Cell, Vol. 40, 37–43, January 1985, Copyright © 1985 by MIT
0092-8674/85/010037-07 $02.00/0

The *engrailed* Locus of Drosophila: Structural Analysis of an Embryonic Transcript

Stephen J. Poole, Lawrence M. Kauvar,
Barry Drees, and Thomas Kornberg
Department of Biochemistry and Biophysics
University of California, San Francisco
San Francisco, California 94143

Summary

cDNA clones originating from the *engrailed* gene of Drosophila have been isolated from recombinant phage libraries that were made using poly(A)$^+$ RNA extracted from early embryos. The DNA sequence of one of these clones includes a homeo box, a 180 bp sequence present in several other Drosophila genes important in formation of body pattern during development. The homeo boxes found in the other Drosophila genes, as well as in cognate sequences from a wide range of segmented animals, including higher vertebrates, are highly conserved. By contrast, the homeo box within the *engrailed* gene diverges substantially and, unlike the other homeo boxes, is interrupted by an intervening sequence. The *engrailed* homeo box is located near the 3′ end of a 1700 bp open reading frame. If translated, this sequence would produce a protein of unusual composition. We also show that a neighboring gene has a large region with strong homology to *engrailed*, and that it also contains a homeo box.

Introduction

engrailed is one of a number of genes involved in specifying the body pattern of Drosophila melanogaster. Embryos homozygous for many alleles of *engrailed* display a severely disrupted segmentation pattern and die prior to hatching (Kornberg, 1981a). In these embryos pairs or large groups of segments fuse because the segmental borders are not maintained. The requirement for wild-type *engrailed* function extends beyond the embryonic period to later developmental stages as well. Adult flies that either bear *engrailed* mutations that are not lethal (e.g., *en*^{1}), or lack *engrailed* function only in mosaic patches of tissue homozygous for embryonic lethal alleles, are unable to develop normally in many areas of the body. Specific defects have been observed in each of the adult segments examined and a remarkable position dependence has been noted. In every segment the abnormalities occur within the posterior portion only (Lawrence and Morata, 1976; Kornberg, 1981a, 1981b; Lawrence and Struhl, 1982), an area that coincides with the posterior developmental compartment (Garcia-Bellido et al., 1973). No *engrailed*-related defect has ever been observed in the cells of the anterior compartments: cells in the anterior compartments develop normally in the absence of *engrailed* function.

Among the many other genes whose mutant phenotypes suggest controlling roles in Drosophila development, two gene clusters, the Antennapedia complex and the Bithorax complex, stand out. Mutations in these "homeotic" genes produce normal structures in abnormal locations. Flies with an *Antennapedia* mutation have distal leg structures in place of normal distal antennal structures (Kaufman et al., 1980). Flies bearing simultaneous mutations in the *bx* and *pbx* regions of the Bithorax complex have a wing in place of the normal haltere (Lewis, 1978). Within these two gene clusters, which apparently perform similar functions in different parts of the animal, several loci also share a common segment of DNA (McGinnis et al., 1984b). A strongly conserved region of approximately 180 bp, present in several genes in both gene clusters, has been designated the homeo box (McGinnis et al., 1984a). This sequence is strongly conserved among a number of other segmented organisms including humans and Xenopus (Levine et al., 1984; Carrasco et al., 1984). The homeo box contains an open reading frame and conservation among the human, frog, and fly sequences is even stronger at the amino acid level than at the nucleotide level. Thus, the homeo box appears to encode a peptide region of about 60 residues conserved over an immense evolutionary span (Levine et al., 1984; Carrasco et al., 1984).

By using an extensive collection of *engrailed* mutations whose precise cytological aberrations had been mapped to the salivary gland polytene chromosomes (Kornberg, 1981a), DNA spanning the *engrailed* locus has been cloned by chromosomal walking (Kuner et al., unpublished data). Within the approximately 40 kb of DNA in which lethal *engrailed* mutations have been localized, a set of developmentally regulated poly(A)$^+$ transcripts has been identified (Drees and Kornberg, unpublished). In this report, we analyze several cDNA clones representing the most abundant of these transcripts. The *engrailed* gene contains an identifiable homeo box sequence that is divergent when compared to all of the other known homeo boxes. Furthermore, the *engrailed* homeo box is set apart by being interrupted by an intervening sequence. A nearby transcribed gene also contains a similarly divergent and split homeo box.

Results

Isolation of *engrailed* cDNA Clones

Isolation of the *engrailed* locus has made possible a molecular analysis of its structure and function. In the map of the *engrailed* region (Figure 1A), arrows mark the breakpoints of alleles with cytologically visible chromosome rearrangements. Because most *engrailed* mutations are embryonic lethals (Kornberg, 1981a), suggesting embryonic expression, DNA fragments (1–4 kb) from this entire region were used as probes for detection of homologous embryonic RNA. Several such probes hybridize on Northern blots to poly(A)$^+$ embryonic RNAs of 3.6, 2.7, and 1.4 kb.

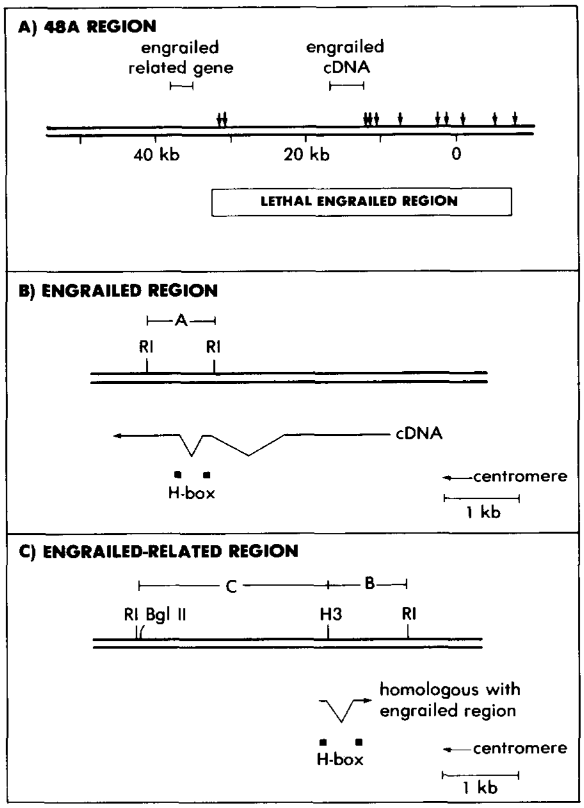

Figure 1. The *engrailed* Chromosomal Region

(A) Overview of the 48A region of the second chromosome. The centromere is to the left of the map, and the zero point is the insertion site of the *en*^{1} transposition (Kuner et al., unpublished). The vertical arrows mark the locations of the mapped *engrailed* chromosomal breakpoint lethal mutations. These mutations are (left to right): en^{SF52}, en^{SF37}, en^{SF42}, en^{SF63}, en^{SF50}, en^{SF32}, en^{2}, en^{SF37}, en^{SF24}, en^{30}, en^{SF61}, and en^{LA3} (Kornberg, 1981a; Kornberg and Ali, unpublished data). Shown above on the right is the region of homology with the *engrailed* cDNA clone. Above and to the left is the location of a related gene that shares some homology with the *engrailed* cDNA clone.

(B) Structure of the *engrailed* cDNA clone c-2.4. The polarity on the chromosome is the same as in (A). The direction of transcription (5′ to 3′) is from right to left. The horizontal lines show the inferred *engrailed* exons that are spliced in the cDNA. Also indicated is the location of the homeo box sequences within the cDNA. The 0.9 kb Eco RI fragment designated A is used as a probe in Figure 6.

(C) Structure of the *engrailed-related* region. The centromere lies to the left and the *engrailed* gene is 17 kb to the right. The regions within the 3.6 kb genomic Eco RI fragment that are homologous to the *engrailed* cDNA clone c-2.4 are shown and include the split homeo box. Direction of transcription of this gene is from left to right. The sequence of the 417 bp presumptive intron splitting the homeo box is not homologous with the 282 bp *engrailed* intron. The fragments designated B and C are used as probes in Figure 6. The Hind III site is within the presumed intron.

cDNA libraries in phage λgt 10 (T. St. John, J. Rosen, and H. Gershenfeld, personal communication) were made from RNA isolated from embryos, larvae, and pupae at various developmental stages (see Experimental Procedures). Several cDNA libraries from early embryonic RNA, as well as one kindly provided by M. Goldschmidt-Clermont and D. Hogness, were screened with a 0.9 kb genomic fragment as probe. This particular fragment was chosen because it contains sequences homologous to the major embryonic transcripts detected with any of the *engrailed* probes tested (see lane A in Figure 6 below). Eleven cDNA clones hybridizing with the 0.9 kb probe were isolated. They range in size from ~1.1 to 2.6 kb. All clones have been examined in detail by digestion with restriction endonucleases and by heteroduplex mapping. Nine are apparently derived from the same transcript; their size and relative abundance suggests that they are copies of the major 2.7 kb RNA, and one of these (c-2.4) has been analyzed in detail.

Structure of the *engrailed* cDNA Clones

Both strands of cDNA clone c-2.4, as well as much of the corresponding genomic DNA and an independently isolated cDNA clone, have been sequenced. With the exception of a few differences attributed to single-base polymorphisms, the homologous sequences are identical. The cDNA clone c-2.4 contains 2449 bp (Figure 2) and has one adenylate-rich end. Comparison with the genomic sequences indicates that two intervening sequences, 1.1 and 0.28 kb in length, had been removed from the RNA template for c-2.4. In the genomic sequences, both the donor splice junction (AGGT) and the acceptor splice junctions (TTXCAG) are similar to other eukaryotic splice junctions (Mount, 1982; see Figure 5 for the splice junctions of the second intron). The splice points of c-2.4 are at positions 1489 and 1587. The location of the introns is diagrammed in Figure 1B.

The sequences at positions 238–336 consist largely of repeating trinucleotides of the form CAX, where X is usually A or G. This CAX repeat is homologous to a repeated element that has been found in numerous Drosophila RNAs, including the *Notch* (S. Artavanis, personal communication) and *Antennapedia* (R. Garber, personal communication) transcripts and may in some instances consist of 100 tandem CAA or CAG repeats (S. Artavanis, personal communication). A restriction fragment that contains this region of the *engrailed* cDNA hybridizes to many genomic restriction fragments (data not shown); other repeating trinucleotides in this region (e.g., GCX, positions 379–420) may also contribute to the multiple bands of hybridization.

The orientation of the poly(A) stretch of c-2.4 and the orientation of c-2.4 with respect to the genomic DNA are consistent with the known direction of transcription of the *engrailed* region (Drees and Kornberg, in preparation). In this direction is a single large open reading frame of about 1700 bp. Although what portion if any of this sequence is translated into protein remains unknown, the first ATG triplet within the cDNA is 90 bp into the open frame (position 178). Translation of the open reading frame, starting from this ATG (see Figure 2), would produce a protein with several unusual features, including stretches of polyglutamine, (positions 253–285, 379–438, 871–897), polyglutamic/aspartic acid (580–630), and polyserine (1135–1152, 1183–1281, 1408–1419).

The nine nucleotides at the 5′ end of c-2.4 do not correspond with the equivalent genomic sequences. Because the 5′ end of the major 2.7 kb transcript has not been unambiguously localized, these nucleotides may be part of a splice junction and an upstream exon. This

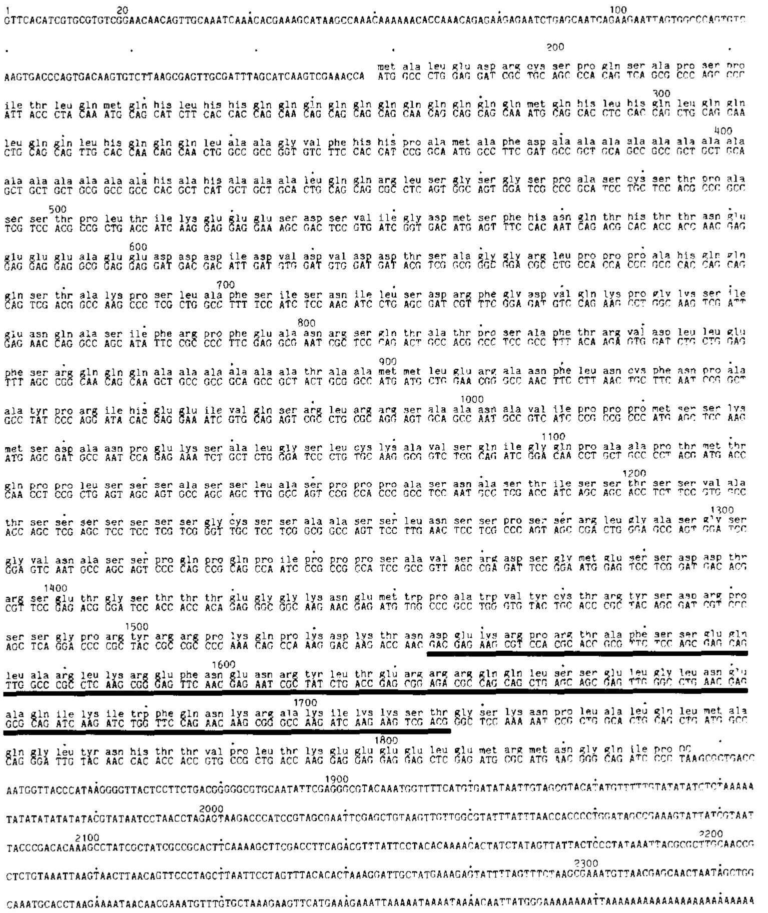

Figure 2. Sequence of the *engrailed* cDNA Clone, c-2.4

The direction of transcription (5′ to 3′) is from left to right. The sequences of both strands were obtained. The first 9 nucleotides of the cDNA do not correspond with the genomic sequences (see text). Translation of the 1700 bp open reading frame starting from the ATG at position 178 is also shown. In the regions upstream and downstream of this large open frame all three frames are closed. The region of the homeo box homology is underlined. The splice junctions of the two introns lie at positions 1489 and 1587.

seems unlikely for two reasons: no consensus splice acceptor site sequences were found in this region, and a second *engrailed* cDNA clone that was sequenced has a different, but also anomalous, extreme 5′ end. The two clones each diverge from the genomic DNA at different points. Similar anomalies have been noted in other cDNA clones isolated from the same library, and these may have resulted from the snapback priming in synthesis of the second strand (Laughton and Scott, 1984).

The *engrailed* cDNA Clone Contains a Homeo Box

The presence of a homeo box located towards the 3′ end of, and in-frame with, the large 1700 bp open reading frame suggests that at least a part of this open reading frame is probably translated. As described above (see Introduction), the homeo box is a potential coding region of ~180 bp that is strongly conserved among several genes responsible for generating the body pattern of *Drosophila*. At the amino acid level, the homeo box sequence is conserved in several vertebrates as well (Levine et al., 1984; Carrasco et al., 1984). The *engrailed* cDNA contains a copy of a homeo box whose sequence diverges from that of all other known homeo boxes: *Drosophila* genes *ftz* and *Antp* (two genes within the Antennapedia complex), and *Ubx* (a gene of the Drosophila Bithorax complex), two human homeo boxes (Hu1 and Hu2; Levine et al., 1984), and a *Xenopus* homeo box (AC1; Carrasco et al., 1984) (Figure 3). Although the *engrailed* homeo box is related to the others, it is also clear that the other *Drosophila* homeo boxes share more homology with the human sequences than with the *engrailed* one. For example, within the 60 amino acids comprising the homeo box, the *ftz* gene shares 50 identical amino acids with the human Hu1 homeo box but only 30 with the *engrailed*. When conservative amino acid substitutions are taken into account and *engrailed* is compared with the others, there are three regions of relatively strong homology separated by regions of weak homology. These strong regions are a short stretch at the N-terminal end of the homeo box and two longer stretches near the middle and the C-terminal end of the homeo box. The strongest region of homology lies within the C-terminal third of the *engrailed* homeo box, where 11 of 14 residues are identical among the *engrailed*, *ftz*, *Antp*, Hu1, and Hu2 homeo boxes. It is in this region that some homology has also been found with the yeast mating-type regulatory proteins a1 and $\alpha2$ (Laughton and Scott, 1984; Shepherd et al., 1984).

Another distinguishing feature of the *engrailed* homeo box is its interruption by an intervening sequence. One of the two intron splice junctions of the *engrailed* cDNA is within the *engrailed* homeo box (see upper line, Figure 4). Thus, the first 70 bp of the homeo box are found at the 3′ end of a 98 bp exon, and in the genomic DNA are separated from the rest of the homeo box sequences by a 282 bp intron. By contrast, all of the previously described homeo boxes are not interrupted.

```
                                       1                                   10
 Hu1    ser arg tyr asp gly pro asp gly lys arg ala arg thr ala tyr thr
 Hu2    thr gly ser ser phe gly pro thr ala gly gly arg gln thr tyr thr
 AC1    gly val gly tyr gly ser asp arg arg arg gly arg gln ile tyr ser
Antp    ile tyr leu glu pro thr glu arg lys arg gly arg gln thr tyr thr
 ftz    met leu thr asp cys lys asp ser lys arg thr arg gln thr tyr thr
 Ubx    ser thr gly thr asn gly leu arg arg arg gly arg gln thr tyr thr
  en    pro lys asp lys thr asn asp glu lys arg pro arg thr ala phe ser
  er    asp gly gly gly val pro glu asp lys arg pro arg thr ala phe ser

                                            20
 Hu1    arg tyr gln thr leu glu leu glu lys glu phe his phe asn arg tyr
 Hu2    arg tyr gln thr leu glu leu glu lys glu phe his tyr asn arg tyr
 AC1    arg tyr gln thr leu glu leu glu lys glu phe his phe asn arg tyr
Antp    arg tyr gln thr leu glu leu glu lys glu phe his phe asn arg tyr
 ftz    arg tyr gln thr leu glu leu glu lys glu phe his phe asn arg tyr
 Ubx    arg tyr gln thr leu glu leu glu lys glu phe his thr asn his tyr
  en    ser glu gln leu ala arg leu lys arg glu phe asn glu asn arg tyr
  er    gly thr gln leu ala arg leu lys his glu phe asn glu asn arg tyr

                    30                                      40
 Hu1    leu thr arg arg arg arg ile glu ile ala his ala leu cys leu ser
 Hu2    leu thr arg arg arg arg ile glu ile ala his ala leu cys leu thr
 AC1    leu thr arg arg arg arg ile glu ile ala asn ala leu cys leu thr
Antp    leu thr arg arg arg arg ile glu ile ala his ala leu cys leu thr
 ftz    ile thr arg arg arg arg ile asp ile ala asn ala leu ser leu ser
 Ubx    leu thr arg arg arg arg ile glu met ala tyr ala leu cys leu thr
  en    leu thr glu arg arg arg gln gln leu ser ser glu leu gly leu asn
  er    leu thr glu lys arg arg gln gln leu ser gly glu leu gly leu asn

                                      50
 Hu1    glu arg gln ile lys ile trp phe gln asn arg arg met lys trp lys
 Hu2    glu arg gln ile lys ile trp phe gln asn arg arg met lys trp lys
 AC1    glu arg gln ile lys ile trp phe gln asn arg arg met lys trp lys
Antp    glu arg gln ile lys ile trp phe gln asn arg arg met lys trp lys
 ftz    glu arg gln ile lys ile trp phe gln asn arg arg met lys ser lys
 Ubx    glu arg gln ile glu ile trp phe gln asn arg arg met lys leu lys
  en    glu ala gln ile lys ile trp phe gln asn lys arg ala lys ile lys
  er    glu ala gln ile lys ile trp phe gln asn lys arg ala lys leu lys

            60
 Hu1    lys asp asn lys phe lys ser met ser leu
 Hu2    lys glu ser lys leu leu ser ala ser gln
 AC1    lys glu arg
Antp    lys glu asn lys thr lys gly glu pro asp
 ftz    lys asp arg thr leu asp ser ser pro glu
 Ubx    lys glu ile gln ala ile asn his leu asn
  en    lys ser thr gly ser lys asn pro leu ala
  er    lys ser ser gly thr lys asn pro leu ala
```

Figure 3. Amino Acid Comparisons of the Homeo Boxes

The first amino acid of the homeo box is marked as 1, and the entire homeo box homology is underlined. Hu1 and Hu2 are two homeo boxes isolated from humans (Levine et al., 1984). AC1 is a Xenopus homeo box (Carrasco et al., 1984). Antp and ftz are from two loci of the Antennapedia complex of Drosophila (Antp is from McGinnis et al., 1984a; ftz is from McGinnis et al., 1984b), and Ubx is from the Bithorax complex (data from Levine et al., 1984). En and er are from the *engrailed* and *engrailed-related* genes (this paper).

A Nearby Related Gene Also Contains a Divergent Homeo Box

The 2.7 kb RNA from which the cDNA clone c-2.4 was derived is transcribed from within a region in which 11 *engrailed* mutations have been localized. When a fragment containing the 3′ half of c-2.4 was hybridized to a Southern blot of the various lambda phage containing the DNA surrounding the *engrailed* locus, hybridization was observed both to the *engrailed* region and to a region that is 17 kb downstream of that of *engrailed* transcription and 4 kb from the most proximal *engrailed* breakpoint mutation (Figure 1A). A partial nucleotide sequence for this region reveals a striking homology to the *engrailed* gene. A comparison of this sequence (hereafter called the *engrailed-related* gene) to the genomic *engrailed* sequences (Figure 4), shows that homology begins at the upstream edge of the *engrailed* homeo box and continues for about 50 bp, diverging at the *engrailed* splice junction. The sequence at this point in the *engrailed-related* sequences is AGGTA, conforming to typical eukaryotic splice junctions (Mount, 1982). The sequences within the 282 bp intron that splits the *engrailed* homeo box share no homology with the *engrailed-related* region. However, 417 bp beyond the point of divergence the homology with *engrailed* resumes. The homology begins at the *engrailed* splice acceptor junction and then continues through the *engrailed* homeo box and 75 bp into the 3′-terminal region of the *engrailed* open reading frame. Nine codons before the end of the *engrailed* open reading frame, the two sequences again diverge. No other homology, either upstream or downstream, between the *engrailed* region and this related region has been detected.

When the *engrailed* open reading frame is used as a guide, and the points of divergence of this related region are taken to be splice junctions, conservation of amino acids between the two is quite strong (Figure 5). The region of homology is within an open reading frame, and 78 of the 91 potential amino acid residues are identical. Of the rest, 8 are conservative substitutions so that 86 out of 91 amino acids are homologous between *engrailed* and the related open frame. The homology between these two related genes extends beyond the homeo box, whereas the homology among the *Ubx*, *ftz*, and *Antp* sequences is strictly confined to the homeo box region.

```
        <------- en Intron I---------
        *     ***        *  *      *    *  *  *     *             .*
en      tatttctcgtgtttttatttattccagGACCCCGCTACCGCCGCCCCAAACAGCCAAAG
er      TCCAGTTCGGCGGCAGGTGGTGGGGGTGGGGCGTCGAGAAGGGGGAGGCCGCCGATGGG

        * *  *      ** ** **  * ** ** ** ** ***  * *  **********  *
en      GACAAGACCAACGACGAGAAGCGTCCACGCACCGCGTTCTCCAGCGAGCAGTTGGCCCGC
er      GGCGGGGTGCCGGAGGACAAAAGGCCGCGAACGGCCTTCAGCGGAACGCAGTTGGCCAGA

              <--------------en Intron II------------------>
        ** ***** ****    * * **         ** *** ** * * ***** ******
en      CTTAAGgtgagattcagttcttttt....attccatttcttccacttacagCGGGAGTTC
er      CTGAAGgtaagatgacttccgttcg....cccccctttttttcgcccacagCACGAGTTC

        1600
        ******************** ***  * ***************** * **  **** ***
en      AACGAGAATCGCTATCTGACCGAGCGGAGACGCCAGCAGCTGAGCAGCGAGTTGGGCCTG
er      AACGAGAATCGCTATCTGACGGAGAAGCGACGCCAGCAGCTGAGCGGGGAACTGGGACTG

                                                1700
        *********************************** ********* * ** *******
en      AACGAGGCGCAGATCAAGATCTGGTTCCAGAACAAGCGGGCCAAGATCAAGAAGTCGACG
er      AACGAGGCGCAGATCAAGATCTGGTTCCAGAACAAACGGGCCAAGCTGAAAAAGTCGAGC

        *** **** *********** ************** ****************** * **
en      GGCTCCAAAAATCCGCTGGCACTGCAGCTGATGGCCCAGGGATTGTACAACCACACCACC
er      GGCACCAAGAATCCGCTGGCGCTGCAGCTGATGGCGCAGGGATTGTACAACCACTCGACG

                            1800
         * *********   **************  **  **    * * * *   *  **
en      GTGCCGCTGACCAAGGAGGAGGAGGAGCTCGAGATGCGCATGAACGGGCAGATCCCCTAA
er      ATACCGCTGACCCGCGAGGAGGAGGAGCTGCAGGAGCTGCAGGAGGCGGCTAGTGCCCGT

        ** *   **       * ** * .  *
en      GCGCTGACCAATGGTTACCCATAAGGG
er      GCCCGTGCCGCCAAGGAGCCCTGCTAG
```

Figure 4. Nucleotide Sequence Comparison of the *engrailed* and *engrailed-related* Regions

The genomic sequences of the *engrailed* gene are compared with the homologous sequences in the related gene 17 kb downstream (Figure 1). The numbering used is the same as in Figure 2; intron sequences are not numbered. Direction of transcription (5′ to 3′) is from left to right in both cases, although the two are transcribed from opposite strands (see Figure 1). Asterisks indicate nucleotides shared between the two genes; the homeo box homology is underlined. The *engrailed* intron sequences and the presumed *engrailed-related* intron are shown as lowercase letters. Homology begins at the upstream edge of the homeo box and continues almost to the end of the *engrailed* open reading frame. The internal intron sequences (282 nucleotides in *engrailed* and 417 nucleotides in the related gene) are not homologous except at the splice junctions and are not shown. In this region, the *engrailed* genomic and cDNA clone c-2.4 sequences are identical except for a single C to T change at position 1584; this is a silent change in this reading frame. en = *engrailed* genomic sequences; er = *engrailed-related* genomic sequences.

The *engrailed-related* gene is transcribed (see Figure 6). Because the *engrailed* locus and the *engrailed-related* sequences are oriented in opposite directions on the chromosome, this expression does not involve the use of an alternative splice site that joins different 3′ ends to the *engrailed* transcript. In order to monitor RNA expression, a genomic fragment that contains the majority of the sequences related to the *engrailed* locus (probe B in Figure 1C) was hybridized to a Northern blot of poly(A)$^+$ RNA isolated from embryos 3–12 hr after egg-laying. Because of the *engrailed* homology, this probe detected the same spectrum of transcripts as did a probe from a similar region of the *engrailed* gene (probe A, Figure 1B), but the intensities of the various bands differ. This spectrum in-

```
                                                              .
                                *
en  (..Intron I of engrailed..)GlyProArgTyrArgArgProLysGlnProLys
er  SerSerSerAlaAlaGlyGlyGlyGlyGlyGlyValGluLysGlyGluAlaAlaAspGly

                      .                   .                   .
                       *  *  *  *  *  *  *  *        *  *  *  *
en  AspLysThrAsnAspGluLysArgProArgThrAlaPheSerSerGluGlnLeuAlaArg
er  GlyGlyValProGluAspLysArgProArgThrAlaPheSerGlyThrGlnLeuAlaArg

     *  *                                                  *  *
en  LeuLys<..........Intron II of engrailed...........>ArgGluPhe
er  LeuLys<......Proposed splice in  related gene.....>HisGluPhe

    1600                   .                   .
     *  *  *  *  *  *  *  *     *  *  *  *  *  *     *  *  *  *
en  AsnGluAsnArgTyrLeuThrGluArgArgArgGlnGlnLeuSerSerGluLeuGlyLeu
er  AsnGluAsnArgTyrLeuThrGluLysArgArgGlnGlnLeuSerGlyGluLeuGlyLeu

                                            1700
       .                   .
     *  *  *  *  *  *  *  *  *  *  *  *  *  *  *     *  *  *
en  AsnGluAlaGlnIleLysIleTrpPheGlnAsnLysArgAlaLysIleLysLysSerThr
er  AsnGluAlaGlnIleLysIleTrpPheGlnAsnLysArgAlaLysLeuLysLysSerSer

       .                   .                   .
     *     *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *     *
en  GlySerLysAsnProLeuAlaLeuGlnLeuMetAlaGlnGlyLeuTyrAsnHisThrThr
er  GlyThrLysAsnProLeuAlaLeuGlnLeuMetAlaGlnGlyLeuTyrAsnHisSerThr

                        1800
       .
        *  *  *     *  *  *  *  *
en  ValProLeuThrLysGluGluGluGluLeuGluMetArgMetAsnGlyGlnIleProStop
er  IleProLeuThrArgGluGluGluGluLeuGlnGluLeuGlnGluAlaAlaSerAlaArg

       .                   .
en  (Past en stop codon)
er  AlaArgAlaAlaLysGluProCysStop
```

Figure 5. Amino Acid Comparison of the *engrailed* and *engrailed-related* Regions

Translation of the *engrailed* genomic sequences is shown, with the positions of the introns also indicated. Translation of *engrailed* is based on homology with the open reading frames of other homeo boxes. The translation of the *engrailed-related* gene is based on assumptions that it is translated in the same frame as *engrailed* and spliced at the same sites. The homeo box homology is underlined, and amino acid identities between the two genes are indicated by asterisks. Numbering is the same as in Figure 2. Homology starts at the upstream edge of the homeo box and continues through the splice junctions until 9 amino acids prior to the first *engrailed* stop codon. There is an in-frame stop codon in the *engrailed-related* gene shortly thereafter.

cludes the major 2.7 kb *engrailed* RNA and other minor RNAs (3.6, 3.4, 2.0, and 1.4 kb). By contrast, when an upstream restriction fragment from the *engrailed-related* region (probe C, Figure 1C) was used as a probe, the 3.4 kb and 2 kb RNAs were the major bands detected (Figure 6C). This probe contains a 60 bp stretch that has 70% homology with the 5′ end of the *engrailed* homeo box, and weak hybridization with the other RNAs was observed. It is possible, of course, that probes B and C are hybridizing to different RNAs that were not resolved by electrophoretic sizing. We consider it more likely that the *engrailed-related* sequences homologous with *engrailed*, as well as upstream sequences not related to *engrailed*, are both present on the same *engrailed-related* transcripts. If so, the transcript produced by this *engrailed-related* gene has 5′ sequences unrelated to *engrailed* but does contain a homeo box and 3′-terminal portion that are closely related. By analogy with the *engrailed* gene, we presume that an intron from within the homeo box is spliced from the *engrailed-related* gene. Confirmation of these points will require S1 nuclease mapping of the transcripts or isolation of *engrailed-related* cDNA clones.

Discussion

This initial molecular analysis of *engrailed* locus function has identified a cDNA sequence with several interesting features. This cDNA has a peptide coding capacity for an extremely unusual polypeptide, it contains a divergent homeo box, and it has a region that shares strong sequence homology with a neighboring gene.

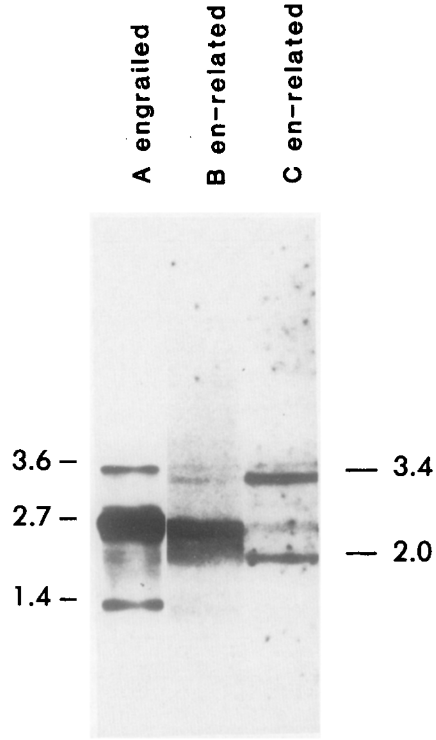

Figure 6. The *engrailed-related* Region is Transcribed

Poly(A)$^+$ RNA from 3–12 hr embryos was fractionated on a formaldehyde–agarose gel, transferred to nitrocellulose, and hybridized with probes from the *engrailed* or *engrailed-related* regions. See Figures 1B and 1C for a representation of the probes used. Lane A: probe A, containing the *engrailed* homeo box region. Lane B: probe B, containing the majority of the *engrailed-related* homeo box sequences. Lane C: probe C, containing for the most part *engrailed-related* sequences upstream of the region of homology, but also having a 60 bp stretch with 70% homology with the 5′ end of the *engrailed* homeo box.

Within the cDNA sequence, a 1700 bp open reading frame is present, which, if translated in its entirety, would produce a polypeptide with stretches of polyglutamine, polyserine, and polyalanine, and few regions with an amino acid composition similar to other known proteins. It is possible that one of the several internal ATG triplets serves as the actual initiation codon. In addition to the ATG at position 90, there are 22 other ATG triplets upstream from the homeo box; half of these are in frame with the homeo box and half are in the frame shifted by +1 bp. However, initiation of translation at most of these internal in-frame ATGs would still result in unusual polypeptides, since the unusual features extend almost to the homeo box. For example, initiation at any of the four in-frame ATGs prior to position 898 would result in a protein containing, among other features, a stretch of polyglutamic/aspartic acid (starting at position 580) and a stretch with the composition (glutamine)$_3$–(alanine)$_6$. Initiation at any of the next six in-frame ATGs would yield stretches extremely high in serine, such as those encoded at positions 1135–1152, 1183–1281, and 1408–1419. The only ATG that avoids these regions lies at position 1438, 31 codons upstream of the start of the homeo box. Of the 22 upstream ATG triplets, none stand out as having excellent homology with the initiation sites of other eukaryotic genes (consensus sequence CCPurCCATG(G); Kozak, 1984). Those having the most homology to this sequence lie at positions 367 (CGGCAATGG), 898 (CGGCCATGA), 901 (CCATGATGC), 1117 (CTACGATGA), and 1372 (CCGGAATGG). All of these are in-frame with the homeo box, and all would produce a protein with some unusual features.

The *engrailed* homeo box is in frame with the large open reading frame, implying that at least a portion of the unusual sequence is translated. However, this homeo box region of the cDNA is found in several *engrailed* transcripts of different sizes whose structures are at present unanalyzed, and it may be that the 2.7 kb transcript is not translated. It is interesting to note that on the opposite strand there is an $\sim$1200 nucleotide open reading frame in the region upstream of the homeo box, but this open frame may be simply a reflection of the repeating trinucleotide structure of the DNA in this region.

The homeo box of the *engrailed* gene is distinct from those within the Antennapedia and Bithorax complexes and in vertebrates. Despite regions of strong homology with the other homeo boxes, the less homologous regions predict amino acid substitutions that would result in charge differences. The *engrailed* homeo box is further set apart from the other homeo box class by the presence of an intron. Were the homeo box to represent a functional protein domain such as a DNA binding domain (Laughton and Scott, 1984; Shepherd et al., 1984), its operation might be modulated in the *engrailed* class by splicing of various upstream exons with partial homeo boxes to the same 3′-terminal homeo box region.

At present, we have identified two examples of the *engrailed* class of homeo boxes, and the two genes are in close proximity. The other identified Drosophila genes that contain homeo boxes all lie within either the Antennapedia complex or the Bithorax complex, both of which are clusters of genes important for proper interpretation of segment identities.

Nearly perfect transformation of parts of one segment into parts of another, the hallmark of the classical homeotic mutations in the Bithorax complex and the Antennapedia complex, does not occur in *engrailed* mutations. Although certain *engrailed* alleles may cause the duplication of some anterior compartment structures in homologous posterior compartments in some segments (Garcia-Bellido and Santamaria, 1972), nonhomologous morphogenetic abnormalities and cell lethality occur in other alleles (Kornberg, 1981a; Lawrence and Struhl, 1982). The unifying feature of the pleiotropic *engrailed* phenotype is the requirement for *engrailed* function in posterior compartment cells to maintain the separation between neighboring anterior and posterior compartments. Although the similarity between the function of the classical homeotic genes and *engrailed* is left unresolved by comparison of mutant phenotypes, the shared homeo box sequence domains suggest common functionality. The homeo box may therefore contribute to a developmental switching function that is more general than choosing between two fully formed structures as in the Antennapedia and Bithorax complexes.

Experimental Procedures

General Procedures

Most experimental methods used are described by Maniatis et al. (1982). Two cDNA libraries in λ gt-10 were screened with a purified 0.9 kb genomic Eco RI fragment (probe A in Figure 1B). One library was made from 1.5–5 hr embryos and kindly provided by M.Goldschmidt-Clermont and D. Hogness; a second, also in λ gt-10, was made from 3–12 hr RNA; second-strand synthesis for the latter library was primed by oligo(dC) hybridized to the dG-tailed first strand rather than by snapback. The protocol used in preparing this library follows very closely that described in detail by St. John et al. in an unpublished manuscript (personal communication). In brief: poly(A)$^+$ RNA isolated from Oregon R flies was primed with oligo(dT) (Collaborative Research), and first-strand cDNA was made with reverse transcriptase (Life Sciences, Inc.). RNA was digested with RNAase A, and the cDNA was separated from primer, deoxy- and ribonucleotides on Sepharose CL-2B. Tailing of the first strand with dG was accomplished using terminal transferase (Ratliffe Biochemicals). Oligo(dC) (Collaborative Research) was annealed to the tailed first strand, and the second strand was synthesized using the Klenow fragment of DNA polymerase I (Boehringer Mannheim). Internal Eco RI sites were protected with Eco RI methylase (Gift of P. Greene). The DNA was phenol-extracted, and Eco RI linkers (New England Biolabs) were ligated with T4 DNA ligase (International Biotechnologies). After cutting with Eco RI (New England Biolabs), the cDNA was phenol-extracted and sized on a Sepharose CL-2B column. Ligation to Eco RI-cut λgt10 DNA and in vitro packaging followed standard procedures (Maniatis et al., 1982). From 2 μg of RNA, $\sim$7 × 10^5 clones were recovered. About 50% of these clones have inserts larger than 750 bp. This library, as well as others prepared in an identical manner from later developmental stages (0–3 hr, 12–24 hr embryo, I and II instar larvae, early and late III instar larvae, 5.5–7.5 day pupae, 7–9 day pupae, adult males, and adult females), are available upon request. Both of the early embryo libraries yielded several clones $\sim$2.5–2.6 kb in length. The sequenced cDNA was isolated from the Goldschmidt-Clermont and Hogness library.

The cDNA clone was sequenced on both strands by the method of Sanger et al. (1977). Subclones for sequencing were obtained either by sonication (Deininger, 1983) or by subcloning defined restriction fragments into either M13mp8/9 (Messing and Vieira, 1982) or pEMBL 8/9 (Dente et al., 1983). Genomic sequences were obtained by subcloning DNA fragments from a Canton S-derived λ recombinant library (Maniatis et al., 1978) into M13.

Northern Blot Hybridization

Poly(A)$^+$ RNA was prepared from 3–12 hr Drosophila melanogaster embryos. Embryos were collected, dechorionated, and stored frozen at −80°C until use, and then powdered with a mortar and pestle in liquid N_2. A batch of 28 g of powdered embryos was stirred into 80 ml of a buffer mixture (5M guanidinium thiocyanate, 10 mM EDTA, 50 mM Hepes, pH 7.6, and 5% 2-mercaptoethanol), homogenized in a Polytron homogenizer (Brinkman Instruments) for 1 min, filtered through cheesecloth, and centrifuged for 10 min at 10,000 rpm. The supernatant was collected, Sarcosyl was added to 4%, and the mixture was layered onto 5 ml of 5.7M CsCl in three SW27 tubes and centrifuged for 20 hr in a Beckman SW27 rotor at 24,000 rpm, 20°C. The resulting clear pellets were rinsed with H_2O, combined, and dissolved in 14 ml of 4M urea, 50 mM Hepes, pH 7.6, 10 mM EDTA. The RNA mixture was extracted twice with phenol-$CHCl_3$ (1:1), once with $CHCl_3$, twice with ether, and precipitated with ethanol. The pellet was dissolved in H_2O, loaded as a batch onto oligo(dT) cellulose (Collaborative Research) and poly(A)$^+$ RNA was eluted as described by Maniatis et al. (1982).

RNA was separated on 1.0% agarose–formaldehyde gels and blotted onto nitrocellulose (Maniatis et al., 1982), except that the running buffer was 0.5M Hepes, pH 7.0, 10 mM EDTA, 50 mM Na-Acetate. Hybridization probes were prepared by nick translation of purified restriction fragments. Hybridizations were carried out at 42°C in 50% formamide, 5× SSC, 10 mM Na-phosphate, 5 mM EDTA, 5× Denhardts, 100 μg/ml carrier DNA.

Acknowledgments

We thank Hugo Martinez for aid in starting the sequence analysis. We thank Gail Martin and Harold Varmus for helpful comments on the manuscript. This work was supported by a postdoctoral fellowship (to S. P.), a training grant (to B. D.), and a grant to T. K. from the National Institutes of Health, and by a fellowship from the Weingart Foundation to L. M. K.

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 October 16, 1984

References

Carrasco, A. E., McGinnis, W., Gehring, W. J., and De Robertis, E. M. (1984). Cloning of an X. laevis gene expressed during early embryogenesis coding for a peptide region homologous to Drosophila homeotic genes. Cell *37*, 409–414.

Deininger, P. (1983). Random subcloning of sonicated DNA: application to shotgun DNA sequence analysis. Anal. Biochem. *129*, 216–223.

Dente, L., Cesareni, G., and Cortese, R. (1983). pEMBL: a new family of single stranded plasmids. Nucl. Acids Res. *11*, 1645–1655.

Garcia-Bellido, A., and Santamaria, P. (1972). Developmental analysis of the wing disc in the mutant engrailed of Drosophila melanogaster. Genetics *72*, 87–104.

Garcia-Bellido, A., Ripoll, P., and Morata, G. (1973). Developmental compartmentalization of the wing disk of Drosophila. Nature New. Biol. *245*, 251–253.

Kaufman, T. C., Lewis, R., and Wakimoto, B. (1980). Cytogenetic analysis of chromosome 3 in Drosophila melanogaster: the homeotic gene complex in polytene chromosome interval 84A-B. Genetics *94*, 115–133.

Kornberg, T. (1981a). *engrailed*: a gene controlling compartment and segment formation in Drosophila. Proc. Natl. Acad. Sci. USA *78*, 1095–1099.

Kornberg, T. (1981b). Compartments in the abdomen of Drosophila and the role of the *engrailed* locus. Dev. Biol. *86*, 363–372.

Kozak, M. (1984). Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. Nucl. Acids Res. *12*, 857–872.

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

Lawrence, P. A., and Morata, G. (1976). Compartments in the wing of Drosophila: a study of the engrailed gene. Dev. Biol. *50*, 321–337.

Lawrence, P. A., and Struhl, G. (1982). Further studies of the *engrailed* phenotype in Drosophila. EMBO J. *1*, 827–833.

Levine, M., Rubin, G., and Tjian, R. (1984). Human DNA sequences homologous to a protein coding region conserved between homeotic genes of Drosophila. Cell *38*, 667–673.

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

Maniatis, T., Hardison, R., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G., and Efstratiadis, A. (1978). The isolation of structural genes from libraries of eucaryotic DNA. Cell *15*, 687–701.

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

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

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

Messing, J., and Vieira, J. (1982). A new pair of M13 vectors for selecting either DNA strand of double-digest restriction fragments. Gene *19*, 269–276.

Mount, S. (1982). A catalogue of splice junction sequences. Nucl. Acids Res. *10*, 459–471.

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

Shepherd, J. C. W., McGinnis, W., Carrasco, A. E., De Robertis, E. M., and Gehring, W. J. (1984). Fly and frog homoeo domains show homologies with yeast mating type regulatory proteins. Nature *310*, 70–71.

Note Added in Proof

We have now isolated an *engrailed*-related cDNA clone whose structure confirms that the *engrailed*-related homeo box contains an intron that is spliced as proposed in Figure 4. The *engrailed* homeo box region has also been isolated by virtue of its homology with the *Ubx* homeo box (W. Gehring, personal communication). We thank W. Gehring for sharing this information prior to its publication.