Volume 16 Number 19 1988 Nucleic Acids Research

Primary structure, developmentally regulated expression and potential duplication of the zebrafish homeobox gene ZF-21

Pål Rasmus Njølstad, Anders Molven, Ivar Hordvik, Jaran Apold and Anders Fjose

Laboratory of Biotechnology, University of Bergen, PO Box 3152 Årstad, N-5029 Bergen, Norway

Received August 3, 1988; Accepted September 8, 1988 Accession nos X12802, X12803

ABSTRACT

We report the molecular cloning and characterization of a cDNA derived from a zebrafish gene (ZF-21) related to the mouse homeobox containing gene Hox2.1. Interesting information about the differential conservation of various domains was gained from comparisons between the putative protein sequences from ZF-21 (275 amino acids) and Hox2.1 (279 aa). A separate DNA binding domain including the ZF-21 homeodomain and 36 additional flanking residues is completely identical to the C-terminal part of Hox2.1. As a consequence, these two mouse and zebrafish proteins must have identical DNA binding properties. A lower level of sequence identity between the N-terminal coding regions of ZF-21 and Hox2.1 reduces the total protein homology to 81%. However, short stretches of perfect homology in these N-terminals suggests that the essential biochemical functions are the same. As expected for true homologues, the ZF-21 and Hox2.1 genes also share extensive similarities with respect to non-coding sequences and temporal expression during embryogenesis. The finding of a potential ZF-21 duplication is discussed in relation to functional and evolutionary aspects of vertebrate homeobox genes.

INTRODUCTION

Mutant analyses led to the discovery of segmentation and homeotic genes that specify the number, polarity and identity of segments in the fruit fly Drosophila melanogaster (reviewed in ref. 1). The homeobox, a 180 base pair (bp) highly conserved region of several of these genes (2, 3, 4), codes for a domain found to bind with sequence specificity to DNA (5, 6, 7). The discovery of vertebrate homeobox containing genes with embryonic expression (8, 9, 10) suggested a possible involvement in developmental control, as previously described for the Drosophila segmentation and homeotic genes. Recent analyses of their temporal and spatial expression patterns in mouse embryos have provided evidence consistent with such a regulatory role (reviewed by ref. 11).

As a model system for the study of embryogenesis in higher animals, the zebrafish, Brachydanio rerio, has several

advantages (12, 13). Firstly, genetic tools for generating and characterizing mutants have been established (14). These techniques have been applied to generate a number of zebrafish developmental mutants (C. Kimmel, personal communication). The recent use of microinjection techniques to establish transgenic lines (15) also provides extended possibilities for molecular analysis of zebrafish gene functions. Finally, the special characteristics of zebrafish embryos, including rapid development (16) and a central nervous system with clear segmental features (17) make this species especially well suited for analyzing the biological functions of homeobox genes.

We have previously reported the zebrafish homeobox sequence ZF-21 (18). This study concerns the primary structure of the corresponding gene. The complete protein coding sequence of the gene is described and comparisons reveal that the murine homeobox gene Hox2.1 (19) is the true homologue of ZF-21. In the zebrafish gene we also identify 5'- and 3'-untranslated sequences that are conserved in homeobox genes from distantly related vertebrate species. Moreover, we present evidence that another zebrafish homeobox gene ZF-54, represents a duplication of ZF-21.

MATERIALS AND METHODS

Construction and screening of a cDNA library

Total RNA was extracted from staged zebrafish embryos (0-48 hours old) and adult fish by the guanidinium/cesium chloride centrifugation method described by MacDonald et al. (20). Poly(A)$^+$ RNA was selected by one passage over an oligo(dT)-cellulose column (New England Biolabs). The cDNA was synthesized using a commercial kit (Amersham). One ug of cDNA was cloned into the phage vector λgt10 by standard techniques (21) resulting in a library of 350 000 recombinants. The library was screened with a nick-translated probe under high stringency conditions.

Screening of a genomic library

500 000 plaques of a λEMBL3 library derived from zebrafish DNA (22) were screened using ^{32}P-labelled homeobox-containing DNA fragments. Phage plaque offprints on nitrocellulose papers (Schleicher & Schuell), performed according to standard methods (21), were hybridized under conditions of reduced stringency (3).

Subcloning and sequencing

Fragments from restriction endonuclease and exonuclease III digests (23) of the lambda clones were subcloned into the vectors

pGEM-3/4 (Promega Biotec) and sequenced by the chain termination method.

Northern analyses

The quality and quantity of the RNA preparations used for Northern analyses were tested by electrophoresis of small aliquots together with RNA standards on EtdBr stained agarose gels. Samples of 30 µg total RNA and 5 µg poly(A)$^+$ RNA were separated by electrophoresis in agarose-formaldehyde gels and transferred to nitrocellulose filters (Schleicher & Schuell; 24). Filters were hybridized (3) with ^{32}P-labelled DNA fragments, and final washing was in 0.2x SSC, 0.1% SDS at 68^{o}C.

RESULTS AND DISCUSSION

cDNA isolation and sequence analysis

An embryonic zebrafish cDNA library was screened with a probe consisting of the homeobox region of the previously characterized zebrafish gene ZF-21 (18). A single positive clone isolated proved to contain a 2.1 kilobase (kb) EcoRI insert (C21c1) which is shown in Fig. 1. Short fragments of this cDNA were subcloned and sequenced by the dideoxy method (Materials and methods). As shown in Fig. 2, the cDNA contains 2073 bp, of which a middle piece of 616 bp (position 581 to 1196) is non coding (see below). The first ATG triplet, preceeded by one in-frame stop codon, is the most probable initiator codon. Multiple stop codons in all three reading frames are present in this region of the ZF-21 cDNA. This was surprising, because the flanking sequences exhibit extensive homology with the murine Hox2.1 mRNA (19). Moreover, perfect consensus acceptor and donator splice site sequences are located at the 5'- and 3'-ends of this part (Fig. 2). We therefore assume that the 616 bp region in the middle of the ZF-21 cDNA is an intron and that the cDNA was synthesized from a partly processed mRNA. In fact, when both cytoplasmatic and nuclear RNA are used for the purification of poly(A)$^+$ RNA, intron containing clones are not uncommonly isolated from the corresponding cDNA libraries (25, 26). When aligned with Hox2.1, the first ATG triplet (position 1, Fig. 2) in the ZF-21 cDNA corresponds directly to the initiator codon predicted for the murine gene. Then follows a single long open reading frame (ORF) of 825 bases, the intron beeing excluded. The 180 bp homeobox is located in the beginning of the last coding exon. Thus, the overall gene structure of ZF-21 is very similar

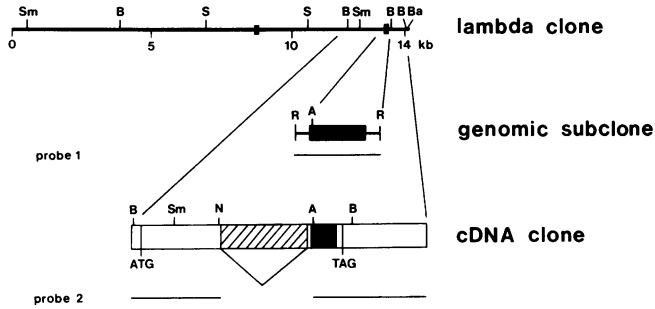


Fig. 1. Structure of the clone λC21 including the two homeobox containing genes ZF-21 and ZF-22 (ref. 18; "B" has been renamed ZF-22). The genomically derived subclone C21 and the cDNA clone C21c1 are outlined. Solid rectangles represent zebrafish homeobox sequences. An intron of the 2.1 kb unspliced cDNA is indicated by the cross-hatched central part. The positions of the initiaton and termination codons of the deduced protein-coding region are shown. Probe 1 and 2 were hybridized to Northern blots (see text). Enzymes used for restiction mapping were ApaI (A), BamHI (Ba), BglII (B), NcoI (N), RsaI (R), SalI (S) and SmaI (Sm).

to that of Drosophila, Xenopus, murine and human homeobox containing genes.

Structure of the putative ZF-21 protein

In a previous report (18) we have by sequencing of genomic DNA demonstrated that the 81 aa COOH-terminal part encoded by the ZF-21 gene is identical to the corresponding domain of the murine Hox2.1 protein (19). The deduced ZF-21 protein, translated from the 825 bp ORF, consists of 275 aa and the Antennapedia-type (Antp; 3) homeodomain reported previously (18) is located close to the COOH-terminus (Fig. 2). Most homeobox containing genes characterized so far belong to this class. In the zebrafish genome, we have till now identified 10 different Antennapedia-type homeoboxes (22). However, we have also described (27) one engrailed-like homeobox (en; 28), and at least one more sequence of this class has been detected in total genomic DNA (27). Thus, the relative distribution of these two categories of homeobox genes (with an excess of Antp-like sequences) is very similar to Drosophila and other vertebrate species (29, 30, 31).

A stretch of 5 residues located upstream of the homeodomain has been reported to be conserved in many homeobox genes from several species. The consensus Ile/Val Phe/Tyr Pro Trp Met is

Fig. 2. Nucleotide and deduced protein sequence of the ZF-21 cDNA clone C21c1. Numbering starts at the first ATG codon which is underlined, and a star indicates the stop of the protein coding region. The intron of the unspliced cDNA is located between arrowheads, consensus acceptor and donator splice sites are underlined. The homeobox and the conserved pentapeptide are boxed. In the 3'-region, a line is located underneath an ATT-repeat and ATTTA/ATTTT motifs, which are also found in the corresponding region of a murine and a human homologue (see text).

found in five murine homeobox containing genes (19, 32, 33, 34, 35), five from Xenopus (36, 37) and three human homeobox genes (38, 39). The same sequence is present in the Drosophila genes Deformed (Dfd; 40), Ultrabithorax (Ubx; 41) and in slightly modified versions in caudal (cad; 42) and fushi tarazu (ftz; 6).

Also the putative ZF-21 protein contains the pentapeptide in the typical position relative to the homeodomain (Fig. 2). Since this sequence is very conserved in proteins from distantly related species, one would expect it to play some important role for the function of the homeobox containing gene products. The discovery of a similar sequence in haemoglobins and myoglobins (43), where it makes a bend which separates two regions of the globin protein (44), has led to the proposal that the pentapeptide separates the DNA binding domain of the COOH-terminal from the remaining part

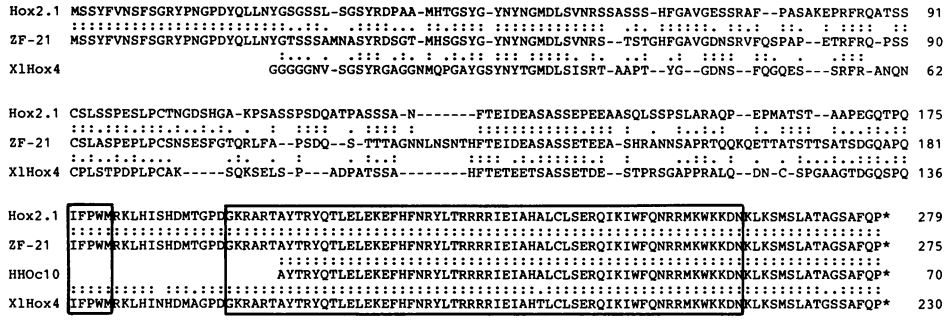


Fig. 3. Comparisons between the deduced proteins from the murine Hox2.1, Xenopus XlHox4 and human HHO.c10 genes and the putative zebrafish ZF-21 protein. Double dots indicate identities, single dots indicate conserved amino acids according to structurally similar groupings, which are nonpolar chain R groups (M, L, I, V, C), aromatic or ring-containing R groups (A, G, S, T, P), acidic and uncharged polar R groups (D, E, N, Q) and basic polar R groups (K, R, H). Hyphens denote deletions while stars show the stops of the proteins. The boxed regions are the pentapeptides and homeodomains.

of the protein (37). If this is correct, also the sequences immediately flanking the homeodomain could be necessary for generating the proper DNA binding capacity. In this connection, it is worth noting that the homeodomain by itself is sufficient for sequence specific DNA binding, however, the binding affinity was found to be relatively low (W. J. Gehring, pers. comm.). Therefore, the remaining COOH-terminal sequences possibly help to increase the binding affinity and/or change the binding specificity.

An alignment between the full-length putative proteins of ZF-21 and Hox2.1 (shown in Fig. 3) reveals that the identity covers the first 25 and the last 96 aa residues of the proteins. Overall, 224 of the 275 aa (81%) in ZF-21 are shared between the two proteins, and the homology is even more extensive (256/275 aa, 93%) when conservative amino acid changes are taken into account (Fig. 3). Interestingly, the sequence identity between the two proteins in the COOH-terminal end starts at the position of the pentapeptide. If, as suggested above, this 96 aa region constitutes a separate DNA binding domain, then the ZF-21 and Hox2.1 proteins must bind with the same affinity to identical DNA sequences. Therefore, the regulatory circuits in which these two proteins are integral parts, most likely contain regulatory

DNA elements that have been highly conserved during the 400 million years of vertebrate evolution.

In contrast, only 59% of the N-terminal part of the ZF-21 protein is identical to Hox2.1. Interestingly, this homology is not uniformly distributed, but rather localized to specific segments of the N-terminal sequence (Fig. 3). The remaining parts are very different and even includes insertions and deletions of amino acid residues. As a consequence of these sequence differences, the N-terminal domains of the ZF-21 and Hox2.1 proteins probably have major differences in tertiary structure. On the other hand, the conserved sequence elements should reflect some functional similarities between the two domains. Several other DNA binding regulatory proteins have been found to contain domains which in spite of strong differences in tertiary structure are functionally equivalent (45). In the light of these observations, one may assume that the ZF-21 and Hox2.1 proteins have essentially the same biochemical functions.

Also a Xenopus gene (XlHbox4) is known to be a potential homologue of Hox2.1 (Fig. 3; ref. 40). However, in this case not all, but 91 out of the 96 COOH-terminal aa residues are identical. Moreover, the homology in the remaining part of the proteins is somewhat less striking. Similar to ZF-21, however, is the presence of short stretches of highly conserved aa residues. Since the most N-terminal part of XlHbox4 has not yet been determined (37), it is premature to conclude whether or not this protein is the true Xenopus homologue of ZF-21/Hox2.1. Related to this question is also the finding that many of the Hox-loci in the mouse are duplicated (35 and A. Graham and R. Krumlauf, pers. comm.). The Hox2.1 gene, for instance, shares a high degree of sequence identity with Hox1.3 (32, 46) and the extent of homology to additional Hox-loci is currently under investigation (R. Krumlauf et al, pers. comm.). When compared to ZF-21, the Hox1.3 protein shares only 52% (142/275 aa) sequence identity (not shown) and are thus less related than ZF-21/Hox2.1. This is similar to the degree of sequence identity observed between Hox2.1/Hox1.3 (53%, 147/279 aa; not shown). However, we have searched for further evidence that the ZF-21 gene is the true homologue of Hox2.1. Such information was obtained from sequence comparisons of nontranslated regions and analyses of the developmental expression pattern.

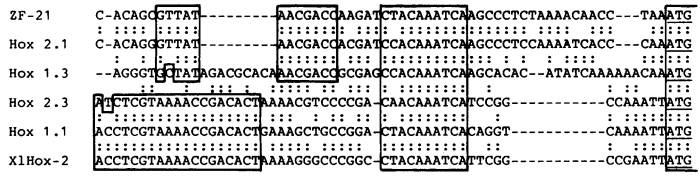


Fig. 4. Conserved sequence elements in the 5'-non coding regions of homeobox containing genes (see text). The initiaton ATG codon is underlined at the right-hand end of the figure. Vertical dots between pairs of sequences indicate identical nucleotides. Common regions of high conservation are boxed.

Sequence conservation in nontranslated regions of ZF-21

Comparison of the 5'-sequence of ZF-21 to the corresponding parts of the murine Hox2.1 (19), Hox1.3 (32), Hox2.3 (34), Hox1.1 (33) and the Xenopus XlHbox2 (47) genes, shown in Fig. 4, discloses a very high degree of sequence conservation. There are several regions within the upstream sequences where the level of homology is especially high. In addition, these regions seem to vary between different homeobox genes, indicating that at least two groups can be distinguished. The similarity between ZF-21 and Hox2.1 is particularly striking, 46 of the 53 nucleotides upstream of the ATG are identical. Moreover, a 48 bp ORF in the 5'-end of the ZF-21 cDNA (Fig. 4) is shared with Hox2.1. Hox1.3 displays a somewhat lower level of sequence identity with ZF-21 and Hox2.1 (Fig. 4), but still these three genes seem to belong to a separate group of related sequences. This conservation of 5'-sequences between genes from distantly related species probably reflects some functional role at the level of translational control (48). The discovery of short upstream ORFs in Hox1.1 and Xhox-36, proposed to encode a conserved peptide (49), supports the possible significance of the 5'-sequence in the regulatory translation of the homeodomain containing protein.

Another interesting characteristic of the C21c1 is that an ATT repeat and stretches of ATTTT and ATTTA located in the 3'-end of the clone (Fig. 2) are also found in the corresponding regions of Hox2.1 and HHO.c10 (10). Worth noting, only two ATTTT and ATTTA sequences and no ATT repeat are present in Hox1.3. ATTTT and ATTTA motifs frequently occur in the 3'-untranslated parts of transiently expressed genes and may be recognition signals for molecules that specifically degrades certain mRNA

molecules (50, 51, 52). Most interestingly, the ATT repeat has recently by cycloheximide experiments been shown to have an influence on Hox2.1 mRNA stability (N. Papalopulu and R. Krumlauf, pers. comm.). Thus, extensive sequence conservation in the untranslated upstream region and even conserved sequences in the 3'-end give additional evidence to the assumption that ZF-21 and Hox2.1 are true homologues. Whether homologous sequences also appear in the probably untranscribed introns must await further investigation since the Hox2.1 intron has not yet been reported.

We could not find any polyadenylation signal or poly(A)$^+$ tail adjacent to the 3'-end of C21c1 (Fig. 2). Thus, when excluding the intron, the ZF-21 cDNA contains only 1457 bases. However, the transcript size as judged from Northern blots (Fig. 5) is 2.3 kb. The murine homologue Hox2.1 contains about 0.5 kb additional 3'-sequence. Hence it seems likely that C21c1 has arisen by specific priming from an internal A-rich region in the ZF-21 mRNA.

Analysis of ZF-21 expression

To investigate the developmental pattern of ZF-21 expression, Northern blots containing RNA from different zebrafish embryonic stages were prepared (Fig. 5). Under stringent conditions, DNA probes including the ZF-21 homeobox region (Fig. 1) were hybridized to blots of total RNA (Fig. 5a; probe 2) and poly(A)$^+$ RNA (Fig. 5b; probe 1).

Shortly after fertilization (0 hours stage; ref. 16) no signal was detected and this makes maternal deposits of ZF-21 mRNA unlikely. A detectable level of the 2.3 kb ZF-21 transcript has accumulated by the 12 1/2 hours stage, when the embryos undergo an early phase of somite formation (3-4 somites; ref. 16). Persistent expression of the gene occurs in later developmental stages, and a peak level of the transcript is obtained about 48 hours after fertilization. At this stage, the morphological differentiation is considerable: the heart functions, the head and tail are freed from their attachments to the yolk sac and body pigmentation has appeared. The ZF-21 probe also displayed a weak signal in a lane of 30 µg total RNA extracted from adult zebrafish (not shown). The finding of hybridization signals in the poly(A)$^+$ RNA fraction and lack of detectable transcipts in poly(A)$^-$ RNA (Fig. 5b) clearly demonstrates that the signals detected on the total RNA blots correspond to mRNA molecules.

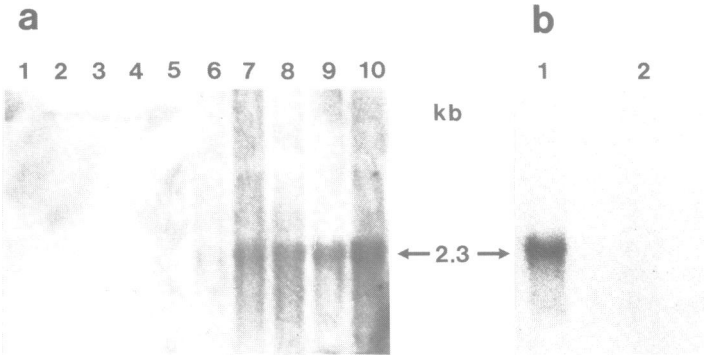
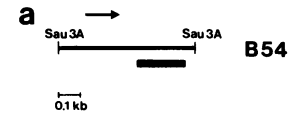


Fig. 5. Expression pattern of the ZF-21 gene during development. (a) Northern blot of 30 µg total RNA from staged embryos. Lanes 1-10 contain RNA from 0, 2 1/2, 5, 7 1/2, 10, 12 1/2, 15, 21, 29 and 48 hours old embryos, respectively. The filter was hybridized with the ZF-21 cDNA probe 2 (see Materials and methods and Fig. 2). (b) Enrichment of ZF-21 transcripts in poly(A)$^+$ RNA from staged embryos. Lane 1 contains 5 µg poly(A)$^+$ RNA from 48 hours embryos and 5 µg non-poly(A)$^+$ RNA from the same stage was loaded in Lane 2. The transcript was detected by probing the filter with the genomic ZF-21 probe 1 (Materials and methods; Fig. 2).

Temporally regulated expression of ZF-21 suggests that the gene is of importance in zebrafish development, as the homeobox containing genes are in Drosophila embryogenesis. Since transcripts from stages prior to somite formation cannot be detected, ZF-21 probably plays no fundamental role in the very early developing zebrafish embryo. Interestingly, no other zebrafish homeobox gene characterized so far is expressed in this presomite period of development either (18, 22, 27, 53 and I. Hordvik et al. and P. R. Njølstad and A. Fjose, unpublished). The finding of a relatively steady level of the ZF-21 transcript from the initiation of somitogenesis and throughout the remaining embryonic period is in accordance with the temporal pattern of transcription reported for the murine and Xenopus ZF-21 homologues (19, 37, respectively). Similar developmental expression patterns have also been observed for several other homeobox genes from mice and frogs (32, 34, 35, 54, 37). However, many Xenopus and most murine and human homeobox genes (including Hox2.1 and Hox1.3) give rise to multiple transcripts with differential patterns of tissue distribution (37, 19, 32, 35, 44, 10, 38, 39). Only a single transcript from ZF-21 can be detected, this also contrasts with the expression



b

```
ZF-54                                                                -15                -1
                                                                      ATG ACT AAG CCT GAT
                                                                      Met Thr Lys Pro Asp
ZF-21/Hox2.1                                                           -   -  Gly  -   -
Hox1.3                                                                Ile Gly Gly  -  Glu

                1                                                                       60
ZF-54           GAA AAC AGG GCT CGA ACA GCT TAC ACT CGC TAC CAA GCG CTC GAG CTG GAG AAA GAG TTC
                Glu Asn Arg Ala Arg Thr Ala Tyr Thr Arg Tyr Gln Ala Leu Glu Leu Glu Lys Glu Phe
ZF-21/Hox2.1    Gly Lys  -   -   -   -   -   -   -   -   -   -  Thr  -   -   -   -   -   -   -
Hox1.3          Gly Lys  -   -   -   -   -   -   -   -   -   -  Thr  -   -   -   -   -   -   -

                                                                                        120
ZF-54           CAT TTC AAT AGA TAT CTG ACG CGC AGG AGG AGA ATA GAG ATA CGT CAC GCT CTC TGC CTC
                His Phe Asn Arg Tyr Leu Thr Arg Arg Arg Arg Ile Glu Ile Arg His Ala Leu Cys Leu
ZF-21/Hox2.1     -   -   -   -   -   -   -   -   -   -   -   -   -  Ala  -   -   -   -   -
Hox1.3           -   -   -   -   -   -   -   -   -   -   -   -   -  Ala  -   -   -   -   -

                                                                                        180
ZF-54           TCA GAG AGG CAA ATC AAA ATC TGG TTC CAA AAT CGC AGG ATG AAA TGG AAG AAG GAC AAT
                Ser Glu Arg Gln Ile Lys Ile Trp Phe Gln Asn Arg Arg Met Lys Trp Lys Lys Asp Asn
ZF-21/Hox2.1     -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -
Hox1.3           -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -   -

                                                                                    231
ZF-54           AAA CTG AAA AGT ATG AGT CTC GCT ACA GCG GGG AGC GCG TTT CAA AAT TGA
                Lys Leu Lys Ser Met Ser Leu Ala Thr Ala Gly Ser Ala Phe Gln Asn  *
ZF-21/Hox2.1     -   -   -   -   -   -   -   -   -   -   -   -   -   -  Pro  *
Hox1.3           -   -   -   -   -   - Met  -  Ala  -   -  Gly  -   -  Arg Pro  *
```

Fig. 6. (a) Localization of the zebrafish homeobox ZF-54 within the 0.6 kb Sau3A subclone B54 which was derived from the clone λB54. The homeobox is indicated by a solid rectangle and an arrow shows the 5'-3'-orientation. (b) Nucleotide and deduced protein sequences of the zebrafish homeobox gene ZF-54. The protein is also compared to the homeodomains deduced from ZF-21 and the two murine genes Hox2.1 and Hox1.3. The homeobox extends from position 1 to 180. Stop codons are indicated by stars, identical amino acids are denoted by hyphens and sequence deviations are illustrated by the substituting amino acid.

pattern of the previously reported Antp-like homeobox sequences from zebrafish (18, 22, 53). However, similar to ZF-21 only a single XlHbox4 derived message has been reported.

The small size makes dissection of zebrafish embryos to investigate a possible tissue-specific expression of homeobox genes difficult. In mice and humans, the main site of expression of these genes has been confined to the CNS and the spatial distribution of transcripts in mice has appeared in a restricted manner (10, 38, 39 and 11). In a previous report we have shown that also a zebrafish homeobox gene is transcribed mainly in the brain and spinal cord of embryos (53). Preliminary data indicate tissue-specific expression of ZF-21 as well, and as observed in mouse embryos, the transcription was found to be spatially restricted within the CNS (P. R. Njølstad and A. Fjose, unpublished). The similarity in the expression pattern of

homeobox genes from mice, zebrafish and Drosophila points to a role for such vertebrate genes in the determination of cell fate, and this role may be to establish domains of positional value in the vertebrate embryo. A more complete understanding of the biological role of homeobox genes can, however, only be elucidated by experiments where the function is directly examined, for example by altering the expression of these genes in transgenic animals. We believe the zebrafish, as a model system, will be of importance in such efforts, since stable lines of transgenic zebrafish and developmental mutants have been generated (15 and C. Kimmel, pers. comm.) and a segmental nature of the CNS reported (17).

The zebrafish genome contains a potential ZF-21 duplication

We have previously reported the isolation of 26 independent zebrafish homeobox containing clones from a genomic library (22). These clones could be classified into at least eight different groups, and from one of the groups, consisting of 4 clones, the clone λB54 was analyzed. The homeobox was shown to be contained in a 0.6 kb Sau3A restriction fragment (Fig. 6a) which subsequently was subcloned into a plasmid vector and the DNA sequence determined (Materials and methods).

The 246 bp homeobox containing region (ZF-54) of B54 shown in Fig. 6b constitutes a single ORF (Fig. 6b). Although only 68% of the nucleotide sequence of ZF-54 (position 1 to 180) is identical to the Antp homeobox, the deduced protein sequences are closely related (85%). This high degree of conservation is more than sufficient to classify the zebrafish gene to the Antennapedia-class. Most interestingly, however, is the close sequence relationship with ZF-21/Hox2.1. As shown in Fig. 6b, 56 of the 60 residues in the homeodomains are shared. Moreover, extensive homology is also present between the homeodomain flanking regions of these proteins: 4 of 5 residues encoded by the 5'-region and 15 of 16 aa derived from the 3'-end are identical. At the DNA level, the ZF-54 and ZF-21 homeoboxes are only 82% homologous, but the differences are mostly due to third base changes (not shown). Also worth noting, the stop codons are situated in the same positions. These close similarities between the putative ZF-54 and ZF-21/Hox2.1 proteins clearly are suggestive of a common origin for the regions encoding these DNA binding domains. However, it is premature to conclude whether this duplication involved the whole gene, since the DNA sequence encoding the remaining N-terminal part of the ZF-54 protein is

not yet available. Supporting the view that a complete gene duplication has occurred is the finding of diverged copies of Hox2.1 and other Hox-genes in the mouse genome. These genes are organized in an almost identical manner in two gene clusters, Hox-1 and Hox-2 (35) and additional clusters are beeing characterized (A. Graham and R. Krumlauf, pers. comm.). In light of this, the finding of closely linked homeobox genes (C21, see Fig. 1) and a potential duplication (ZF-21/ZF-54) in the zebrafish genome is suggestive of a similar situation as in mice and humans (35 and R. Krumlauf et al., pers. comm.). If this assumption is correct, then it is very likely that all other vertebrate species also have groups of duplicated Hox-like genes. This would clearly contrast the situation in Drosophila where the extent of gene duplication events have been more limited (40).

Accumulating evidence now favours the developmental regulatory role of vertebrate homeobox genes (35, 55, 56, 57 and P. W. H. Holland, pers. comm.) and therefore the structural organization of the corresponding gene complexes is of particular interest when discussing the evolvement of different species. As compared to Drosophila, a common characteristic of vertebrate organisms is the high degree of morphological complexity, and probably an extensive increase in the number of developmental regulatory genes was a necessary prerequisite for their evolution. This would imply that multiplication of homeobox genes occurred in early vertebrate or prevertebrate ancestors and possibly additional duplication events took place at later periods of evolution. Answers to these questions can be found by comparing how homeobox genes are structurally organized in the genomes of distantly related vertebrate species. In this connection, the characterization of zebrafish homeobox genes is particularly useful and these analyses may help to identify the more fundamental regulatory molecules which are likely to be the most conserved.

ACKNOWLEDGEMENTS

We dedicate this paper to the late Professor Kjell Kleppe for his support and encouragement during our work. We thank Rein Aasland for help with the computer analyses and Drs. Anthony Graham, Nancy Papalopulu, Robb Krumlauf, Peter Holland, Charles Kimmel and Walter Gehring for sharing unpublished results. A.F.,

A.M. and P.R.N. were supported by fellowships and grants from the Norwegian Research Councils NAVF and NTNF.

REFERENCES

1. Scott, M.P. and Carroll, S.B. (1988) Cell 51,689-698.
2. Gehring,W.J. (1987) Science 236,1245-1252.
3. McGinnis,W., Levine,M., Hafen,E., Kuroiwa,A. and Gehring,W.J. (1984) Nature 308,428-433.
4. McGinnis,W., Garber,R.L., Wirz,J., Kuroiwa,A. and Gehring,W.J. (1984) Cell,403-408.
5. Desplan,C., Theis,J. and O'Farrell,P.H. (1985) 318,630-635.
6. Laughon,A. and Scott,M.P. (1984) Nature 310,25-31.
7. Hoey,T. and Levine,M. (1988) Nature 332,858-861.
8. Carrasco,A.E., McGinnis,W.J., Gehring,W.J. and De Robertis,E.M. (1984) Cell 37,409-414.
9. McGinnis,W., Hart,C.P., Gehring,W.J. and Ruddle,F.H. (1984) Cell 38,675-680.
10. Simeone,A., Mavilio,F., Bottero,L., Giampaolo,A., Russo,G., Faiella,A., Boncinelli,E. and Peschle,C. (1986) Nature 320,763-765.
11. Feinberg,A.A., Utset,M.F., Bogarad,L.D., Hart,C.P., Awgulewitsch,A., Ferguson-Smith,A., Faisod,A., Rubin,M. and Ruddle,F.H. (1987) Current topics in murine developmental biology 23,233-256.
12. Marcey,D. and Nüsslein-Volhard,C. (1986) Nature 321,380-381.
13. Kimmel,C. and Warga,R.W. (1988) Trends Genet. 4,68-74.
14. Streisinger,G., Walker,C., Dower,N., Knauber,D. and Singer,F. (1981) Nature 291,293-296.
15. Stuart,G.W., McMurray,J.V. and Westerfield,M. (1988) Development 103,403-412.
16. Hisaoka,K.K. and Battle,H.I. (1958) J. Morph. 102,311.321.
17. Hanneman,E., Trewarrow,B., Metcalfe,W.K., Kimmel,C. and Westerfield,M. (1988) Development 103,49-58.
18. Njølstad,P.R., Molven,A. and Fjose,A. (1988) FEBS Lett. 230,25-30.
19. Krumlauf,R., Holland,P.H., McVey,J.H. and Hogen,B.L.M. (1987) Developmet 99,603-617.
20. MacDonald,R.J., Swift,G.H., Przybyla,A.E. and Chorgwin,J.M. (1987). In Berger,S.L. and Kimmel.A.R. (eds), Methods in Enzymology, Academic Press,Inc., Orlando, Florida, Vol 152, pp.219-226.
21. Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) Molecular Cloning. A Laboratory Manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
22. Eiken,H.G., Njølstad,P.R., Molven,A. and Fjose,A. (1987) Biochem. Biophys. Res. Comm. 149,1165-1171.
23. Henikoff,S. (1987) In Abelson,J.N. and Simon,M.I. (eds). Methods in Enzymology, Academic Press,Inc., Orlando, Florida, Vol 155, pp.156-165.
24. Lehrach,H., Diamond,D., Wozney,S.M. and Boedtker,H. (1977) Biochemistry 16,4743-4751.
25. Verde,P., Stoppelli,M.P., Galeffi,P., Di Nocera,P. and Blasi,F. (1984) Proc. Natl. Acad. Sci. USA 81,4727-4731.
26. Sharpe,P.T., Miller,J.R., Evans,E.P., Burtenshaw,M.D. and Gaunt,S.J. (1988) Development 102,397-407.
27. Fjose,A., Eiken,H.G., Njølstad,P.R., Molven,A. and Hordvik,I. (1988) FEBS Lett. 231,355-360.
28. Fjose,A., McGinnis,W. and Gehring,W.J. (1985) Nature 313,284-289.
29. Poole,S.J., Kauvar,L.M., Drees,B. and Kornberg,T. (1985) Cell 40,37-43.
30. Coleman,K.G., Poole,S.J., Weir,M.P., Soeller,W.C. and Kornberg,T. (1987) Genes Develop. 1,19-28.
31. Joyner,A.L. and Martin,G.R. (1987) Genes Develop. 1,29-38.

32. Odenwald,W.F., Taylor,C.F., Palmer-Hill,F.J., Friedrich,V.,Jr., Tani,M. and Lazzarini,R.A. (1987) Genes Develop. 1,482-496.
33. Kessel,M., Schulze,F., Fibi,M. and Gruss,P. (1987) Proc. Natl. Acad. Sci. USA 84,5306-5310.
34. Meijlink,F., de Laaf,R., Verrijzer,P., Destree,O., Kroezen,V., Hilkens,J. and Deschamps,J. (1987) Nucl. Acids Res. 15,6773-6786.
35. Graham,A., Papalopulu,N., Lorimer,J., McVey,J.H., Tuddenham,E.G.D. and Krumlauf,R. (1988) Genes Develop., in press.
36. Harvey,R.P., Tabin,C.J. and Melton,D.A. (1986) EMBO J. 5,1237-1244.
37. Fritz,A. and De Robertis,E.M. (1988) Nucl. Acids Res. 16,1453-1469.
38. Mavilio,F, Simeone,A., Giampaolo,A., Faiella,A., Zappavigna,V., Acampora,D., Poiana,G., Russo,G., Peschle,L. and Boncinelli,E. (1986) Nature 324,664-668.
39. 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.
40. Regulski,M., Harding,K., Kostriken,R., Karch,F., Levine,M. and McGinnis,W. (1985) Cell 43,71-80.
41. Wilde,L.D. and Akam,M. (1987) EMBO J. 6,1393-1401.
42. Mlodzik,M., Fjose,A. and Gehring,W.J. (1985) EMBO J. 4,2961-2969.
43. Stryer,L. (1981) Biochemistry, W.H.Freeman and Company, San Francisco.
44. Fermi,G. (1975) J. Mol. Biol. 97,237-256.
45. Sigler,P.B. (1988) Nature 333,210-212.
46. Fibi,M., Zink,B., Kessel,M., Colberg-Poley,A.M., Labeit,S., Lehrach,H. and Gruss,P. (1988) Development 102,349-359.
47. Wright,C.V.E., Cho,K.W.Y., Fritz,A., Bürglin,T.R. and De Robertis,E.M. (1987) EMBO J. 6,4083-4094.
48. Bürglin,T.R., Wright,C.V.E. and De Robertis,E.M. (1987) Nature 320, 701-702.
49. Kessel,M. and Gruss,P. (1988) Nature 332,117-118.
50. Shaw,G. and Kamen,R. (1986) Cell 46,659-667.
51. Brawerman,G. (1987) Cell 48,5-6.
52. Clemens,M.J. (1987) Cell 49,157-158.
53. Njølstad,P.R., Molven,A., Eiken,H.G. and Fjose,A. (1988) Gene, in press.
54. Le Mouellic,H., Condamine,H. and Brulet,P. (1988) Genes Develop. 2,125-135.
55. Cho,K.W.Y., Goetz,J., Wright,C.V.E., Fritz,A., Hardwicke,J. and De Robertis,E.M. (1988) EMBO J. 7,2139-2149.
56. Harvey,R.P. and Melton,D.A. (1988) Cell 53,687-697.
57. Keynes,R.J. and Stern,C.D. (1988) Development 103,413-429.