Volume 279, number 2, 300–302 FEBS 09430 February 1991

ADONIS 001457939100138W

Cloning and sequencing of a leech homolog to the *Drosophila* engrailed gene

Cathy J. Wedeen, David J. Price* and David A. Weisblat

Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA

Received 26 November 1990

We have cloned and sequenced a homolog (ht-en) to the *Drosphila* engrailed (*en*) gene from the glossiphoniid leech, *Helobdella triserialis*. Amino acid comparisons of the ht-en homeodomain and C-terminal residues with the corresponding residues encoded by *en*-class genes of other species reveal 75–79% sequence identity. In addition, the ht-en sequence appears to have a serine-rich region 16 residues C-terminal from the homeodomain, which by analogy to *Drosophila* may be a target site for phosphorylation. The leech gene encodes some amino acid substitutions for residues that are highly conserved in other species. These are found within the second and third of the three putative helices of the homeodomain, and in both of the intervening turn regions.

Engrailed; Homeobox gene; *Helobdella triserialis*

1. INTRODUCTION

The homeobox gene, engrailed (*en*), encodes a DNA-binding protein that is necessary to establish the 'identity' of the posterior compartment within each segment in *Drosophila* [1–3]. The *en* gene encodes a serine-rich protein that has been shown to be the target of serine phosphorylation [4]; it has been proposed that other segmentation genes (e.g. fused) may regulate *en* function by phosphorylation [5]. A closely related gene in *Drosophila* is invected (*inv*), for which no function has yet been determined. Both *en* and *inv* are transcribed concurrently in the same tissues during embryogenesis [6]. In addition, *en* is expressed later in development in certain neurons of the central and peripheral nervous systems [7–10].

En-class genes of divergent species are defined as a subfamily of homeobox-containing genes having an especially distinct and highly conserved homeobox region. This high degree of conservation has led to the identification and cloning of homologs from divergent species. In the fruit fly, honeybee, mouse, chicken, zebrafish, and human, two copies of *en*-class genes have been identified; in other species (grasshopper and sea urchin) only one *en*-class gene has been found [10–17]. Thus, it may be that a single *en* gene was present in a common ancestor to the arthropods, echinoderms and chordates and that this gene was duplicated independently in two, and maybe more, separate lines (i.e. the chordates and the insects).

We have previously reported an *en*-class gene in the leech, *Helobdella triserialis* [18]. We have now cloned and sequenced the homeobox and 3′ nucleotides of this gene (ht-en) and we compare this sequence with those of other *en*-class genes.

Correspondence address: C.J. Wedeen, Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA

**Present address:* Department of Physiology, University Medical School, Teviot Place, Edinburgh, EH8 9AG, UK

2. MATERIALS AND METHODS

2.1. *Library screening*

The phage, λHt-en1, was one of 10 recombinants obtained by screening 6.8×10^4 plaque forming units from a *Helobdella triserialis* genomic library [19] using the low stringency hybridization conditions described by McGinnis et al. [20]. The probe used was a 250 bp *Pvu*II-*Sal*I en cDNA fragment containing the homeobox and upstream region from the clone en-HB 1 [3].

2.2. *DNA sequencing*

Both strands of the 500 bp *Pvu*II fragment (Fig. 1) were sequenced. Most of the sequence reported in Fig. 2 (i.e. the 3′ 147 bp of the homeobox and the downstream region preceding the first termination codon) is a subset of these data. Homeobox sequence 5′ to the *Pvu*II site was obtained from a subclone of the 3.5 kb *Hpa*I fragment, using oligonucleotide primers designed to anneal to already sequenced portions of the clone. All sequencing was done using the dideoxy chain termination method.

3. RESULTS AND DISCUSSION

3.1. *The ht-en sequence is highly conserved*

A recombinant clone homologous to *Drosophila en* was obtained by low stringency hybridization to a *Helobdella triserialis* library (Fig. 1, and section 2). The nucleotide and deduced amino acid sequence of the ht-en homeobox and C-terminal flanking region are given in Fig. 2. Given the probe used to clone λHt-en1 contained the *Drosophila en* homeobox and 5′ sequences,

 Published by Elsevier Science Publishers B.V.

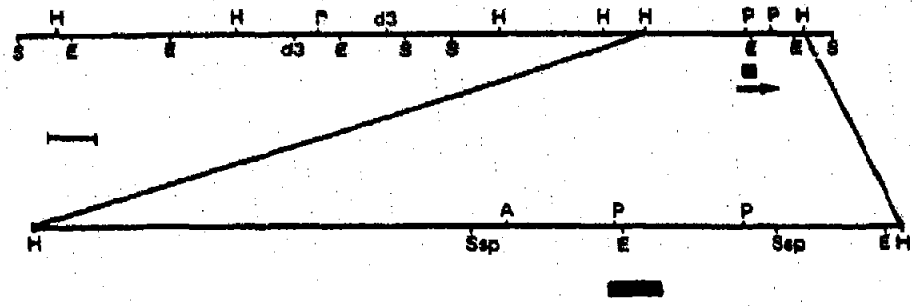

Fig. 1. Restriction map of genomic clone λHt-en1. The upper line shows the map of a 17 kb fragment. The *Sall* sites at the ends of the clone are from the polylinker of EMBL3 [21]. A blow-up of the 3.5 kb *Hpa*I fragment containing the homeobox is shown on the lower line. The position of the homeobox is shown by the filled boxes below each line. The arrow below the upper line designates the putative direction of transcription. The scale bar is equivalent to 1 kb for the upper line and 200 bp for the lower line. Key: A, *Apa*LI; d3, *Hin*dIII; E, *Eco*RI; H, *Hpa*I; P, *Pvu*II; S, *Sal*I; Ssp, *Ssp*I.

it was expected that the homeodomain portion of the cloned leech gene should be homologous to *en*. We do indeed observe this expected homology, but in addition, there is extensive homology extending 19 residues C-terminal to the homeobox, in a region not represented by the probe (Fig. 3). By these criteria we designate ht-en as an *en* homolog. The inferred amino acid sequence of the entire conserved region of ht-en was compared to the corresponding region of the other *en*-class genes from the species listed in Fig. 3. The ht-en amino acid sequence is 75–79% identical to the other *en* homologs. Given the short sequence length, the high degree and close range of sequence identity, we are unable to make any significant correlations between the degree of sequence identity and the time since evolutionary divergence of the leech from any of these species.

```
1                                                                   48
CAG GAC GAA AAG AGA CCT CGA ACA GCA TTC ACG GGC GAT CAG CTG GCG
GLN ASP GLU LYS ARG PRO ARG THR ALA PHE THR GLY ASP GLN LEU ALA
1                                   10
49                                                                  96
AGG TTG AAG CGT GAA TTC AGC GAG AAC AAA TAC CTG ACG GAG CAG AGG
ARG LEU LYS ARG GLU PHE SER GLU ASN LYS TYR LEU THR GLU GLN ARG
          20                                  30
97                                                                 144
AGA ACA TGT CTG GCG AAG GAA CTG AAC TTG AAC GAG AGC CAG ATC AAA
ARG THR CYS LEU ALA LYS GLU LEU ASN LEU ASN GLU SER GLN ILE LYS
                              40
145                                                                192
ATC TGG TTC CAG AAC AAG AGG GCC AAG ATG AAC AAG GCG AGT GGC GTG
ILE TRP PHE GLN ASN LYS ARG ALA LYS MET ASN LYS ALA SER GLY VAL
    50                                  60
193                                                                240
AAG AAT CAG TTG GCT CTG CAA CTC ATG GCA CAG GGC CTC TAC AAC CAC
LYS ASN GLN LEU ALA LEU GLN LEU MET ALA GLN GLY LEU TYR ASN HIS
                         70                                        80
241                                                                288
TCA TCA TCA TCA TCT TCT TCT TCC TCC TCC TCC TCT TCG ATC TTC CTC
SER SER SER SER SER SER SER SER SER SER SER SER ILE PHE LEU
                                    90
    290
CTC GCA TAA
LEU ALA
      98
```

Fig. 2. Nucleotide and deduced amino acid sequence of the ht-en homeobox and 3′ flanking region. The 294 nucleotide sequence of the putative open reading frame containing the homeobox and 3′ sequences and ending in a stop codon, TAA, is shown on the upper line; the homeo box (nucleotides 3–186) is underlined. The first and last nucleotides of each line are numbered above the line. The corresponding amino acid sequence of the putative open reading frame is given on the lower line; every 10 amino acids are numbered below the line.

3.2. *ht-en encodes amino acid substitutions in the homeodomain*

X-Ray diffraction has been used to determine the structure of an *en* homeodomain/DNA complex [23]. The proposed structure of the *en* homeodomain is similar to that proposed for the *Antennapedia* homeodomain on the basis of nuclear magnetic resonance [24]. The *en* homeodomain contains 3 α-helices and an N-terminal arm. Helices 1 and 2 pack against each other in an antiparallel arrangement and make few contacts with the DNA; helix 3 lies perpendicular to helices 1 and 2 and, as the 'recognition helix', makes extensive contacts with the major groove of the DNA. The residues composing each of the helices are designated in Fig. 3. In the ht-en homeodomain several amino acid changes are observed. Some of these amino acid differences have been reported earlier in a discussion of the epitope for a monoclonal antibody, mab4D9, directed against a portion of the invected homeodomain [10]. Here we describe the substitutions in the ht-en protein with respect to the proposed homeodomain structure. One change occurs at residue 58 within the 'recognition helix', number 3. This residue is isoleucine in every *en*-class protein except *Drosophila inv*, where it is leucine, and in the ht-en homeodomain, where it is a methionine. The other changes occur in helix 2 and in the turn regions between helices 1 and 2, and between helices 2 and 3. Substitutions within helix 2 occur at residues 34 and 35. One or both of these is always glutamine except in the sea urchin, where they are arginine and serine, and in leech, where they are threonine and cystine. In the turn regions, residue 26 is always arginine except in the sea urchin, where it is asparagine, and in leech, where it is lysine; residue 41 is often glycine in *en*-class genes but in sea urchin and in mouse this residue is replaced by the more sterically restricting threonine and serine, respectively, and in leech an asparagine, a nonconservative substitution, is found in this position. None of these amino acid substitutions occurs in a position that has

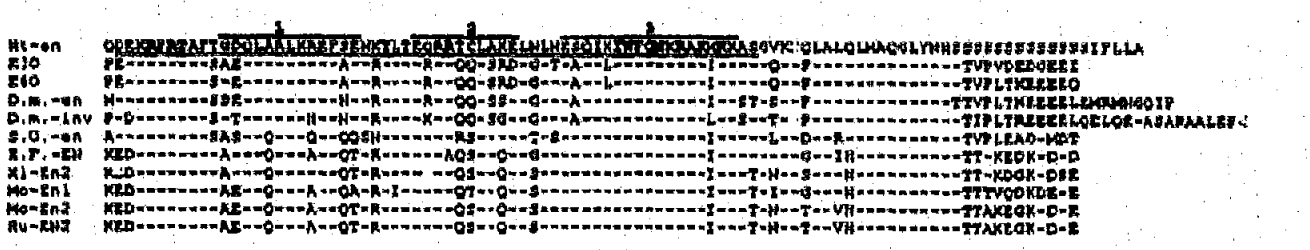

Fig. 3. Amino acid comparison of conceptual translation products from some *en*-class genes. The 98 amino acids translated from the ht-en nucleotide sequence are designated by the one letter code and shown on the top printed line. The sequence begins with the one amino acid N-terminal to the homeodomain and ends with the last amino acid before the first termination codon. The homeodomain is underlined. Residues putatively participating in α-helixes 1-3 are designated by bold lines above the ht-en sequence. Residues thought to make contact with DNA [22] are designated in bold type. The sequence is aligned with sequences of *en*-class genes from honeybee (E30, E60), *Drosophila* (en, inv), sea urchin (S.U.-en), zebrafish (ZF-EN), frog (X1-En2), mouse (Mo-En1, Mo-En2), and human (Hu-EN2) [9-16,22]. Dashes represent amino acids identical to those of ht-en.

been observed to make direct DNA or protein contact. Therefore, despite these changes in the sequence, the DNA target and overall structure of ht-en are likely to be very similar to that determined for *en*.

3.3. *Intron position in the homeobox is not conserved between fruit fly and leech*

Introns have been found at nucleotide 60 (Fig. 2) in the *en* and *inv* genes of *Drosophila* but are not present in the E30 and E60 genes of the honeybee. Intron sites have also been identified at a position 39 nucleotides 5′ to the homeobox in En-1 and En-2 of the mouse and a putative intron has been identified at this position in the zebrafish gene, ZF-EN. Within the portion of the ht-en gene that we have sequenced, we find no indication of any introns. We identify a putative open reading frame encompassing the homeobox and extending 99 nucleotides 3′ to the homeobox which aligns by homology with the nucleotide sequences of other *en* homologs. However, because our DNA sequence does not extend 5′ to the homeobox, we do not know whether ht-en contains any intron sites upstream from the homeobox.

3.4. *The ht-en gene encodes a potential phosphorylation site*

In addition to the other structural properties of the *en*-class genes, the *en* gene has been shown to encode several serine-rich stretches in regions of the gene 5′ to the homeobox. It has been demonstrated that *Drosophila* en is the target of a serine-threonine protein kinase [4]. Our deduced amino acid sequence for ht-en reveals a stretch of 13 serine repeats near the putative carboxy-terminal. By analogy to *Drosophila*, we speculate that these might serve as a phosphorylation site for the regulation of ht-en protein function.

Acknowledgements: This work was supported by Damon Runyon Cancer Fund Fellowship no. DRG925 to C.J.W., a traveling fellowship from the Medical Research Council (UK) to D.J.P. and NIH no. R29 HD23328 to D.A.W. We thank Ali Hemmati-Brivanlou and Richard Harland for sharing their sequence data prior to publication and Rich Kostriken for helpful comments and suggestions during the course of this work and on the manuscript.

REFERENCES

[1] Lawrence, P.A. and Morata, G. (1976) Dev. Biol. 50, 321-337.
[2] Kornberg, T. (1981) Proc. Natl. Acad. Sci. USA 78, 1095-1099.
[3] Desplans, C., Theis, J. and O'Farrell, P.H. (1985) Nature 318, 630-635.
[4] Gay, N.J., Poole, S.J. and Kornberg, T. (1988) NAR 16, 6637-6647.
[5] Preat, T., Therond, P., Lamour-Isnard, C., Limbourg Bouchon, B., Tricoire, H., Erk, I., Mariol, M-C. and Busson, D. (1990) Nature 347, 87-89.
[6] Coleman, K.G., Poole, S.J., Weir, M.P., Soeller, W.C. and Kornberg, T. (1987) Genes Dev. 1, 19-28.
[7] DiNardo, S., Kuner, J.M., Theis, J. and O'Farrell, P.H. (1985) Cell 43, 59-69.
[8] Brower, D.L. (1986) EMBO J. 10, 2649-2656.
[9] Doe, C. and Scott, M.P. (1989) Trends Neurosci. 11, 101-106.
[10] Patel, N., Martin-Blanco, E., Coleman, K.G., Poole, S.J., Ellis, M.C., Kornberg, T.B. and Goodman, C.S. (1989) Cell 58, 955-968.
[11] Dolecki, G.J. and Humphries, T. (1988) Gene 64, 21-31.
[12] Joyner, A. and Martin, G. (1987) Genes Dev. 1, 29-38.
[13] Fjose, A., Eiken, H.G., Njolstad, P.R., Molven, A. and Hordvik, I. (1988) FEBS Lett. 231, 355-360.
[14] Darnell, D.K., Kornberg, T. and Ordahl, C.P. (1986) J. Cell Biol. 103, 311a.
[15] Walldorf, U., Fleig, R. and Gehring, W.J. (1989) Proc. Natl. Acad. Sci. USA 86, 9971-9975.
[16] Poole, S.J., Law, M.L., Kao, F. and Lau, Y. (1989) Genomics 4, 225-231.
[17] Logan, C., Willard, H.F., Rommens, J.M. and Joyner, A.L. (1989) Genomics 4, 206-209.
[18] Weisblat, D.A., Price, D.J. and Wedeen, C.J. (1988) Development 104 Suppl., 161-168.
[19] Wedeen, C.J., Price, D.J. and Weisblat, D.A.. (1990) in: The Cellular and Molecular Biology of Pattern Formation (Stocum, D. and Karr, T. eds) pp. 145-167, Oxford University Press, New York.
[20] McGinnis, W., Levine, M.S., Hafen, E., Kuroiwa, A. and Gehring, W.J. (1984) Nature 308, 428-433.
[21] Frischauf, A.H., Lehrach, A., Poutska, A. and Murray, N. (1983) J. Mol. Biol. 170, 827-842.
[22] Hemmati-Brivanlou, A., de la Torre, J.R., Holt, C. and Harland, R.M. (1990) Development (in press).
[23] Kissinger, C.R., Liu, B., Martin-Blanco, E., Kornberg, T.B. and Pabo, C.O. (1990) Cell 63, 579-590.
[24] Quian, Y.Q., Billeter, M., Otting, G., Muller, M., Gehring, W.J. and Wutrich, K. (1989) Cell 59, 573-580.