Arch. Biol. Med. Exp. 21: 443-447 (1988)
Printed in Chile

Sequence conservations in vertebrate homeo-box mRNAs

Conservación en las secuencias homeo de mRNA de vertebrados

E. M. De ROBERTIS, T. R. BURGLIN, A. FRITZ, G. OLIVER, K. CHO and C. V. E. WRIGHT

Department of Bioligical Chemistry, University of California
Los Angeles, CA 90024-1737
USA

Since the initial discovery of the homeobox in *Drosophila* homeotic genes (1, 2) and its subsequent finding in *Xenopus* (3), this 180-nucleotide region has been found in genes from a wide range of organisms (reviewed in 4). Recently the deduced sequences of several complete vertebrate homeodomain-containing proteins have been published. We have searched for regions conserved outside of the homeobox in cDNA sequences from frogs (5-7), mice (8-10), and humans (11, 12), and now report three observations. First, the amino-terminal 20 to 45 amino acids of many different vertebrate and insect genes are conserved. Second, some vertebrate homeodomain proteins are similar throughout the length of the whole protein, defining families of highly conserved proteins that presumably arose by gene duplication. Finally, we find that the 5' non-coding sequences of several homeobox genes are highly conserved within a 100-nucleotide stretch upstream of the initiator AUG. These unexpected similarities in the 5' mRNA leaders are sometimes better conserved than the protein-coding homeobox regions, and may be pointing to a translational control mechanism common to many homeobox genes.

Initially we compared the deduced amino acid sequence of our *Xenopus* X1Hbox 2 (7) gene to that of the human cDNA HHO.cl recently reported by Boncinelli's group (12). As shown in the top two lines of Fig. 1, the genes share extensive conservations throughout the whole coding region, to such an extent that we believe that they may represent the frog and human homologues of the same gene. Starting with this alignment we compared the other available sequences, all of which are genes different from each other, using the methods described in Fig. 1. The vertebrate genes, all of which contain an *Antennapedia*-type homeobox, fall into three families that we call A, B and C (Fig. 1). Amino acids that are shared between members of several families are indicated by diamonds in Fig. 1. Similarities to the *Drosophila melanogaster* genes *Antennapedia* (13), *Deformed* (14), *Ultrabithorax*, *caudal*, and *fushi tarazu* are also indicated.

The vertebrate genes have a conserved amino terminus, and similarities extend in some cases for the first 45 amino acids. Phenylalanines (F) and Tyrosines (Y) seem to appear periodically in the amino terminus. Less extensive amino-terminal similarities (mainly MSS-YF-N-) had been noted previously in other studies (7, 9, 11, 14). A second region present in most homeodomain-containing proteins is the conserved IYPWM pentapeptide, which is found at variable distances (10 to 22 amino acids) in front of the homeobox (7, 12, 18), and always located in the exon preceding the homeobox. The *Drosophila* homeobox genes *caudal* and *fushi tarazu* lack the amino-terminal conservation but have a divergent form of the IYPWM pentapeptide (Fig. 1).

In the regions between the amino terminal conservation and the IYPWM pentapeptide, as well as downstream of the homeodomain, many conserved stretches can be found within each of the three families shown in Fig. 1. These similarities are outlined by boxes in Fig. 1 and, unless also indicated by diamond signs, they arc not shared by other homeodomain protein families. The gene nomenclature used in Fig. 1, while confusing, is that of the various laboratories that isolated them. Only those

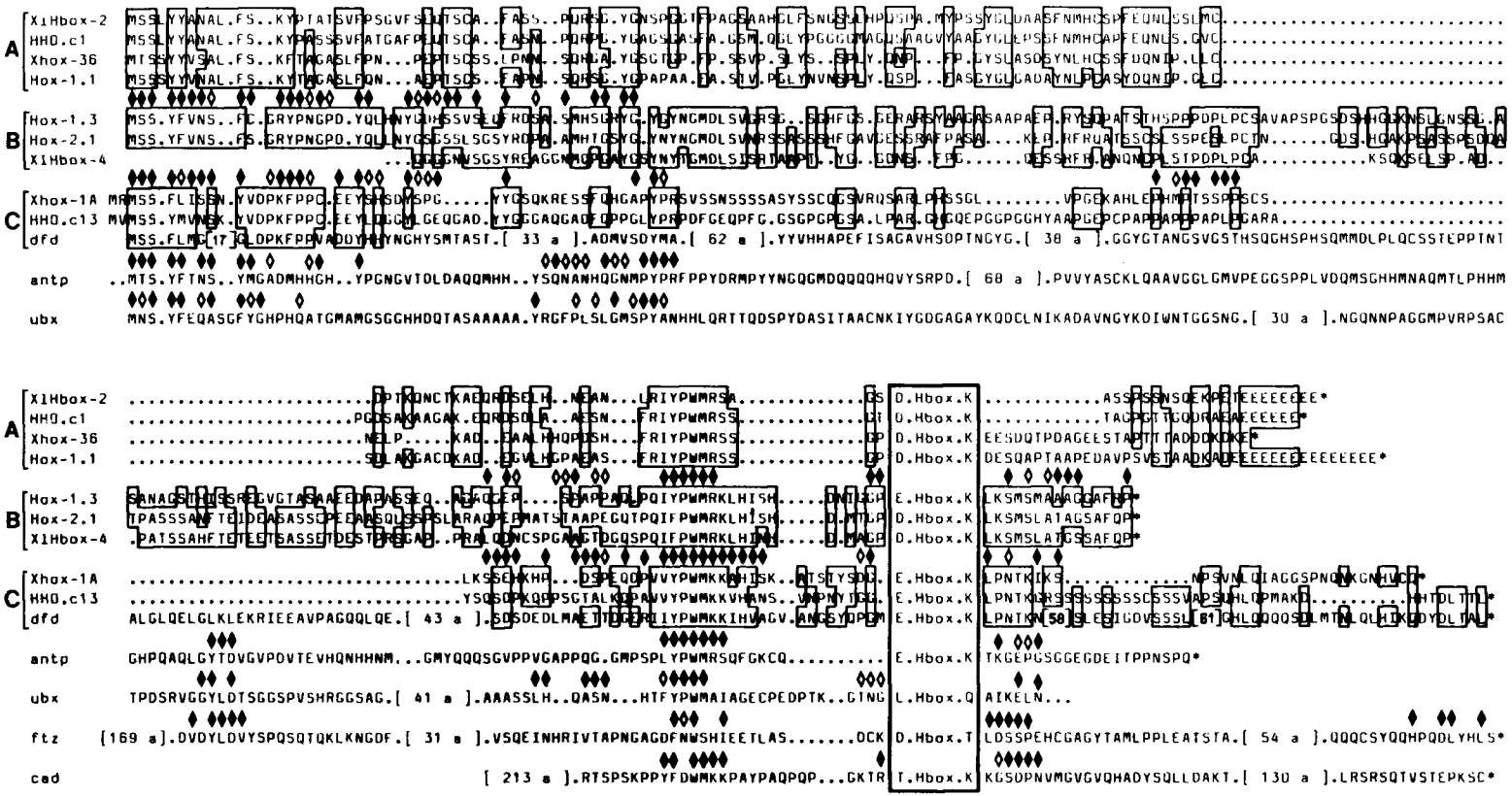

Fig. 1: **Regions of protein homology in 14 homeobox-containing genes.** Alignment of the deduced proteins for the *Xenopus* genes X1Hbox-2 (7), Xhox-36 (6), H1Hbox-4 (18), Xhox-1A (5), the mouse genes Hox-1.1 (10), Hox-1.3 (9), Hox-2.1 (8), the human genes HHO.cl (12), HHO.c13 (11), and the *Drosophila melanogaster* genes *Deformed* (dfd,14), *Antennapedia* (antp,13), *Ultrabithorax* (ubx, 15,1), *fushi tarazu* (ftz,17), and *caudal* (cad,16). Closely related genes are grouped into families A, B, and C. Boxes indicate a majority of identical or similar amino acids within a family. The following amino acids are defined as similar as adapted from the Mutational Data Matrix (20) : (K,R), (E,D), (Y,F), (I,L,V,M), (S,T), (H,Q). Solid diamonds indicate identities or similarities (as defined above) to one or several of the other families displayed above it. Open diamonds represent lesser similarites (20,22) as follows: (S,A,N,G,P,T), (N,Q,H), (F,L,I). Alignment in the central variable region may not be definitive, and no diamonds were placed in regions of no obvious homology. The homeobox amino acids are not displayed and the homeobox region is boxed as a whole. Numbers in brackets indicate the number of amino acids that were taken out in that position to avoid large gaps in the shorter vertebrate proteins. These deletions in *Drosophila* genes were made, if possible, in places which contained homopolymeric peptide stretches. In the case of *Antennapedia* the first methionine shown here is the eleventh amino acid of the deduced amino acid sequence (13). For all genes the gap between the homeobox and the IYPWM pentapeptide has been introduced at the intron position.

Methods: Nucleotide sequences were entered and translated using the SEQGEL program (21). Sequences were individually aligned using the ALIGN program (generously provided by the National Biomedical Research Foundation). The sequences were then further aligned by eye using the LINEUP program (22).

mouse genes that have been mapped to specific chromosomal loci (Hox gene complexes) have definitive names (19).

Family A comprises 4 genes, two of which (X1Hbox 2 and HHO.cl) are probably homologues. We have previously shown that X1Hbox 2 has similarities to the yeast gene MAT *al* in the region upstream of the homeobox (7), and these regions are also conserved in the other members of the A family (not shown). The recent finding that X1Hbox 2 has alternatively processed transcripts lacking a homeobox (7) suggests that these conserved upstream regions may have a function even in the absence of the homeodomain. The proteins from family A have a highly charged carboxy-terminus, usually polyglutamic acid (16 consecutive Glu residues in mouse Hox 1.1, eight in frog X1Hbox 2 and six in human HHO.cl). Family B consists of three very related proteins; mouse Hox 2.1 and Hox 1.3 have identical homeobox regions (8, 9), while X1Hbox 4 has a single change (18). Family C consists of three members: a *Xenopus* gene (5), a human gene (11), and the *Drosophila* gene *Deformed* (14). The similarity between the four members of family A was previously unreported, but previous workers had noted a close relationship between genes of families B (8, 9) and C (14), that is now extended in the more detailed alignments of the entire proteins presented in Fig. 1.

Within any of the families, the homeodomains are so similar that it would not be possible to say from a homeobox sequence comparison whether or not one deals with a previously reported gene. The best example is provided by Hox 2.1 and Hox 1.3, which have identical homeoboxes and many similarities in the immediately flanking regions, despite being two distinct genes that map on different mouse chromosomes (9). The degree of similarity between genes of a particular family suggests that vertebrates may have evolved duplicated copies of homeobox genes. Such duplications of genes, containing conservations throughout the entire length of the protein, have not been detected in *Drosophila* proteins of the *Antennapedia*-type (4). The vertebrate gene families might be considered comparable to the situation existing in the *Drosophila* gene network consisting of the gene *paired* and two *gooseberry* genes, which have extensively homologous amino-terminal domains joined to a *paired*-type homeobox (20).

Examination of the various vertebrate sequences shown in Fig. 1 suggests a general structure for *Antennapedia*-type homeodomain proteins. A conserved amino-terminus (of up to 45 amino acids) is followed by a long variable region (which may nevertheless have substancial conservations within a family of duplicated genes). This is followed by the IYPWM pentapeptide, which is separated from the homeodomain by a very short (10-22 amino acids) variable region, and finally a variable carboxy-terminus.

We also compared the cDNA sequences at the nucleotide level. Much to our surprise, we found that there are extensive nucleotide conservations in the 5' *non-coding* regions immediately adjacent to the initiator AUG codon. As shown in Fig. 2, the 5' non-coding sequence conservation is maximal among members of family A: for X1Hbox 2 and HH0c1 it is 91% (over 99 nucleotides), for Xhox-36 and Hox 1.1 it is 77% (over 102 nucleotides), and for Hox 1.1 and X1Hbox 2 it is 77% (over 100 nucleotides). In family B the conservation is somewhat less: for Hox 2.1 and Hox 1.3 it is 63% (over 114 nucleotides). Furthermore, the *Drosophila* gene *Deformed* (14) has a short block of conservation close to the AUG that is also present in vertebrate mRNAs.

The conservations shown in Fig. 2 are unlikely to be due to a translational reading frame in this 5' leader region, because in all sequences there are frequent stop codons in all reading frames, upstream AUGs preceding an open reading frame are lacking, and frequent insertions that would produce frameshifts are present.

Why are non-coding nucleotides conserved in many genes from such diverse organisms? It could be argued that it merely represents an evolutionary relic due to the common origin of homeobox genes. On the other hand, the sequences may remain invariant because they serve a func-

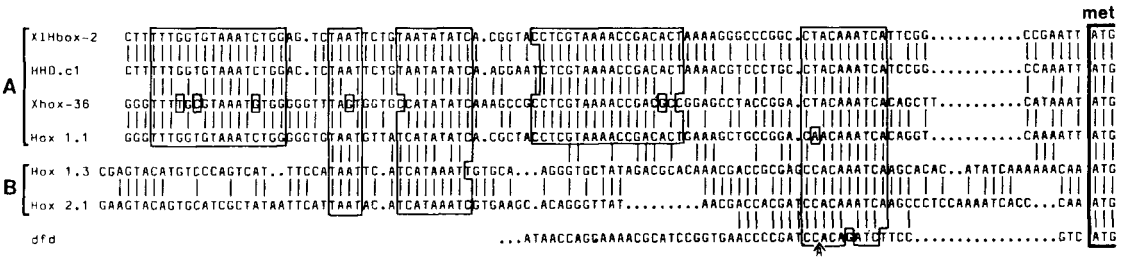

Fig. 2: **Conserved regions in the 5' non-coding regions in homeobox-containing genes.** The initiator AUG is at the right hand end of the figure. The sequences are those described in the legend to Figure 1. Vertical lines between pairs of sequences indicate identical nucleotides. Regions of high conservation are boxed. *Fushi tarazu* (17) has a small stretch of nucleotides (*TCTGATTTTGCTATATAT*, approximately −100 upstream of the AUG) identical to the second and third boxed regions of X1Hbox 2 and HHO.cl (not shown). Percentage of identity between different genes in the regions shown in the figure is as follows: X1Hbox-2 to HHO.cl: 91%; X1Hbox-2 to Hox1.1: 77%; X1Hbox-2 to Xhox-36: 63%; Hox 1.1 to Xhox-36: 77%; Hox 2.1 to Hox 1.3: 63%.
Methods: Searches for sequences similarities and alignments were made using a matrix program by Paollela (23), the programs GAP, BESTFIT, COMPARE, DOTPLOT and LINEUP (22), and a matrix program by T. Bürglin (unpublished).

tion common to the genes shown in Fig. 2. The best argument in favor of strong evolutionary pressure to maintain the region preceding the initiator AUG unchanged is provided by the homologues X1Hbox 2 (7) and HHO.cl (12). These two genes have been evolving for at least 350 million years, since amphibians and the mammalian ancestors separated. Their 5' non-coding regions next to the AUG (Fig. 2) are 91% conserved. This compares to a conservation of only 81% in the two homeobox regions, in which the protein reading frame is under strong pressure to remain invariant. The 3' non-coding regions do not have any detectable sequence conservation (this also applies to all the genes in Fig. 2, not shown). Even in the coding region between the conserved amino-terminus and the IYPWM pentapeptide there is very little conservation at the nucleotide level between these two genes. Since the 5' non-coding region is the most conserved part of all, we think it is conserved not just because the genes have a common evolutionary origin, but rather because it has a function.

A number of functions can be envisaged for this region, for example providing protein binding sites for transcriptional regulation. However, since the 5' leader conservations are closely associated with the initiation codon, an attractive hypothesis is that these sequences may convey some type of translational control shared by these genes. Translational control has been implicated in the expression of homeobox genes in *Drosophila,* and the best example is provided by the maternal gene *caudal,* whose mRNA is uniformly distributed in the egg but is only translated in the posterior part of the embryo (16, 21). A possibility of translational regulation has been shown for X1Hbox 2 in an experiment by Fritz and De Robertis (18), in which the deletion of most of the 5' leader of a cDNA clone, leaving behind only 26 nucleotides in front of the AUG, stimulated *in vitro* translation of SP6 mRNAs over 20-fold in both the reticulocyte and wheat germ systems. In addition, the proteins translated from both constructs were of the same size (18), providing further support to the view that the 5' conserved region is not translated. Clearly the function of these conserved non-translated regions will have to be tested further. for example by attaching them to reporter mRNAs such as globin. In the meantime, the present observations may help find similar conservations in other, non-homeobox, gene families. In addition, the conserved 5' sequences may provide useful probes for the isolation of related homeobox-containing genes by low stringency hybridization

ACKNOWLEDGEMENTS

We thank A. Fritz, K. Cho and G. Oliver for discussions, W. Wickner, L. Zipursky and K. Calame for comments on the manuscript, R. Harland for a preprint of his work, and B. Blumberg, M. Gribskov and J. Baumer for their help with the computer analysis. This work was supported by a grant from the NIH.

REFERENCES

1. McGINNIS, W.; GARBER, R.L.; WIRZ, J.; KUROIWA, A. and GEHRING, W.J. (1984) *Cell, 37:* 403-408.
2. SCOTT, M.P. and WEINER, A.J. (1984) *Proc. Natl. Acad. Sci., 81:* 4115-4119.
3. CARRASCO, A.E.; McGINNIS, W.; GEHRING, W.J. and De ROBERTIS, E.M. (1984) *Cell, 37:* 409-414.
4. GEHRING, W.J. (1987) *Science, 236:* 1245-1252.
5. HARVEY, R.P.; TABIN, C.J. and MELTON, D.A. (1986) *EMBO J., 5:* 1237-1244.
6. CONDIE, B.G. and HARLAND, R.M. (1987) *Development, 101:* Nº 3, in press.
7. WRIGHT, C.V.E.; CHO, K.W.Y.; FRITZ, A.; BURGLIN, T. and De ROBERTIS, E.M. (1987) *EMBO J., 6:* Nº 13, in press.
8. KRUMLAUF, R.; HOLLAND, P.W.; McVEY, J.H. and HOGAN, B.L.M. (1987) *Development, 99:* 603-617.
9. ODENWALD, W.F.; TAYLOR, C.F.; PALMER-HILL, F.J.; FRIEDRICH, V. Jr.; TANI, M. and LAZZARINI, R.A. (1987) *Genes Dev., 1:* 482-496.
10. KESSEL, M.; SCHULZE, F.; FIBI, M. and GRUSS, P. (1987) *Proc. Natl. Acad. Sci. USA, 84:* 5306-5310.
11. MAVILIO, F.; SIMEONE, A.; GIAMAOLO, A.; FAIELLA, A.; ZAPPAVIGNA, V.; ACAMPORA, D.; POIANA, G.; RUSSO, G.; PESCHLE, C. and BONCINELLI, E. (1986) *Naure, 324:* 664-668.
12. SIMEONE, A.; MAVILIO, F.; ACAMPORA, D.; GIAMPAOLO, A.; FAIELLA, A.; ZAPPAVIGNA, V.; D'ESPOSITO, M.; PANNESE, M.; RUSSO, G.; BONCINELLI, E. and PESCHLE, C. (1987) *Proc. Natl. Acad. Sci. USA, 84:* 4914-4918.
13. SCHNEUWLY, S.; KUROIWA, A.; BAUMGARTNER, P. and GEHRING, W.J. (1986) *EMBO J., 5:* 783-739.
14. REGULSKI, M.; McGINNIS, N.; CHADWICK, R. and McGINNIS, W. (1987) *EMBO J., 6:* 767-777.
15. WILDE, C.D. and AKAM, M. (1987) *EMBO J., 6:* 1393-1401.
16. MLODZIK, M. and GEHRING, W.J. (1987) *Cell, 48:* 465-478.
17. LAUGHON, A. and SCOTT, M.P. (1984) *Nature, 310:* 25-31.
18. FRITZ, A. and De ROBERTIS, E.M. (1987) submitted for publication.
19. MARTIN, G.R. *et al.* (1987) *Nature, 325:* 21-22.
20. BOPP, D.; BURRI, M.; BAUMGARTNER, S.; FRIGERIO, G. and NOLL, M. (1986) *Cell, 47:* 1033-1040.
21. MacDONALD, P.M. and STRUHL, G. (1986) *Nature 324:* 537-545.
22. DAYHOFF, M.O.; SCHWARTZ, R.M. and ORCUTT, B.C. (1979) In Atlas of Protein Sequence and Structure, Vol. 5, suppl. 3 (ed. Dayhoff, M.O.), 345-362 (National Biomedical Research Foundation, Washington, D.C.).
23. BURGLIN, T.R. (1986) *CABIOS, 2:* 99-101.
24. DEVEREUX, J.; HAEBERLI, P. and SMITHIES, O. (1984) *Nucleic Acids Res., 12:* 387-395.
25. PAOLLELA, G. (1985) *CABIOS, 1:* 43-49.