Available online at www.sciencedirect.com

SCIENCE DIRECT

Gene 346 (2005) 51–61

GENE
SECTION
EVOLUTIONARY GENOMICS

www.elsevier.com/locate/gene

Organization and base composition of tilapia *Hox* genes: implications for the evolution of *Hox* clusters in fish☆

Simona Santini*, Giorgio Bernardi

Molecular Evolution Laboratory, Stazione Zoologica "A. Dohrn", Villa Comunale, 80121 Napoli, Italy

Received 16 July 2004; received in revised form 5 October 2004; accepted 21 October 2004
Available online 29 January 2005
Received by Takashi Gojobori

Abstract

Hox genes encode DNA binding proteins that specify cell fate in the anterior–posterior axis of metazoan animal embryos. While each *Hox* cluster contains the same genes among the different mammalian species, this does not happen in ray-finned fish, in which both the number and organization of *Hox* genes and even *Hox* clusters are variables. Ray-finned fish are believed to have undergone an additional genome duplication that led to the presence of 8 *Hox* clusters (four twin pairs) in their ancestor. Here we describe the Tilapia (*Oreochromis niloticus*) *Hox* genes set in terms of gene content, clusters organization and base composition and compare it with those of pufferfish and zebrafish. We observed that in all these fish, when paralogous genes are conserved in both the twin clusters, the gene which has a lower GC level generally: (i) belongs to the less gene-rich (less conserved) cluster; (ii) has a reduced field of embryonic expression; or (iii) is a pseudogene. The relationship between the decrease of GC level and the loss of conservation and function of one of the paralogous genes from twin clusters is discussed.
© 2004 Elsevier B.V. All rights reserved.

Keywords: Homeobox; Cichlid; Pufferfish; Zebrafish; GC level

1. Introduction

Hox genes encode DNA binding proteins that specify the cell fate in the anterior–posterior (a–p) axis of metazoan animal embryos (Gehring, 1993; McGinnis and Krumlauf, 1992). They are characterized by their arrangement in genomic clusters, and by their colinearity, i.e., the correlation between chromosomal organization, time of activation, and domain of expression along the a–p axis (e.g. McGinnis and Krumlauf, 1992). The *Hox* genes have been classified into 13 paralogous groups according to their sequences homology and their positions in the clusters (Scott, 1992). The origin of vertebrates seems to be associated with an increase of the total gene number, possibly as a consequence of two rounds of genome duplication, the so-called '2R' hypothesis (Ohno, 1970). This would have resulted in two duplications of the entire ancestral *Hox* cluster to the four-cluster situation represented nowadays by the tetrapods (reviewed by Holland and Garcia-Fernandez, 1996). Most vertebrate paralogous groups do not have a full set of four genes anymore, as a result of secondary gene losses and often members of paralogous group have partially redundant functions (McClintock et al., 2001, Bruce et al., 2001; Amores et al., 2004). Mammalian genomes (particularly mouse, rat and human) contain 39 *Hox* genes that are organized in four clusters, namely A, B, C and D, in which different numbers and combination of genes have been maintained. The same situation is delineated by the isolation of many *Hox* genes from the tetrapods' developmental model systems, frog (*Xenopus laevis*) and chick (*Gallus gallus*). Therefore, there is no evidence to suggest differences from the 39 genes–4 cluster mammalian organization for the non-mammalian land vertebrates (reviewed by McGinnis and Krumlauf, 1992). The mammalian *Hox* organization was initially assumed as being general for all

☆ Accession numbers: AY757306–AY757355.

* Corresponding author. Tel.: +39 81 5833312; fax: +39 81 7641355.
E-mail address: santini@szn.it (S. Santini).

0378-1119/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2004.10.027

jawed vertebrates (*Gnathostomata*) and early PCR-based screens on teleosts failed to highlight any major discrepancy from this condition (Misof and Wagner, 1996). Successively, more extensive analysis of *Danio rerio* (zebrafish) *Hox* set revealed the existence of at least 48 *Hox* genes organized in seven clusters in this model ray-finned (Actynopterigian) fish (Amores et al., 1998). Further research showed that, while each *Hox* cluster contains the same genes among the different mammalian species (International Human Genome Sequencing Consortium, 2001; Mouse Genome Sequencing Consortium, 2002; Rat Genome Sequencing Project Consortium, 2004), this does not happen in ray-finned fish (Aparicio et al., 2002; Amores et al., 1998, 2004; Naruse et al., 2004). Both the number of conserved *Hox* genes and even the number of clusters are variables among the extant species of ray-finned fishes. There are at least 51 *Hox* genes arranged in nine clusters in the pufferfish *Takifugu rubripes* (Aparicio et al., 2002; Amores et al., 2004), 50 *Hox* genes organized in seven clusters have been described in the pufferfish *Spheroides nephelus* (Amores et al., 2004), 51 *Hox* genes organized in seven clusters exist in the zebrafish *D. rerio* (Amores et al., 1998) and at least 33, organized in at least seven clusters, have been identified in the medaka *Oryzias latipes* (Naruse et al., 2004; Kondo et al., unpublished). Ray-finned fish are believed to have undergone a third *taxon*-specific genome duplication (Amores et al., 1998), that led to the presence of eight *Hox* clusters (four twin pairs) in the ray-finned fish ancestor: Aa, Ab, Ba, Bb, Ca, Cb, Da and Db. This hypothesis seems to be more parsimonious than the possibility of multiple *Hox* cluster duplication events within the different teleosts lineages. Molecular clocks estimates suggest that this third genome duplication in ray-finned fish occurred about 350 My (Taylor et al., 2001).

Classical models (e.g. Haldane, 1933) predict that few duplicates should be retained in the genome over the long term. In the Lynch and Conery (2000) model, silencing and subsequent loss of duplicated genes are common and relatively rapid evolutionary events, occurring in the time span of few millions of years. Nevertheless, Nadeau and Sankoff (1997) estimated that around half of all duplicated genes have been maintained in extant vertebrate genomes. Many gene duplicates have been retained for tens of millions of years after the duplication event, indicating the existence of preservation mechanisms. To explain the retention and functionality of many pairs of duplicated genes, it has been proposed that, assuming that many genes may have multiple, often pleiotropic, and separable functions, the duplicates will retain complementary functions by acquiring complementary loss-of-function mutation in independent sub-functions (sub-functionalization). Both duplicates will then be required to recapitulate the original gene function (duplication–degeneration–complementation model, DDC, Force et al., 1999). The DDC model ensures that both copies go to fixation in the genome. Evidences have been found that support the DDC model of evolution in ray-finned fishes. For instance, the zebrafish *Hox*B5a and *Hox*B5b underwent sub-functionalization and their expression summarizes that of the ancestral gene, represented by the mouse *Hox*B5 gene, which did not undergo duplication.

Changes in number and in genomic organization of *Hox* genes are believed to have played an important role in metazoan body-plan evolution and it has been hypothesized that the genome duplication event contributed to the vast radiation of the teleosts (Amores et al., 1998) through the differentially resolved complementation of duplicated genes functions mentioned above. In fact, ray-finned fish species constitute almost half the total number of vertebrate species. An estimated over 20,000 living teleosts species, on a total of over 43,000 recognized vertebrate species, have been described (Nelson, 1994). The fact that different species of fish possess a different set of *Hox* genes makes them the most suitable system to understand the mechanisms behind such an unequal conservation of duplicated copies.

To evaluate the evolutionary modifications of *Hox* genes organization in ray-finned fish, we isolated and sequenced the *Hox* genes in *Oreochromis niloticus* (Tilapia), a *taxon* thought to be widely separated from those of medaka and zebrafish and close to pufferfish (Santini and Tyler, 1999). Then we compared the *Hox* gene number, organization in clusters and base composition in these four fish species, all of which have been proved to have at least seven *Hox* clusters and then to have been subjected to a further duplication event relative to the tetrapods. Our working hypothesis is that the gene number and organization of fish *Hox* clusters are still changing and in this view, we try to speculate about the fate of the evolving *Hox* genes in terms of the variability in their expression in the embryo and in their base composition proposing a model for the evolution and decay of duplicated genes. Our rationale to formulate such a model is based on the evidences that mutations are strongly AT-biased in many organisms (Gojobori et al., 1982; Echols et al., 2002; Alvarez-Valin et al., 2002). The AT bias is particularly evident in the third position of codons, which, due to the redundancy of the genetic code, is the most tolerant to mutations (Alvarez-Valin et al., 2002). As pseudogenes are apparently subject to no functional constraints, all mutations in them would be selectively neutral and would become fixed in the population with equal probability. Thus, the pattern of nucleotide substitutions in pseudogenes would reflect the pattern of spontaneous substitution mutations (Gojobori et al., 1982). In particular, when a gene is duplicated and only one copy remains functional, while the other copy experiences a reduction of constraints related to the loss of function, a decrease in the GC level of the latter ensues, towards a GC-poorness that makes it similar to the surrounding intergenic DNA sequences, generally lower in GC than coding sequences (cds) (Bernardi, 1995; Echols et al., 2002).

In the fish genome, *Hox* clusters are present in twin pairs (see above). In a few cases, both paralogous genes of the

twin *Hox* clusters are conserved, whereas in even fewer cases, one of the twin paralogous genes is still present as a pseudogene. Therefore, these *Hox* twin pairs are ideal candidates to evaluate the relationship between the decrease in GC level and the loss of function of one member of the pair of genes belonging to twin clusters. We observed that the genes belonging to the gene-richer cluster have a higher GC level than the genes on the gene-poorer cluster in each twin pair. Where one of the paralogous genes of the twin clusters had become a pseudogene, it presents a lower GC level compared to the corresponding functional gene on the twin cluster. Finally, in those cases in which a gene belonging to the gene-richer cluster has a GC level which is lower than the corresponding gene on the twin gene-poorer cluster, we observed that it also presents a reduction of its field of expression in the embryo and might therefore be in the process of reducing its function and finally becoming a pseudogene. We propose a mechanism in which the duplicated sub-functionalized genes, being subject to different selective pressure and different functional constraints, differently accumulate mutations. Therefore, these genes can be preserved for a long time after the duplication event and then become pseudogenes and get lost, as a consequence of the reduction of function and field of expression. The model we propose offers an explanation for differential retention of duplicated *Hox* genes among ray-finned fish over a period of time much longer than that expected on the basis of the traditional models.

2. Materials and methods

2.1. Amplification and sequencing of Hox genes

Genomic DNA of Tilapia was extracted from muscle tissue by following standard protocols. On the basis of the fish *Hox* genes sequences, we designed degenerate primers (available as supplemental material at the journal website) to amplify the complete DNA sequence of Tilapia *Hox* genes. In case of divergence of sequence of a particular gene among different species, the primers were designed based on the pufferfish sequence. This choice was due to the previous observation that Tilapia *Hox*Aa genes were highly similar to the corresponding ones in pufferfish (Santini et al., 2003). PCR was carried out in 25 ml reaction volumes, by using Taq DNA polymerase (Roche) under the following standard cycle: one denaturation step at 95 °C for 5 min, 35 cycles of 95 °C for 30 s, 45 °C for 45 s and 72 °C for 2 min, followed by a final elongation step at 72 °C for 7 min. PCR products were sequenced on a 48 capillars 3730 DNA Analyzer (AB Biosystems). The sequences were BLASTX-searched against the NCBI protein database (http://www.ncbi.nlm.nih.gov/BLAST/) and against the TBLASTX-searched against the pufferfish genomic database maintained at the MRC Rosalind Franklin Centre for Genomics Research (http://Fugu.hgmp.mrc.ac.uk/).

2.2. Phylogenetic analyses

The amino acid sequences were aligned by using the ClustalX 1.81 program (Thompson et al., 1997). Regions of sequences that were difficult to align were removed from the alignment. Neighbor-joining trees were constructed from these unambiguously aligned protein sequences by using the program PAUP (Swofford, 2000). Bootstrap values for the nodes were determined by analyzing 1000 bootstrap replicate data sets to estimate the strength of the groupings. The alignments are available online as supplemental material at the journal website.

2.3. Base composition analyses

We calculated the guanine–cytosine (GC) percentage in the whole cds and the percentage of GC in the third position of codons of the genes. The analyses were carried out by using the program CodonW available at the website http://bioweb.pasteur.fr/seqanal/interfaces/codonw.html.

3. Results and discussion

3.1. Analyses of PCR products

The amplification products obtained by PCR as described in Section 2.2 have been searched against the NCBI protein database (http://www.ncbi.nlm.nih.gov/BLAST/), against the Fugu genome database maintained at the MRC Rosalind Franklin Centre for Genomics Research (http://Fugu.hgmp.mrc.ac.uk/blast/), against the Tetraodon genome database at the Genoscope (http://www.genoscope.cns.fr/cgi-bin/recherche.cgi) and against the Danio genome database at the Wellcome Trust Sanger Institute (http://www.ensembl.org/Multi/blastview?species=Danio_rerio). Our analyses identified 50 different *Hox* genes, 38 of which were complete sequences. We were able to assign these genes to different paralogous groups based on sequence comparison and BLASTX analyses (Fig. 1).

3.2. Phylogenetic analyses

To confirm the cluster affiliation and orthology of the Tilapia *Hox* genes, we generated phylogenetic trees based on the alignments of the amino acid sequences encoded by known chordate *Hox* genes.

Phylogenetic trees were generated separately for individual paralogous groups by using the *Hox* sequences from amphioxus (*Branchiostoma floridae*) as an outgroup. The resulting trees for each paralogous group are shown in Fig. 2A. To confirm the groupings observed in these trees, we also generated the combined tree for all of the *Hox* genes (Fig. 2B).

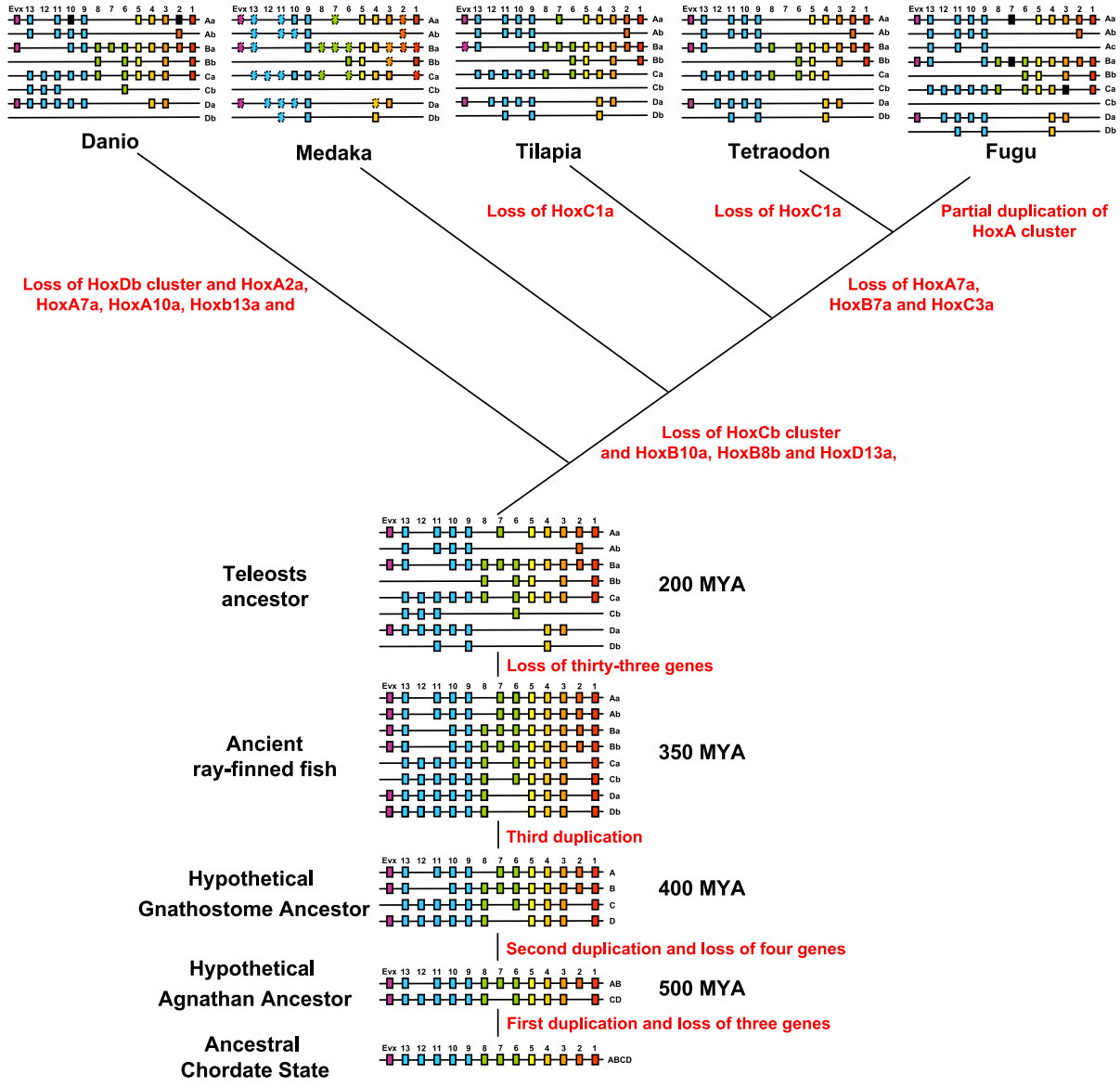

Fig. 1. Schematic of the evolution of Hox genes. The first genome duplications that occurred in metazoans 500 My led to the evolution of two clusters in the Hypothetical Agnathan Ancestor and the second genome duplication 400 My led to the evolution of four clusters in the Hypothetical Gnathostome Ancestor. From this, the Tetrapods evolved. A third genome duplication is believed to have occurred in the ray-finned fish lineage that led to a situation of eight clusters in the ray-finned fish Ancestor (350 My). The loss of some genes led to the Teleosts Ancestor (200 My), from which the modern Teleosts radiated in the last 200 My. Each square represents a gene. Black squares in Fugu and zebrafish *Hox* clusters are pseudogenes.

3.3. Hox gene set in tilapia and evolution of the Hox cluster organization in fish

Previous PCR surveys and genomic library screenings had identified an interesting variability in *Hox* genes contents among fishes (Amores et al., 1998, 2004; Aparicio et al., 2002; Naruse et al., 2004), therefore, the first goal of this work was to learn the organization of the *Hox* clusters of the Tilapia *O. niloticus*. Tilapia possesses a set of *Hox* genes more similar to the pufferfish and the medaka than to the zebrafish, according with its evolutionary position with respect to these groups (Fig. 1). Tilapia has an almost complete *Hox*Aa cluster (only genes 6, 8 and 12 are missing), and no lineage-specific gene losses relative to other teleost fishes were observed (Santini et al., 2003). The Tilapia *Hox*Aa cluster retains the *Hox* 2, 7, and 10 genes, which are absent or non-functional in the Danio *Hox*Aa cluster. As already noticed (Santini et al., 2003), the *Hox* A7a gene, functional in Tilapia, is present only as a pseudogene in Fugu and absent in Tetraodon. The *Hox*Ab cluster is strongly reduced and contains only genes 2, 9, 10, 11 and 13. Genes 8 and 12 in clusters A have been probably lost very early during vertebrate evolution, in fact they are absent in all fishes and mammals. Gene 6 underwent a specific loss in ray-finned fishes and its loss must have happened early in ray-finned evolution, because this is a character shared by all the teleosts examined so far (but it was conserved in chondrichthyes, crossopterygians and

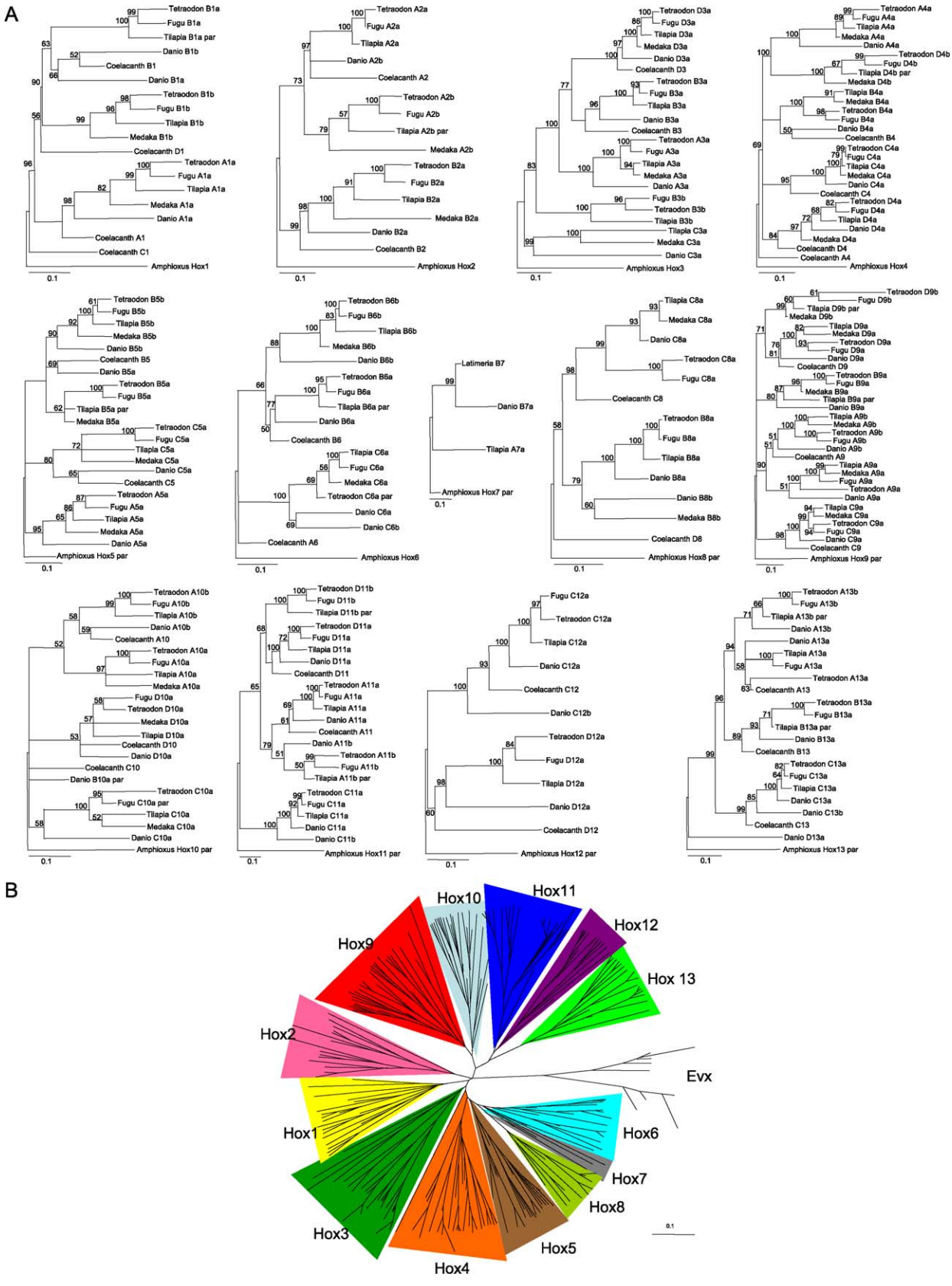

Fig. 2. (A) Neighbor-joining trees for the different Hox paralogous groups. The number at each node represents the bootstrap value recovered in 100 replicates. The Hox genes from Amphioxus, that possesses a single Hox cluster, were used as outgroup. Partial sequences are indicated by the suffix "par". Hox sequences from amphioxus, zebrafish, Fugu, Tetraodon and medaka were retrieved from the database listed in Materials and methods. (B) Unrooted tree of all Hox sequences analyzed in this work. Different colors refer to different paralogous *Hox* groups.

tetrapods). Recent studies from Koh et al. (2003) and from Powers and Amemiya (personal communication) show that gene 6 is present in coelacanth *Hox*A cluster. Unfortunately, the details of coelacanth *Hox* genes organization are not available yet: Koh et al. (2003) found genes belonging to at least four different clusters, even if the physical linkage between genes was not studied for all of them.

The *Hox* Ba cluster is again almost complete and it does not present specific losses with respect to pufferfish *Hox* Ba cluster. Genes 10, 11 and 12 are absent, as in pufferfish, while gene 10 is present in Danio genome. The Tilapia Bb cluster is strongly reduced, in terms of gene contents and it contains only genes 1, 3, 5 and 6. This condition resembles the organization of pufferfish *Hox*Bb cluster. Danio *Hox*Bb cluster lacks gene 3 but possesses gene 8, while both pufferfish (Fugu and Tetraodon) and Tilapia clusters possess gene 3 and lack gene 8. Genes 3 and 8 of cluster *Hox*Bb are not described in medaka. Unfortunately, the information concerning the organization of medaka *Hox* clusters still presents many gaps. It is therefore impossible to ascertain the degree of similarity between medaka and Tilapia *Hox* Bb clusters.

Tilapia *Hox* Ca cluster is almost complete in gene content and lacks only genes 1, 2 and 7. This condition differs from pufferfish *Hox*Ca that possesses gene 1 and for which gene 3 is a pseudogene (Amores et al., 2004). To this extent, we must point out that we could not find the *Hox* C1a gene in Fugu database, nor in Tetraodon database, therefore, we accept the information from Amores et al. (2004). The *Hox* C3a gene appears to be well conserved and possibly functional in Tilapia, as well as in Danio and possibly in medaka, for which only partial cds is available. In medaka, *Hox*C1a is not yet described. If the *Hox* C1a absence in medaka was confirmed, the organization of *Hox*Ca cluster of Tilapia would resemble medaka more than pufferfish and Danio, at least in the anterior part of the cluster. Although this can be expected for Danio, in consideration of the evolutionary distance that exists between these two fish, it would be surprising concerning the pufferfish, which are considered the closest relative of Tilapia among the species in this study (Fig. 1). Tilapia completely lacks *Hox* Cb cluster. This feature is also shared by pufferfish and medaka, while Danio possesses *Hox*Cb cluster. The Tilapia *Hox* Da cluster contains more genes than pufferfish, in fact gene 13 appears to have been conserved. In this case, the organization of *Hox*Da cluster of Tilapia would resemble Danio more than pufferfish, at least in its posterior part. This feature is quite surprising because it would have been expected that Tilapia *Hox* gene number was more similar to pufferfish, which is more closely evolutionary related to Tilapia than Danio is (Fig. 1). Nevertheless, it must be noticed that Tilapia possesses *Hox* Db cluster, which is absent in Danio but present in both the pufferfish and in medaka, although very reduced to only genes 4, 9 and 11.

In conclusion, the last common ancestor of ray-finned fish must have had a *Hox* set as in Fig. 1. Following the first duplication, the AB cluster lost the *Hox*12 gene and the CD cluster lost the *Hox*2 and *Hox*7 genes (Hypothetical Agnathan Ancestor in Fig. 1). After the second duplication, the cluster A lost the *Hox*8 gene, the cluster B lost the *Hox*11 gene, the cluster C lost the Evx gene and the cluster D lost the *Hox*6 gene (Hypothetical Gnathostome Ancestor in Fig. 1). The divergence of the tetrapods lineage followed the second duplication and the tetrapods underwent specific gene losses, while the ancient ray-finned fish ancestor underwent the third duplication and lost the genes: *Hox*6 in cluster Aa, *Hox*1, 3, 4, 5, 6, 7, Evx in cluster Ab, *Hox*2, 4, 7, 9, 10, 13, Evx in cluster Bb, *Hox*1, 3, 4, 5, 8, 9, 10 in cluster Cb, *Hox*1, 5, 8 in cluster Da and *Hox*1, 3, 5, 8, 9, 12, 13, Evx in cluster Db evolving in the teleosts ancestor (Fig. 1). The teleosts underwent further gene losses and in one case (Fugu), a further partial duplication, independently. This reconstruction is purely speculative and based uniquely on the most parsimonious hypothesis of independent gene losses.

3.4. Base composition analyses

We calculated the GC level in the *Hox* genes coding sequences (cds) in Tilapia, Fugu, Tetraodon and zebrafish (available as supplemental material at the journal website). The average GC level in the *Hox* clusters of the three fish species is within the range 48.2–61.8%, while a lower level was calculated for the zebrafish.

We compared pair of paralogous genes from twin clusters to analyze if difference in composition was present (Fig. 3). In most of the cases, genes located in the Xa (Xa=Aa or Ba or Ca or Da) cluster, the most gene-rich and therefore the most conserved cluster, had a higher GC level than their counter part in cluster Xb (Xb=Ab, Bb, Cb, Db; Table 1). Because the Xa clusters are more conserved than their counterparts Xb, it is reasonable to assume that the GC level in Xa clusters genes resembles the original situation of *Hox* genes. The cases in which *Hox* genes from the Xb cluster are GC-richer than their counterparts in the Xa clusters concerned the genes: A2, A10 and A11 in Fugu, Tetraodon and Danio, B3 in Tilapia, Fugu and Tetraodon, C13 in Danio and D4 in Tilapia, Fugu and Tetraodon. The gene A2a is a pseudogene in Danio, but in Fugu and Tetraodon is still expressed, although its expression is limited to a light striped pattern in rhombomeres 1 and 2 (r1 and r2 Amores et al., 2004).The *Hox* A2b gene is expressed in a quite extensive embryonic field in Fugu and Tetraodon, from r2 to r5 (Amores et al., 2004). Because ectopic expression of *Hox*A2 in r1 in chicken induces the development of motor neuron in this rhombomere which normally has no motor neurons, Amores et al. (2004) hypothesized that the conservation and functionality of gene *Hox*A2a in pufferfish might have induced the origin of new motor neurons that could be involved in the evolutionary invention of the

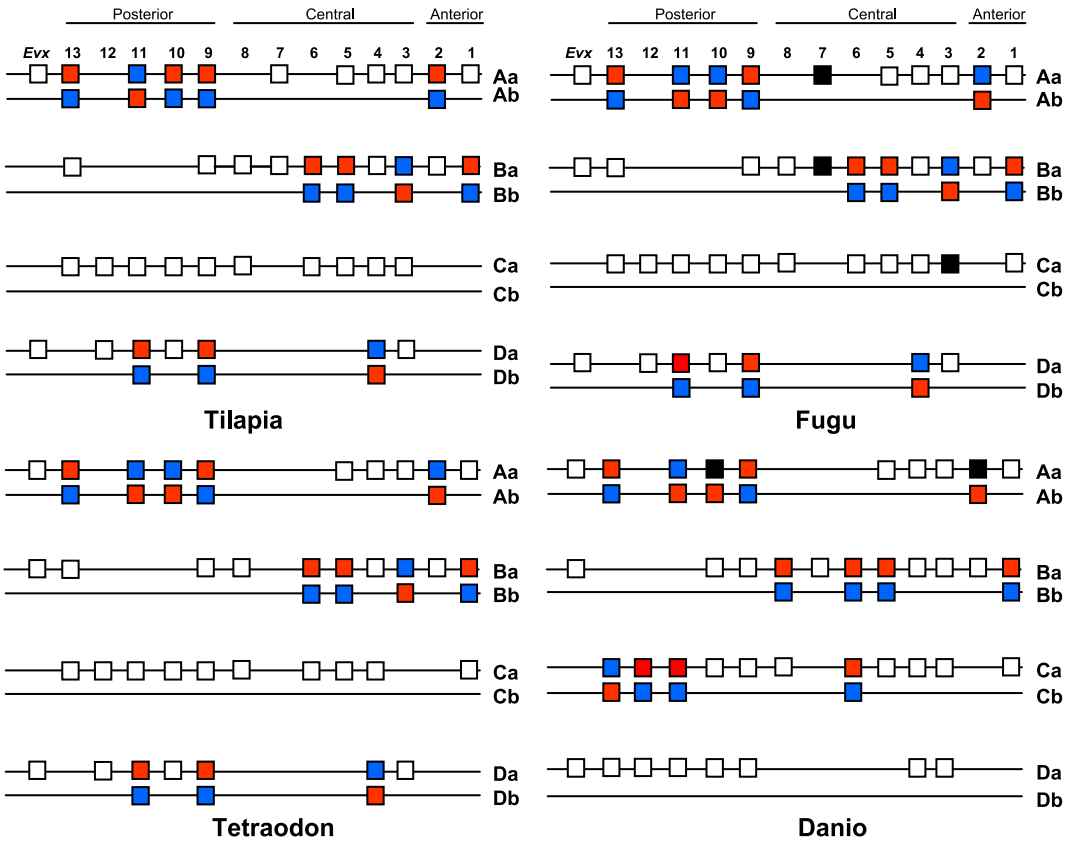

Fig. 3. Comparisons between the four pair of twin Hox cluster. Each square represents a gene. In the genes, pairs from twin clusters (fish-specifically duplicated clusters), the GC-richer gene of the pair is represented by a red square, the GC-poorer by a blue square. Pseudogenes in Fugu and zebrafish *Hox* clusters are represented by a black square.

buccal pump that permits to the pufferfish to inflate its stomach with water to increase its size in response to aggression.

Although very appealing, this hypothesis does not seem to be likely. In fact, Tilapia has a *Hox* A2a gene (89% identities to Fugu *Hox*A2a and 85% identities to Tetraodon A2a) and a *Hox* A2a promoter very well conserved (Santini et al., 2003). Therefore, there is no reason to believe that *Hox*A2a gene in Tilapia is not functional. In the case of Tilapia, the persistence of *Hox*A2a gene expression cannot be related to any evolutionary invention such as the buccal pump of pufferfish. Nevertheless, the variability of sequence composition and function among the three species of fishes under consideration is interesting. It could be hypothesized that GC level in the gene sequence is related to the loss or even to a simple reduction of the expression field of the gene in the embryo. If this was true, pufferfish would represent an intermediate situation between a full functional *Hox*A2a, as possibly Tilapia's, and a completely non-functional *Hox*A2a as Danio pseudogene is. The *Hox*A2a gene in both the pufferfish would have undergone a reduction of its GC level as of its field of embryonic expression. *Hox*A2a gene is expected to be GC-richer than *Hox*A2b gene, because it belongs to the gene-richer and better conserved of the *Hox*A clusters, therefore, any difference from this trend has to be evaluated.

There are two possible explanations for a similar phenomenon (Fig. 4): after the duplication of the ancestral *Hox*A cluster either the *Hox*A2a gene preferentially accumulated mutations in comparison to the *Hox*A2b gene (Fig. 4A) or the *Hox*A2a and A2b genes underwent a sub-functionalization event that put the two *Hox*A2 genes under different selective pressures. As a consequence of the new and different constraints, the *Hox*A2a gene reduced its field of expression in the embryo, becoming progressively less important than its counterpart *Hox*A2b, and therefore more free of accumulating mutations, until it became a pseudo-gene in Danio (4B). What we see nowadays in Fugu and Tetraodon is the intermediate situation, in which the *Hox*A2a gene has already reduced its field of expression in the embryo (Amores et al., 2004) and has already started to preferentially accumulate A–T biased mutations that reduced its GC level as shown in this study (Table 1). Finally, in Tilapia the *Hox*A2a gene still has a higher GC level than its counterpart *Hox*A2b, as expected for a gene belonging to the gene-richer, most conserved cluster. Tilapia *Hox*A2a GC level possibly resembles the original situation following the duplication of the *Hox*A cluster and the subsequent sub-functionalization of A2a and A2b genes. This hypothesis is also supported by the fact that the same condition occurs between the *Hox*A10a and A10b genes: *Hox*A10a is a pseudogene in Danio, is apparently still

Table 1
Comparison of the GC and GC3 level in twin genes pairs

Species	Cluster Xa	%GC	%GC3	Cluster Xb	%GC	%GC3	ΔGC	ΔGC3
Tilapia	A2a	57.1	68.4	A2b*	53.6	62.5	3.5	5.9
	A9a	57.9	69.1	A9b	50.5	56.0	7.4	13.1
	A10a	55.5	63.3	A10b	52.8	60.3	2.7	3.0
	A11a	60.1	67.9	A11b*	59	71.1	1.1	−3.2
	A13a	49.6	69.9	A13b*	44.4	39.5	5.2	30.4
	B1a*	63.9	78.4	B1b	53.6	65.4	10.3	13.0
	B3a*	**55.4**	59.5	**B3b**	**59**	72.8	**−3.6**	−13.3
	B5a*	62.8	90.2	B5b	53.9	71.0	8.9	19.2
	B6a*	60.2	97.5	B6b	55	79.6	5.2	17.9
	D4a*	**47.5**	57.9	**D4b**	**58.6**	78.1	**−11.1**	−20.2
	D9a*	54.2	58.3	D9b	53.5	80.3	0.7	−22.0
	D11a*	47.7	59.8	D11b	43.5	53.6	4.2	6.2
Fugu	**A2a**	**55**	61.7	**A2b**	**55.2**	65.0	**−0.2**	−3.3
	A9a	56.8	66.3	A9b	55	67.1	1.8	−0.8
	A10a	**56.6**	66.5	**A10b**	**56.8**	68.3	**−0.2**	−1.8
	A11a*	**52**	58.5	**A11b**	**58.5**	73.9	**−6.5**	−15.4
	A13a	56.2	67.9	A13b	51.7	59.9	4.5	8.0
	B1a	59	66.0	B1b	58.4	75.9	0.6	−9.9
	B3a	**58.9**	67.1	**B3b**	**60.7**	76.0	**−1.8**	−8.9
	B5a	60	70.6	B5b	57.5	69.9	2.5	0.7
	B6a	63.4	79.2	B6b	57.7	75.1	5.7	4.1
	D4a	**51.6**	61.7	**D4b**	**58.4**	71.5	**−6.8**	−9.8
	D9a	56.4	69.2	D9b	56.2	70.6	0.2	−1.4
	D11a	51.8	69.2	D11b	49.4	54.3	2.4	14.9
Tetraodon	**A2a**	**55.2**	64.0	**A2b**	**55.6**	67.1	**−0.4**	−3.1
	A9a	59.2	64.4	A9b	57.5	72.1	1.7	−7.7
	A10a	**55.5**	64.3	**A10b**	**58.5**	70.7	**−2.9**	−6.4
	A11a	**53.8**	63.6	**A11b**	**59.6**	76.1	**−5.8**	−12.5
	A13a	63.5	70.5	A13b	54	65.4	9.5	5.1
	B1a	**61.1**	73.0	**B1b**	**62.9**	84.3	**−1.8**	−11.3
	B3a	**58**	66.9	**B3b**	**62.3**	76.8	**−4.3**	−9.9
	B5a	62.2	72.9	B5b	59.3	71.5	2.9	1.4
	B6a	65.9	83.2	B6b	59.7	82.5	6.2	0.7
	D4a	**54.3**	69.0	**D4b**	**60.2**	76.3	**−5.9**	−7.3
	D9a	60.3	74.2	D9b	55.3	59.5	5.0	14.7
	D11a	56.7	74.4	D11b	49.9	56.1	6.8	18.3
Danio	**A2a ps.**	**47.4**	48.7	**A2b**	**54.3**	52.3	**−6.9**	−3.6
	A9a	49.7	51.6	A9b	48.3	49.2	1.4	2.4
	A10a ps.	**40.2**	41.1	**A10b**	**50.2**	52.4	**−10.0**	−11.3
	A11a	**48.1**	53.0	**A11b**	**49.9**	59.7	**−1.8**	−6.7
	A13a	51.7	51.3	A13b	49.6	52.6	2.1	−1.3
	B1a	52	51.6	B1b	48.3	51.1	3.7	0.5
	B5a	52.4	58.5	B5b	49	54.0	3.4	4.5
	B6a	52.3	59.2	B6b	46.7	49.6	5.6	9.6
	B8a	52.4	57.1	B8b	52.1	60.3	0.3	−3.2
	C6a	48.2	53.2	C6b	47	50.2	1.2	3.0
	C11a	48.4	54.8	C11b	47.8	55.4	0.6	−0.6
	C12a	50.1	53.7	C12b	46.8	47.7	3.3	6.0
	C13a	**48.7**	50.0	**C13b**	**51.2**	55.7	**−2.5**	−5.7
Sum							48.1	2.0

Partial sequences are indicated by an asterisk. The gene pairs in bold are those for which the gene belonging to the Xb cluster is GC-richer than the twin gene in the Xa cluster.

functional but GC-poorer than the twin *Hox*A10b gene in both the pufferfish and is apparently functional and GC-richer than A10b in Tilapia.

To our knowledge, this is the first reported evidence of an A–T bias in the pattern of base substitution in fish. In fact, all the studies concerning the pattern of base substitution have so far focused on mammals, fly, worms and yeast (Gojobori et al., 1982; Echols et al., 2002; Alvarez-Valin et al., 2002). In these studies, it was highlighted a similar A–T bias among those groups of animals. The fact that a similar

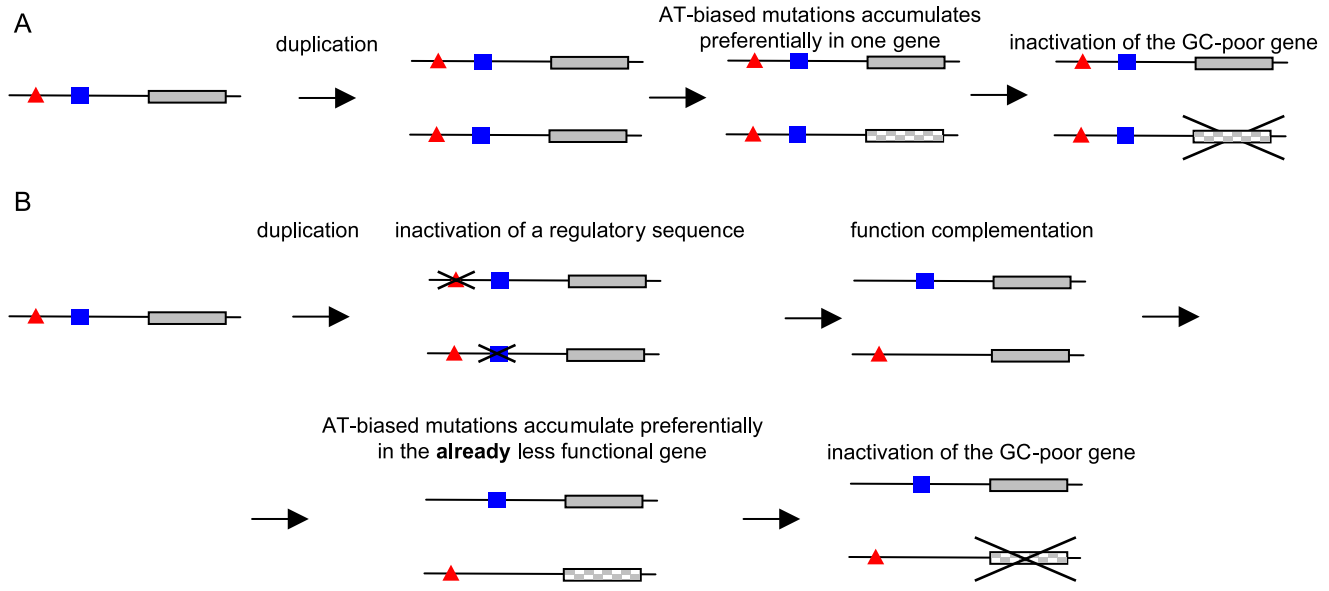

Fig. 4. Proposed models to explain the different retention of genes in duplicated Hox clusters. (A) One of the two duplicates preferentially accumulates AT-biased mutations that lead to a loss of its functionality and becomes a pseudogene. (B) The two duplicates undergo sub-functionalization that makes them complementary in recreating the ancestral parent gene functions. The two duplicates are no longer equivalent, therefore are subjected to different selective pressures. The accumulation of AT-biased mutations will be preferential in the gene that had its field of expression already reduced. The red triangle and the blue square are regulatory elements, the grey square represents the coding sequence.

bias exists also in fishes strongly argues for it to be an universal trend in mutations.

To further support this evidence, we analyzed the pattern of substitution of Fugu pseudogenes A7a, B7a and C3a in comparison with the corresponding functional orthologues in Tilapia. The pseudogene HoxC3a from Fugu does not show a GC level reduction in comparison with the corresponding functional gene from Tilapia, but it could be due to an extensive deletion the Fugu gene underwent. In fact, the similarity between Fugu HoxC3a pseudogene and Tilapia HoxC3a gene is only 40%. Both functional Tilapia A7a and B7a genes show a higher GC level than the corresponding pseudogenes. Fugu HoxA7a pseudogene shows a sequence similarity of 56% to Tilapia HoxA7a gene and Fugu HoxB7a pseudogene shows a sequence similarity of 50% to Tilapia HoxB7a gene. In the case of HoxA7a, the functional gene from Tilapia has 56.4% GC against 54.7% GC of the pseudogene from Fugu. In the case of HoxB7a, the functional gene from Tilapia has 56.4% GC, while the pseudogene from Fugu has 42.7% GC. These results strongly support the pattern of substitution delineated above, in fact show that the most mutated (respect to the functional gene from Tilapia) pseudogene Fugu HoxB7a reduced its GC level more markedly than the less mutated (respect to the functional gene from Tilapia) Fugu HoxA7a. Therefore, the pattern of mutational substitution in fish gene seems to be A–T biased.

Experiments carried out on zebrafish and pufferfish embryos confirm a wider field of expression for those genes of the twin pairs that exhibit a higher GC level. Bruce et al. (1999) showed that, in the pair of sub-functionalized genes B5a (GC=52.4%) and B5b (GC=49.0%) of zebrafish, B5a has a wider field of expression in the embryo. McClintock et al. (2001) found that, in the pair of sub-functionalized genes B1a (GC=52.0%) and B1b (GC=48.3%) of zebrafish, B1a has a later but prolonged expression in comparison to B1b, which has a precocious onset of expression that ends within the first 10 h post-fertilization. Finally, Amores et al. (2004) described a wider field of expression of D4b (GC=58.4%) in comparison to its counterpart D4a (GC=54.3%) in Fugu.

The importance of this finding is in fact that it can tentatively explain how *Hox* clusters continue to evolve in ray-finned fish. Amores et al. (2004) showed that silencing and gene loss can continue far longer than previously thought. They examined the *Hox* gene set in two species of pufferfish, *S. nephelus* and *T. rubripes*, and found that the *Hox*B7a gene is present in *S. nephelus*, but absent in *T. rubripes*. The *Hox*B7a gene must have remained intact from the time of the duplication event about 350 My (Taylor et al., 2001) until the divergence of *Spheroides* and *Takifugu* lineages, which is estimated to have occurred only 5–35 My (Santini and Tyler, 1999). Thus, the *Hox*B7a gene must have been maintained for about 300 My before being lost in the *Takifugu* lineage, but not in the *Spheroides* lineage, after their respective divergence. This is hundreds of millions years longer than the permanent preservation of gene duplicates has been thought to take. It is not at all surprising then the fact that some genes of *Hox* set of ray finned fish are still evolving in their base composition, reflecting the different selective pressure duplicated gene, is subjected to and suggested the different fate of the duplicates.

Table 1 also shows the GC3 levels for the twin genes pairs. All the GC-poorer genes of each twin pair also have a lower GC3 level than the twin gene, confirming that they are preferentially subjected to accumulate AT-biased mutations. The third position of codons is the most tolerant to mutations, therefore, it can be expected that it loses GC in a more conspicuous way than the other positions. To understand whether the differences in GC level of the pairs of twin *Hox* genes were due to different contributions from one of the two exons to the total GC level (i.e. preferential accumulation of mutations in one of the two exons), we analyzed the base composition of exon 1 and exon 2 separately (data not shown). The mutations did not favor one exon compared to the other. It appears that both the exons are subjected to the same mutational pressure, this finding supporting the theory that mutations intervene on genes that have an expression already reduced in the embryo and that are, therefore, freer to mutate in the whole of their length.

3.5. Conclusions

The present work highlighted several issues:

(i) The cichlid fish Tilapia possesses a set of *Hox* genes and clusters that resembles those of medaka and pufferfish more than that of zebrafish, which appears to be, so far, a very specific case. Further investigations will be necessary that describe the organization of *Hox* genes in species more closely related to zebrafish.

(ii) Duplicated *Hox* genes, that have been conserved for several tens of million years presumably through sub-functionalization, may undergo a late non-functionalization.

(iii) This fate of late non-functionalization seems to be preceded by a reduction of the expression field of the considered *Hox* gene in the embryo.

(iv) A *Hox* gene that has an already reduced field of expression in the embryo may become less important and therefore freer to accumulate (AT-biased) mutations than the paralogous gene sharing the original function of the ancestral unduplicated gene.

(v) A reduction of the GC level in functional *Hox* genes in comparison with the paralogous genes can be a hint to identify those genes that are on the way to non-functionalization.

(vi) These events may possibly occur in other pairs of duplicated and sub-functionalized genes, not only *Hox* genes.

Acknowledgements

The authors wish to acknowledge the Molecular Biology service in the Stazione Zoologica of Naples for the sequencing of Tilapia genes, all the members of the Molecular Evolution Laboratory for the useful and constructive criticism and the European Commission Sixth Framework Programme (ERG Grant to SS).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2004.10.027.

References

Alvarez-Valin, F., Lamollea, G., Bernardi, G., 2002. Isochores, GC3 and mutation biases in the human genome. Gene 300, 161–168.

Amores, A., et al., 1998. Zebrafish hox clusters and vertebrate genome evolution. Science 282, 1711–1714.

Amores, A., et al., 2004. Developmental roles of pufferfish Hox clusters and genome evolution in ray-finned fish. Genome Res. 14, 1–10.

Aparicio, S., 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science 297, 1301–1310.

Bernardi, G., et al., 1995. The human genome: organization and evolutionary history. Annu. Rev. Genet. 29, 445–476.

Bruce, A., Oates, A., Prince, V.E., Ho, R.K., 2001. Additional hox clusters in the zebrafish: divergent expression belies conserved activities of duplicate hoxB5 genes. Evol. Dev. 3, 127–144.

Consortium, I.H.G.S., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Consortium, M.G.S., 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562.

Consortium, R.G.S.P., 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428, 493–521.

Echols, N., et al., 2002. Comprehensive analysis of amino acid and nucleotide composition in eukariotic genomes, comparing genes and pseudogenes. Nucleic Acids Res. 30, 2515–2523.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.-L., Postlethwait, J.H., 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151, 1531–1545.

Gehring, W.J., 1993. Exploring the homeobox. Gene 135, 215–221.

Gojobori, T., Li, W.-H., Graur, D., 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. 18, 360–369.

Haldane, J.B.S., 1933. The part played by recurrent mutation in evolution. Am. Nat. 67, 5–9.

Holland, P.W.H., Garcia-Fernandez, J., 1996. Hox genes and chordate evolution. Dev. Biol. 173, 382–395.

Koh, E.G., Lam, K., Christoffels, A., Erdmann, M.V., Brenner, S., Venkatesh, B., 2003. Hox gene clusters in the Indonesian coelacanth, *Latimeria menadoensis*. Proc. Natl. Acad. Sci. U. S. A. 100, 1084–1088.

Lynch, M., Conery, J.S., 2000. The evolutionary fate and consequences of duplicate genes. Science 290, 1151–1155.

McClintock, J.M., Carlson, R., Mann, D.M., Prince, V.E., 2001. Consequences of Hox gene duplication in the vertebrates: an investigation of the zebrafish Hox paralogue group 1 genes. Development 128, 2471–2484.

McGinnis, W., Krumlauf, R., 1992. Homeobox genes and axial patterning. Cell 68, 283–302.

Misof, B.Y., Wagner, G.P., 1996. Evidence for four Hox clusters in the killifish *Fundulus heteroclitus* (teleostei). Mol. Phylogenet. Evol. 5, 309–322.

Nadeau, J.H., Sankoff, D., 1997. Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. Genetics 147, 1259–1266.

Naruse, K., Tanaka, M., Mita, K., Shima, A., Postlethwait, J.H., Mitani, H., 2004. A medaka gene map: the trace of ancestral vertebrate proto-chromosomes revealed by comparative gene mapping. Genome Res. 14, 820–828.

Nelson, J.S., 1994. Fishes of the World. Wiley-Interscience, New York.

Ohno, S., 1970. Evolution by Gene Duplication. Springer-Verlag, Berlin.

Santini, F., Tyler, J.C., 1999. A new phylogenetic hypothesis for the order Tetraodontiformes (Teleostei, Pisces), with placement of the most fossil basal lineages. Am. Zool. 39, 10A.

Santini, S., Boore, J.L., Meyer, A., 2003. Evolutionary conservation of regulatory elements in vertebrate Hox gene clusters. Genome Res., 13.

Scott, M.P., 1992. Vertebrate homeobox gene nomenclature. Cell 71, 551–553.

Swofford, D., 2000. PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). Sinauer, Sunderland, MA. ed. 4.0b10.

Taylor, J.S., van de Peer, Y., Braasch, I., Meyer, A., 2001. Comparative genomics provides evidence for an ancient genome duplication event in fish. Philos. Trans. R. Soc. Lond., B 356, 1661–1679.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25, 4876–4882.