGACACGGCT  660
ACTTGCTAAGATATGATCCCAGTTCTTTGTAGCACGGTAACTCTGCACCTCGCAGCACTA  720
TCGATTGTTTACATGAAAATAATCGAGTCGGACGACTATCGACGTACGCAGAATTGTAAA  780
CCAGAAGTTAATTGCAAATAAAACGAATAATAAAACGTTCCGAGAAGATTATTAATTAAA  840

AATGACTGGTCGAGGCAAGGGGAGTAAGGGGAAGTTGGGGCGCGACAATGCGACCGACGA  900
 M  T  G  R  G  K  G  S  K  G  K  L  G  R  D  N  A  T  D  D

TCCAGTCGATCTAGTGTACGCGGCTGAGAAAATCATCCAAAAGCGCGTTAAGAAGGTGAG  960
 P  V  D  L  V  Y [A  A  E  K  I  I  Q  K  R  V  K  K ]

ATGAGCGTAAACAAATTCGAAAAAGGCGACCAAAACGGGATCGTCCGCTAAAAATGGCAA 1020
ATGCAATATGCGCAGCACACGCAGGATATATGTATAGTCCACAAGGTCACATGTCCTTCG 1080
ACATTTTTTGGCCAGAGGACATCTGCTGCTGCAACCACCATTTGCACGATGCATGTGTGT 1140
GTGCGTAGTGCGTTCGTATGTGTGTGTGTGTGTATGTGAGGCTGTTTTTGTTTCGCTGAA 1200
TGGAAAATCAATTAGTTGTCTAAGAAAATGCAATCCGACCCCCACTCCTTCACTTACTCG 1260

ATTTTCCAGGGCGTCGTGGAGTACCGTGTCAAGTGGAAGGGCTGGAACCAGCGCTACAAC 1320
          [G  V  V  E  Y  R  V  K  W  K  G  W  N  Q  R  Y  N ]

ACCTGGGAACCGGAGGTAAACATCCTGGATCGCCGCCTCATCGACATCTACGAACAAACG 1380
[T  W  E  P  E ]V  N  I  L  D  R  R  L  I  D  I  Y  E  Q  T

AACAAATCCTCCGGAACTCCCTCCAAGCGAGGCATTAAGAAGAAGGAGAAAGAACCCGAT 1440
N  K  S  S  G  T  P  S  K  R  G  I  K  K  K  E  K  E  P  D

CCGGAGCCGGAATCCGAGGAGGATGAGTACACCTTCACAGAAAACGATGTGGACACGCAT 1500
P  E  P  E  S  E  E  D  E  Y  T  F  T  E  N  D  V  D  T  H

CAAGCCACCACCTCATCGGCTACCCACGATAAGGAGTCGAAGAAGGAGAAGAAGCACCAT 1560
Q  A  T  T  S  S  A  T  H  D  K  E  S  K  K  E  K  K  H  H

CACCACCACCACCATCATCACCACATCAAGTCCGAACGCAACAGTGGACGGCGCTCGGAA 1620
H  H  H  H  H  H  H  H  I  K  S  E  R  N  S  G  R  R  S  E

TCTCCGCTGACCCACCATCATCATCACCACCACCACGAGTCCAAGCGTCAGCGCATTGAT 1680
S  P  L  T  H  H  H  H  H  H  H  H  E  S  K  R  Q  R  I  D

CACAGCTCCTCCTCGAACAGCAGCTTCACGCACAACTCATTTGTCCCCGAGCCGGACAGC 1740
H  S  S  S  S  N  S  S  F  T  H  N  S  F  V  P  E  P  D  S

AACTCCTCCAGCTCCGAGGATCAGCCGCTGATCGGCACCAAACGCAAAGCCGAGGTGCTC 1800
N  S  S  S  S  E  D  Q  P  L  I  G  T  K  R  K  A  E  V  L

AAGGAATCGGGCAAGATTGGAGTTACGATTAAGACCTCCCCAGATGGCGCCACTATTAAG 1860
K  E  S  G  K  I  G  V  T  I  K  T  S  P  D  G  P  T  I  K

CCTCAGCCGACCCAGCAGGTAACTCCCAGCCAGCAGCAGCCCTTCCAGGATCAGCAGCAA 1920
P  Q  P  T  Q  Q  V  T  P  S  Q  Q  Q  P  F  Q  D  Q  Q  Q

GCGGAAAAGATTGCCAGCGAGGCTGCAACGCAGCTGAAATCTGAGCAGCAGGCCACCCCA 1980
A  E  K  I  A  S  E  A  A  T  Q  L  K  S  E  Q  Q  A  T  P

CTGGCCACAGAAGCCATAAACACAACGCCCGCAGAGTCCGGAGCTGAGGAAGAAGAAGTA 2040
L  A  T  E  A  I  N  T  T  P  A  E  S  G  A  E  E  E  E  V
```

FIG. 2. *(Figure continues on the opposite page.)*

```
GCCAACGAGGAAGGCAACCAACAGGCGCCACAGGTTCCTTCCGAGAACAATAACATACCA 2100
A  N  E  E  G  N  Q  Q  A  P  Q  V  P  S  E  N  N  N  I  P

AAACCGTGCAACAACCTGGCTATCAACCAGAAACAGCCGCTTACTCCTCTTTCGCCGCGT 2160
K  P  C  N  N  L  A  I  N  Q  K  Q  P  L  T  P  L  S  P  R

GCCCTTCCGCCGCGCTTCTGGCTGCCGGCCAAATGCAACATATCCAACCGGGTGGTGATC 2220
A  L  P  P  R  F  W  L  P  A  K  C  N  I  S  N  R  V  V  I

ACAGACGTCACCGTAAACCTGGAGACCGTTACCATACGCGAGTGCAAGACGGAGCGCGGG 2280
T  D  V  T  V  N  L  E  T  V  T  I  R  E  C  K  T  E  R  G

TTTTTTCGTGAGCGCGACATGAAAGGCGATTCGTCGCCAGTAGCTTGAGCTTCGACCCAA 2340
F  F  R  E  R  D  M  K  G  D  S  S  P  V  A

ACAATGTCGGGAAAACAAACCGGAAAACTGATACAATGTAAATAGACAAACCAAAATGCG 2400
AGAAAGTTGAGAGAAGAATTATGCAGAGAAATCCACACTAATGTTGCTAAGCATTTTCTC 2460
TTTAGGACTTAACTAATTTATTGTATTATTTTTAGCGTAGGTTTTAGTTTGTTGTTTTAT 2520
TACAATGTTTAATGTGCAAGAGTAGTTTAAGCAAGCGCAAGTCGGGTATTTACTAATGAC 2580
TTTATATATACATAAATTTTCGCTATTTTTTTATCAGCATTGCAATGAATATTTTCCTAT 2640
TTAAATTAAAAATTTATTTTTCTGCAATATCACTTTTTAAGATTAATCTGGTTGTCATTT 2700
AGTGAATAGGCGAAATGCTGCCCGTGCTTAACTTGCCAATGCAATAGATTGTAAAACTTG 2760
CAGCGTAGATAAATATGTATGTACATGTTACATATATTTACCACATATTGTATATTTTAT 2820
AACACTTGAATCTGATTGAAAACAACACAAGACACTCAAATGCTAAGAAAACTCCATGCG 2880
AAACCGAAGCAATGTTGATTACGCTTACAAATATATACAATAGAAGATGCATATATTCGA 2940
GTAGTCACCACCACGTTTGCCGAGTTGGCTATCCCAATGTTGTAATTGTTTAATATTATA 3000
TTTCCTCGCGGCCGATGATTTTACAAAAGAAAACCCCAAAAACATATACACAATTGTAGA 3060
TTTACAAATATTTTTAGACTTTAGCTAACCAAAATCTCTACGTAATATTTATTTTAACAT 3120
GTTCTCTTAGCAGTCATTCAAGACTCAAGTGCTGCCTTTATGGGTTCAGAGCGGTGCAGT 3180
GCACAGATAATTCCAAATCGAGCCCATAGACTCATTACTTAAATCCTATGTCTATTTCAC 3240
AATTTTTTTTCACGCTATGAAACAGAAAACCTACGTGGGAGCTGCGGGGGAATTTACAAA 3300
CTCCTGCTTATTTCCCATGGGTCCTGTGCGTATGTTTCTATTTGTCTTAGCTCTTTATTG 3360
GCTAGTTTTAGTTACGGCGTAAATTCCACACAAAACAGCAATTCAAAGCGTATATTTTAA 3420
GCTTTTTAATATAATTTTAACATTAATGTTATTGCCGGTATTGATTGTACATTAAAAAAA 3480
AAATAGAAGCAACCAATTTTCCAATGTTTAGGTATTGTCTACAACTGAGATTAATTTTCA 3540
TATAATATCCAGTAAATATCATCAATCTGCGTTGGGTTAAATTAAACATTAGGAGTAAGG 3600
TACACGAATCCGCTAGTTTGTACACGGGTCTATATGTCGAGGTCTATGTTAAGGCGGTCT 3660
CATAAAGGCCGGTACCTATTCATGAAGGAGTTGATGATAAGGCGCCACTGCA          3712
```

FIG. 2. Nucleotide sequence of the *Pc* gene and amino acid sequence of the *Pc* protein. The genomic sequence of the *Pst* I fragment encoding the entire *Pc* gene is shown. The sequence starts at the first nucleotide after the restriction enzyme cut and ends at the last nucleotide before the restriction enzyme cut. The sequence that corresponds to the Pc-12c cDNA is underlined. Partial sequencing of the Pc-6c cDNA showed the same structure as Pc-12c, except for the starting nucleotide, which was displaced by two nucleotides to position +735 and a shorter poly(A)$^+$ tail. The two overlapping poly(A) signals at positions +3473 and +3478 are shaded. The ORF is shown by the deduced *Pc* protein sequence. The starting methionine codon is positioned at +842. Using the Framescan program of Staden and McLachlan (27) and a set of standard *D. melanogaster* codon usages (K. Burtis and D.S.H.), the indicated ORF exhibited a typical *D. melanogaster* bias. The chromo domain defined by the homology to the HP1 protein encoded by the *Su(var) 205* gene is boxed. The three potential nuclear targeting signals are underlined. The last signal is followed by the first histidine repeat between +1555 and +1584. The second shorter histidine repeat is located between +1633 and +1656. The *Pc* protein contains six potential sites for N-linked glycosylation at positions +887 (aa 16), +1381 (aa 76), +1696 (aa 181), +1741 (aa 196), +1999 (aa 282), and +2208 (aa 348).

been deposited during oogenesis (R.P., J. Lauer, and D.S.H., unpublished data). At cellular blastoderm the *Pc* transcripts are expressed in all cells of the embryo, though slightly less expression is visible toward the poles. Also at later embryonic stages transcripts are essentially homogeneously distributed, but present in a much lower amount than in early stages. At no stage do we see any evidence for a graded distribution in the anterior-posterior axis of the *Pc* transcript, as had been proposed for a regulatory function of *Pc* (1). Our findings suggest that *Pc* does not exercise its complex role in the homeotic expression pattern by being transcribed in a complicated spatial pattern. Rather, it would appear that *Pc* is needed by every cell and only its interaction with other more segment- or tissue-specific factors results in the differential regulation of the homeotic genes. Lack of *Pc* gene product would then disturb a large network of interacting gene products, leading to the multiple effects that are seen in *Pc*$^-$ phenotypes. In addition, the complex phenotype also results from the sum of multiple, cross-regulatory interactions of ectopically expressed transcription factors, in particular those of the Antennapedia (ANT-C) and Bithorax (BX-C).

The structure of the 2.0-kb and 1.0-kb *Pc* mRNAs has not yet been determined. Their function during development would appear, however, to be minor. In contrast to the 2.5-kb mRNA, the 2.0-kb transcript is strictly maternally expressed (R.P., J. Lauer, and D.S.H., unpublished data). Genetic experiments have shown that the lack of the maternal *Pc* component can be rescued by a wild-type paternal gene, resulting in normally developed embryos (8, 10). The 1.0-kb transcript has so far been found only in one particular tissue, the larval salivary glands (26).

The amino acid sequence of the *Pc* protein encoded by the 2.5-kb mRNA revealed several distinct features. The high content of polar amino acids gives the protein a very strong hydrophilic appearance in calculated hydrophobicity plots. Three potential nuclear targeting signals are localized in the N-terminal half of the protein (31). The two long histidine repeats in the middle of the protein are an addition to the many different homopolymeric stretches found in other *Drosophila* nuclear proteins, which also includes the similar histidine-rich *paired* repeat (32). It is still an open question what specific functions, if any, these repeats encode. The

```
                ▼
26YAAEKIIQKRVKKGVVEYRVKWKGWNQRYNTWEPEVN62
  :: ::::::::::::: ::: .::::. .  :::::: :
24YAVEKIIDRRVRKGKVEYYLKWKGYPETENTWEPENN60
                ▲
```

FIG. 3. Homologous domains between the *Pc* protein and the HP1 (C1A9) protein (*Pc* is the upper sequence). The homology extends over 37 aa between residues 26 and 62 in the *Pc* protein and residues 24 and 60 in the HP1 (C1A9) protein (19). Identical residues are marked with two dots, whereas conservative changes are marked by one dot. The chromo domain is defined by a 65% (24/37) amino acid identity and 7/37 conservative changes. The potential nuclear targeting signal is conserved in both proteins and is underlined. The position of the intron separating the coding sequence of the chromo box in *Pc* and *Su(var) 205* is shown by the arrowhead.

FIG. 4. Distribution of *Pc* transcripts. Whole-mount wild-type embryo at the syncytial blastoderm stage (stage 4) hybridized with a digoxigenin-labeled *Pc* probe (*Ava* I–*Hin*dIII in Fig. 1). Anterior is to the left and dorsal is to the top. The focus is on the middle of the embryo. *Pc* transcripts are distributed essentially homogeneously over the entire embryo.

most striking feature in the sequence of the *Pc* protein, however, is the homology to the HP1 protein encoded by the suppressor of variegation *Su(var) 205*. This homology is confined to a 37-aa region defining the chromo domain. The 313-bp intron of the *Pc* gene splits the chromo box, encoding this domain just after the coding sequences for the first potential nuclear targeting signal. It is interesting to note that the position of the intron in the HP1 chromo box is in approximately the same position (Fig. 3). The potential nuclear targeting signal of the chromo domain appears to be conserved in both proteins (Fig. 3). The function of these chromo domains, defined in this work by their homology, is unknown since functional domains have not been identified in either protein. It will be interesting to see whether the chromo domain is retained in other *Drosophila* proteins, or even in other organisms, and to determine the functional relatedness of such proteins.

The molecular relationship that we find between a *Pc*-group gene and a gene of the *Su(var)* class could point to a common mechanism of gene regulation. The HP1 protein is associated with heterochromatin (19). The finding that it is encoded by the *Su(var) 205* gene suggests that it is a component of the multimeric complexes proposed to be involved in the packaging and in the clonal heritability of heterochromatic chromosomal domains. Polycomb together with other proteins of the *Pc* group might use similar mechanisms for the stable transmission of a determined state of a cell, as defined by its specific set of expressed determinants (i.e., homeotic transcription factors) (33). This process of cellular memory could use heterochromatin-like multimeric domains to keep homeotic selector genes repressed in regions of the embryo where they were not activated by the early maternal and segmentation genes. The homeotic expression pattern locked into such a stable conformation of chromatin would be maintained through the rest of development by the same heritable features used by the heterochromatin. It will be of interest to define the functional role of the chromo domain within such a chromatin complex and to see whether it gives us a molecular handle to isolate other components involved in the phenomenon of cellular memory.

We thank Christine Beisel and Ken Relloma for technical assistance and Kenneth Burtis and Thomas Hoffmann for help with the computer searches. R.P. was supported by a fellowship of the Swiss National Science Foundation while at Stanford and by grants from the Bundesministerium für Forschung und Technologie and the Deutsche Forschungsgemeinschaft in Heidelberg. The work at Stanford was supplemented by National Institutes of Health grants to D.S.H.

1. Lewis, E. B. (1978) *Nature (London)* **276**, 565–570.
2. Kaufman, T. C. (1983) in *Time, Space, and Pattern in Embryonic Development*, eds. Jeffrey, W. R. & Raff, R. A. (Liss, New York), pp. 365–383.
3. Duncan, I. M. (1987) *Annu. Rev. Genet.* **21**, 285–320.
4. Duncan, I. M. & Lewis, E. B. (1982) in *Developmental Order: Its Origin and Regulation*, eds. Subtelny, S. & Green, P. B. (Liss, New York), pp. 533–554.
5. Capdevila, M. P. & Garcia-Bellido, A. (1981) *Wilhelm Roux Arch. Dev. Biol.* **190**, 339–350.
6. Jürgens, G. (1985) *Nature (London)* **316**, 153–155.
7. Denell, R. E. (1982) *Dev. Genet.* **3**, 103–113.
8. Haynie, J. L. (1983) *Dev. Biol.* **100**, 399-411.
9. Sato, T. & Denell, R. E. (1985) *Dev. Biol.* **110**, 53–64.
10. Lawrence, P. A., Johnston, P. & Struhl, G. (1983) *Cell* **35**, 27–34.
11. Struhl, G. (1981) *Nature (London)* **293**, 36–41.
12. Wedeen, C., Harding, K. & Levine, M. (1986) *Cell* **44**, 739–748.
13. Beachy, P. A., Helfand, S. L. & Hogness, D. S. (1985) *Nature (London)* **313**, 545–551.
14. Kuziora, M. A. & McGinnis, W. (1988) *EMBO J.* **7**, 3233–3244.
15. Struhl, G. & Akam, M. (1985) *EMBO J.* **4**, 3259–3264.
16. Spofford, J. B. (1976) in *Genetics and Biology of Drosophila*, eds. Ashburner, M. & Novitski, E. (Academic, New York), Vol. 1C, pp. 955–1018.
17. Gartler, S. M. & Riggs, A. D. (1983) *Annu. Rev. Genet.* **17**, 155–190.
18. Locke, J., Kotarski, M. A. & Tartof, K. D. (1988) *Genetics* **120**, 181–198.
19. James, T. C. & Elgin, S. C. R. (1986) *Mol. Cell. Biol.* **6**, 3862–3872.
20. Poole, S. J., Kauvar, L. M., Drees, B. & Kornberg, T. (1985) *Cell* **40**, 37–44.
21. Dente, L., Cesareni, G. & Cortese, R. (1983) *Nucleic Acids Res.* **11**, 1645–1654.
22. Henikoff, S. (1984) *Gene* **28**, 351–359.
23. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
24. Lipman, D. J. & Pearson, W. R. (1985) *Science* **227**, 1435–1441.
25. Tautz, D. & Pfeifle, C. (1989) *Chromosoma* **98**, 81–85.
26. Zink, B. & Paro, R. (1989) *Nature (London)* **337**, 468–471.
27. Staden, R. & McLachlan, A. D. (1982) *Nucleic Acids Res.* **10**, 141–156.
28. Kyte, J. & Doolittle, R. F. (1982) *J. Mol. Biol.* **157**, 105–132.
29. Eissenberg, J. C., James, T. C., Foster-Hartnett, D. M., Hartnett, T., Ngan, V. & Elgin, S. C. R. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 9923–9927.
30. Capdevila, M. P., Botas, J. & Garcia-Bellido, A. (1986) *Wilhelm Roux Arch. Dev. Biol.* **195**, 417–432.
31. Chelki, D., Ralph, R. & Jonak, G. (1989) *Mol. Cell. Biol.* **9**, 2487–2492.
32. Frigerio, G., Burri, M., Bopp, D., Baumgartner, S. & Noll, M. (1986) *Cell* **47**, 735–746.
33. Paro, R. (1990) *Trends Genet.*, in press.