Cloning and sequencing of a leech homolog to the *Drosophila* en*grailed* 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 *Drosophila* en*grailed* (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 genes, en*grailed* (en), encode 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 inv*ersed* (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.

2. MATERIALS AND METHODS

2.1. Library screening

The plasmid, XHt-en1, was one of 10 recombinants obtained by screening 6.8 x 10⁶ 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 *PvuII* fragment containing the homeobox and upstream region from the clone en-HB1 [3].

2.2. DNA sequencing

Both strands of the 500 bp *PvuII* 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 *PvuII* site was obtained from a subclone of the 3.5 kb *HpaI* 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,
Fig. 1. Restriction map of genomic clone Ht-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 Hpal fragment containing the homebox is shown on the lower line. The position of the homebox 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, Apal; d3, HindIII; E, EcoRI; H, Hpal; P, PvuII; S, Sall; Sp, Ssp.

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 homebox, 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-class homologs.

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 invaded 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 the corresponding residue is leucine.

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 invaded 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 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 cysteine. In the turn regions, residue 26 is always glutamine except in the sea urchin, where it is asparagine, and in leech, where it is cysteine. In the turn regions, residue 26 is always glutamine except in the sea urchin, where it is asparagine, and in leech, where it is cysteine.
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 homedgene 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. R39 HD23328 to D.A.W. We thank Ali Hemmati-Brivanlou and Richard Harland for sharing their sequence data prior to publication and Rich Kostrikis for helpful comments and suggestions during the course of this work and on the manuscript.

REFERENCES