The Shark HoxN cluster is homologous to the Human HoxD cluster

Sonja J. Prohaska[‡,†], Claudia Fried[‡,†], Chris T. Amemiya[⊕], Frank H. Ruddle[⊗], Günter P. Wagner[¶], and Peter F. Stadler[‡,†,§]

[‡]Bioinformatik, Institut für Informatik,
Universität Leipzig, Kreuzstraße 7b, D-04103 Leipzig, Germany

[†]Institut für Theoretische Chemie und Molekulare Strukturbiologie Universität Wien,
Währingerstraße 17, A-1090 Wien, Austria

[⊕]Virginia Mason Research Center, Benaroya Research Institute,
Molecular Genetics Dept. 1201 Ninth Avenue, Seattle, WA 98101 USA

[⊗]Department of Department of Cellular and Developmental Biology
Yale University, New Haven, CT, USA

[¶]Department of Ecology and Evolutionary Biology
Yale University, New Haven, CT, USA

[§]The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

[*]Address for correspondence:
Peter F. Stadler, Bioinformatik, Institut für Informatik, Universität Leipzig, Kreuzstraße 7b, D-04103 Leipzig, Germany. Phone: ++49 341 149 5120; Fax: ++49 341 149 5119; Email: peter.stadler@bioinf.uni-leipzig.de.

Abstract.

The statistical analysis of phylogenetic footprints in the two known horn shark Hox clusters and the four mammalian clusters shows that the shark *HoxN* cluster is *HoxD*-like. This finding implies that the most recent common ancestor of jawed vertebrates had at least four Hox clusters, including those which are orthologous to the four mammalian Hox clusters.

Keywords. Hox clusters, phylogenetic footprints

1. Introduction

Hox genes code for homeodomain containing transcription factors which are homologous to the genes in the Drosophila homeotic gene clusters (McGinnis and Krumlauf, 1992). Vertebrates, in contrast to all invertebrates examined, have multiple Hox gene clusters that presumably have arisen from a single ancestral cluster in the most recent common ancestor of chordates, i.e. amphioxus and vertebrates (Garcia-Fernández and Holland, 1994; Kappen *et al.*, 1989). The timing of the Hox cluster duplication events in vertebrate phylogeny is still somewhat unclear. The most popular hypothesis is that the common ancestor of sharks and bony fish (which also include the land vertebrates such as human and mouse) had four clusters homologous to the mammalian ones. (Holland and Garcia-Fernandez, 1996). To test this idea, two nearly complete Hox clusters have recently been isolated and sequenced, called *N* and *M* (Kim *et al.*, 2000). While the *M* cluster is clearly homologous to the human *HoxA* cluster, it was more difficult to assign the homology to the *HoxN* cluster. In the original description *HfHoxN* was identified as homologous to the human *HoxD* cluster, but there is also evidence consistent with homology to the *HoxC* cluster (Málaga-Trillo and Meyer, 2001).

2. Materials and Methods

In this contribution we perform a statistical analysis of conserved non-coding sequences utilizing a new software called `tracker` (Prohaska *et al.*, 2003). This program is based on `BLAST` (Altschul *et al.*, 1990) for the initial search of all pairs of input sequences. Comparisons are (optionally) restricted to homologous intergenic regions. The resulting list of pairwise sequence alignments is then assembled into groups of partially overlapping regions that are subsequently passed through several filtering steps. Individual phylogenetic footprints (PFs) are defined in Tagle's original paper (Tagle *et al.*, 1988) as blocks of at least 6bp of DNA sequence that is 100% conserved in taxa that have an additive evolutionary time of 250 million years. PFs are considered to be putative transcription factor binding sites. Typically `tracker` detects clusters of such footprints which are termed *cliques*. The decomposition of cliques into individual footprints is often ambiguous. Our statistical analysis below is therefore based on the total length of significantly homologous non-coding sequence fragments between pairs of clusters. This measure is roughly proportional to the number of individual footprints. Homologous footprints are necessarily co-linear (disregarding the possibility of local transpositions or inversions which cannot be resolved with the present analysis method due to the highly diverged sequence outside the footprint clusters). Non-colinear `tracker`-hits are therefore disregarded (marked by × in the supplemental material).

The `tracker` program produces alignments of the footprint cliques using `dialign` (Morgenstern, 1999). These are padded with "gap" characters in those sequences that do not take part in a particular clique and then concatenated. The resulting "alignment" is sparse in the sense that the "gap" character is the most frequent letter. The reconstruction of phylogenies from such a dataset has to take three complications into account: (1) gene loss will cause almost certainly the loss of all the the associated regulatory sequences. In the extreme case, presence-absence data of footprints might just reflect that presence-absence pattern of the genes. (2) We cannot expect to have detected *all* footprints in all species. (3) Gain and loss of footprints are not symmetric processes: in fact footprint loss is much easier than the *de novo* creation. These complications can be circumvented by considering only mutations within conserved non-coding regions, i.e., within the footprint cliques detected by the `tracker` program. The distance of two clusters is therefore derived from the frequency of mutations within cliques that are shared by the two clusters. Technically, this amounts to treating "gaps" as missing data rather than as an additional character state.

We use different distance-based and parsimony-type approaches here: Neighbor joining method (Saitou and Nei, 1987) (implemented in the `phylip` package, version 3.6) (Felsenstein, 1989), the canonical split decomposition (Bandelt and Dress, 1992), Buneman trees (Buneman, 1971), parsimony splits and P-trees (Bandelt and Dress, 1993). With the exception of NJ these methods are implemented in the `splitstree` package (version 3.1) (Huson, 1998). The split-based methods are particularly suitable for our purposes because they are known to be very conservative in that they tend to produce multifurcations rather than poorly supported edges (Semple and Steel, 2003). In addition we use `MacClade 4.0` (Maddison and Maddison, 2000) for standard maximum parsimony analysis.

The following sequences are used for the analysis: Shark (*Heterodontus francisci*): M-cluster HfM = AF479755, N-cluster HfN = AF224263; Human (*Homo sapies*): HsA = AC004080.2rc + AC010990 [201-6508]rc + AC004079 [75001-end]rc, HsB = NT_010783 [931646-1263780]rc, HsC = NT_009563 [580371-708054]rc, HsD = NT_037537 [4075338-end]. Rat (*Rattus norvegicus*): RnA = NW_043751 [910030-1194462]rc, RnB = NW_042671 [264022-581839], RnC = NW_044048 [722873-1060956] RnD = NW_042732 [1061702-1217610]rc. Fugu (*Takifugu rubripes*) sequences are taked from the Fugu database DOE Joint Genome Institute: TrAa = scaffold_47 of release_3.0, TrAb = scaffold 1874 of release 2.0, TrD = scaffold_3959+scaffold_214[160440-end]rc. Here "rc" means that the reverse complement of the database entry has been used (after extracting the indicated interval).

3. Results

A comparison of the protein sequences of the shark *HoxN* cluster with mammalian Hox protein sequences is consistent with *D*-likeness, although the data in Table 1 do not show an unambiguous picture. In particular, the *HoxD* proteins are not always the ones with the highest degree of sequence identity, see Table 1. In a similar vein, the analysis of Hox genes and of genes linked to the Hox clusters such as collagens does not yield an unambiguous picture for the branching order of the four mammalian Hox clusters (Bailey *et al.*, 1997).

Let us now turn to the analysis of the conserved parts of the non-coding sequences. Table 2 summarizes the results of pairwise comparisons of shark and human (or rat) Hox clusters. It should be noted that the sequence of the shark *HoxN* cluster is incomplete, spanning only the sequence from *evx* to (almost) *Hox-4*. There is a particularly high conservation of non-coding sequences between shark *HoxM* and

Table 1. Best correspondences of Hox proteins with the *HoxN* sequence of the hornshark. Numbers are percentages of sequence identities in protein alignments obtained with clustalw (Thompson *et al.*, 1994). *Italics* and sans serif fonts indicate that the best match is by the *human* or rat sequence, respectively. Empty fields indicate that there was no good match, a dash — denotes genes that do not exist in the mammalian Hox clusters.

Cluster	evx	13	12	11	10	9	8	5
A		70	—	*57*		63	—	*53*
B	—		—	—	—		68	*48*
C	—		*48*	54		63	*69*	44
D	*81*	68	*48*	*57*	*69*	*61*	*71*	—

mammalian *HoxA* sequences in the range from *Hox-4* to *Hox-1*. As a consequence, the counts for *HoxN* are significantly smaller. In table 2 we therefore display the data for both the full length clusters and the restriction to the region between *evx* and *Hox-4*. The total length of sequences conserved between shark and mammalian clusters in this region is comparable between *HfM* and *HfN*.

The homology of the shark *HoxM* and the mammalian *HoxA* clusters is obvious from these data. For the *HoxN* sequence we find little distinction when counting colinear cliques and only a moderate signal in the numbers of co-linear clusters. The total length of the conserved regions, however, is more than twice as large with *HoxD* than with *HoxC* and about 50% longer in *HoxD* compared to *HoxA*. The location and distribution of the footprint cliques, Fig. 1 also strongly argues for a homology with *HoxD* rather than *HoxC*.

Table 2. Pairwise comparison of non-coding sequences in the shark Hox clusters with mammalian Hox clusters. In addition we report the comparison with preliminary *HoxC* and *HoxD* cluster sequences (obtained from version 3.0 of the Fugu database (DOE Joint Genome Institute; Aparicio *et al.*, 2002); see (Prohaska *et al.*, 2003) for details). For the duplicated *HoxA* and *HoxB* clusters we list the number and lengths of cliques that hornshark shares with at least one of duplicates.

		Shark *HoxM*				Shark *HoxN*			
	.	*HoxA*	*HoxB*	*HoxC*	*HoxD*	*HoxA*	*HoxB*	*HoxC*	*HoxD*
		evx to *hox-4* only							
Cliques	Homo	50	30	15	8	16	20	20	25
	Rattus	54	19	13	7	17	19	21	24
	Fugu	36	14	14	5	19	15	9	17
Length	Homo	2932	1554	736	475	958	851	858	1872
	Rattus	3781	1008	669	537	1031	815	970	1465
	Fugu	2415	782	716	227	891	609	370	1000
		Complete cluster							
Cliques	Homo	95	35	17	16				
	Rattus	95	25	17	15				
Length	Homo	7369	1995	791	859				
	Rattus	7077	1525	827	868				

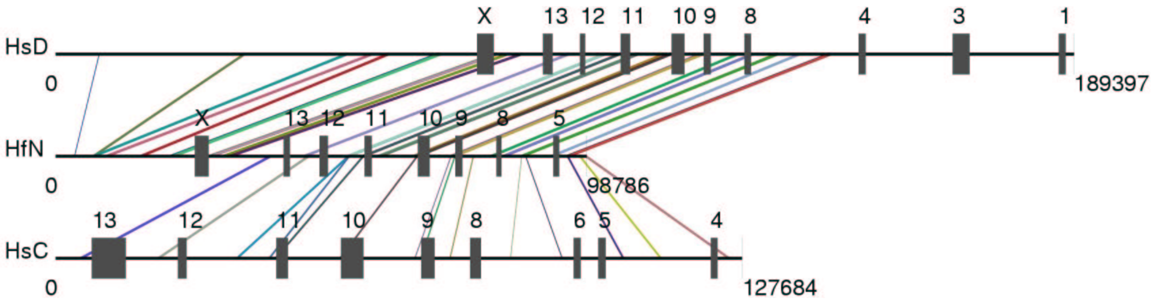

Figure 1. Overview of the phylogenetic footprint cliques produced by `tracker` for the comparison of the horn shark *HoxN* sequence (HfN) and the human *HoxC* (HsC) and *HoxD* (HsD) sequences, respectively. X denotes the *Evx* genes.

A comparison of *HfHoxN* with the fugu (*Takifugu rubripes*) *HoxCα* and *HoxD* sequences also places *HfHoxN* with the *D* rather than *C* cluster. These data must be interpreted with caution: (i) The Fugu sequences are preliminary constructs combining two or three scaffolds and hence not complete. (ii) Even though the current version 3.0 of the Fugu genome database (DOE Joint Genome Institute) does not contain evidence of a *Cβ* cluster, it is most likely that the teleost *C* cluster was duplicated since the zebrafish (*Danio rerio*) does have both a *HoxCα* and a *HoxCβ* cluster (Amores *et al.*, 1998). The duplication event might have caused the additional loss of a substantial number of footprints. Nevertheless, we find that the counts for the hornshark-pufferfish comparisons are similar to the shark-mammal comparisons.

The sensitivity of the `tracker` method is increased by including more sequences. In particular, homologous footprints can be identified between two sequences even if they do not yield a significant signal when the two sequences are compared directly. We have therefore performed a complete analysis of both shark clusters and all four human Hox clusters. The supplemental material lists all footprint cliques in the range from *evx* to *hox-1* that appear in at least one shark and at least one human cluster. The statistics of the conserved regions between clusters is summarized in Table 3.

Treating phylogenetic footprint cliques as presence/absence characters in a parsimony framework also supports the hypothesis that *HfHoxN* is more closely related to *HsHoxD* than to *HsHoxC*. The tree ((A,M),(C,(D,N)) is seven steps shorter than ((A,M),((C,N),D) (tree length = 402, $CI = 0.57$, $RI = 0.18$). This result is based on the assumption of a ((A,B),(C,D)) scenario, which is favored from the analysis of Hox sequences, see for instance (Amores *et al.*, 1998). The alternative, which is supported by the analysis of genes linked to the Hox clusters by Bailey *et al.* (1997), leads to considerably shorter trees. The tree (((A,M),(D,N)),(C,B)) has 374 steps ($CI = 0.62$, $RI = 0.43$) compared to (((A,M),D),((C,N),B)), which is 44 steps longer. While we do not want to get into the question of which cluster phylogeny applies to the human Hox clusters here, we just want to note that in either scenario a tree with *HfHoxN* most closely related to *HsHoxD* is more parsimonious than any other phylogenetic position of *HfHoxN* (data not shown).

These data clearly indicate that the shark *HoxN* cluster is *HoxD*-like at least as far as the non-coding sequences are concerned. In fact, based on total size of the footprints

Table 3. Comparison of phylogenetic footprints from a `tracker` run of both shark and all four human clusters. Only co-linear cliques in range between *evx* and *hox-1* are counted. The data contain six cliques (484, 485, 486, 513, 514, 515 in the supplement) of which at most three are consistent with co-linearity. These are counted with a weight 1/2.

	Shark *HoxM*				Shark *HoxN*			
	HoxA	*HoxB*	*HoxC*	*HoxD*	*HoxA*	*HoxB*	*HoxC*	*HoxD*
	evx to *hox-4*							
Cliques Homo	**47**	21	13	9	15	10	20	**25**
Length Homo	**3847**	1905	646	1065	1728	961	1148	**1995**
	Complete cluster							
Cliques Homo	**79**	24	13	16				
Length Homo	**6937**	2268	1142	1598				

that are shared between clusters, the next candidate would be the mammalian *A*-cluster, not the *C*-cluster as proposed in (Málaga-Trillo and Meyer, 2001).

To test whether *HfHoxN* is a true homologue of the mammalian *HoxD* clusters we consider the co-occurrences of the 49 footprint cliques that are present in *HfHoxN*, Table 4. In particular, there are 14 cliques that *HfHoxN* shares uniquely with human *HsHoxD*, compared to 10 cliques shared with the *HoxC* cluster and only 5 to 6 that are only shared with the *HoxA* and *HoxB* clusters, respectively. On the other hand, about 30% of the footprints are shared between *HfHoxD*, one of the Human clusters and at least one other mammalian cluster. The footprints shared between *HoxN* and either *HoxA* or *HoxB* are to 90% also shared with *HoxM*. Together, these data strongly suggest that *HfHoxN* is not only most similar to the mammalian *HoxD* clusters but is a true homologue.

Table 4. Footprints shared between shark *HoxN*, one of the four human clusters, and other Hox clusters.
There are 49 footprint cliques with a total length of 3042nt involving *HfHoxN*. We obtain two different counts for the number of clusters exclusively shared between *HfHoxN* and *HsHoxB* arising from the mutually incompatible cliques marked with ♣ in the supplement.

	shared with the human cluster			
	HoxA	*HoxB*	*HoxC*	*HoxD*
exclusively				
cliques	5	5-6	10	14
length	234	∼200	267	640
precentage	8	7	9	21
plus at least one other mammalian cluster				
cliques	9	5	10	11
length	1166	741	850	1051
percentage	38	24	28	35
plus *HfM*				
cliques	7	4	5	5
length	149	38	248	359
percentage	87	95	71	66

Further evidence for this claim can be obtained from the phylogenetic analysis of the combined footprint cliques of the four mammalian clusters for either human or rat together with the two available shark sequences. Both distance-based, Fig. 2 and parsimony-based methods, Fig. 3, agree on this interpretation. We have chosen a variety of split-based algorithms for this analysis because these techniques are known to produce multifurcations rather than poorly supported edges. For comparison standard neighbor-joining trees are shown in Fig. 2.

All data presented in Figs. 2 and 3 either support the conclusion that the shark *HoxN* cluster is homologous with mammalian *HoxD* cluster or are at least consistent with this conclusion (whenever the *HfHoxN*-*HoxD* node is a multifurcation).

4. Discussion

The evidence presented in this paper supports the original hypothesis, namely that the shark HoxN cluster is orthologous to the mammalian *HoxD* cluster (Kim *et al.*, 2000). The method employed is novel, namely to use the distribution and extent of non-coding sequences for phylogenetic inferences. Below we discuss the implications of the present finding for our understanding of Hox cluster evolution in vertebrates.

Conserved non-coding sequences have long been used to find candidate cis-regulatory elements, see (Duret and Bucher, 1997) for a review. Identification of putative cis-regulatory sequences requires long stretches of sequence from distantly related species (Tagle *et al.*, 1988) or a set of species which have sufficient additive divergence among them (Sumiyama *et al.*, 2001). More recently this method has been used to trace the non-coding sequence divergence after *HoxA* cluster duplication in teleosts (Chiu *et al.*, 2002). In this paper it has been shown that non-coding sequences can remain highly conserved in the absence of Hox gene cluster duplication, as documented between the shark *HoxM* and the mammalian *HoxA* cluster (see also this paper). Hence it is possible to treat the loss and the acquisition of conserved non-coding sequences as potentially apomorphic characters. Thus they contain phylogenetic information. The congruence between the structural and coding sequence evidence and the comparison on non-coding sequence conservation for *HoxM* and *HoxA* cluster validates this assumption. In the case of the shark *HoxN* cluster the evidence from coding sequence and structural organization is less strong and we thus rely on the evidence from non-coding sequence conservation. While the signal is still not as strong as for the HoxM each analysis is at least consistent and in many cases positively supportive of orthology between shark *HoxN* and mammalian HoxD cluster.

The conclusion that both the shark *HoxM* as well as the *HoxN* clusters are directly orthologous to the mammalian *HoxA* and *HoxD* clusters, respectively, has important implications for the history of Hox cluster duplications. It follows that the most recent common ancestor of cartilaginous fish and the bony fish clade (which includes mammals) had at least four Hox clusters orthologous to the four mammalian Hox clusters. It is thus likely that sharks have two more clusters than those currently described. This evidence also confirms the hypothesis of Peter Holland that the four cluster situation typical for most major gnathostome lineages has arisen before the most recent common ancestor of all recent gnathostomes (Garcia-Fernández and

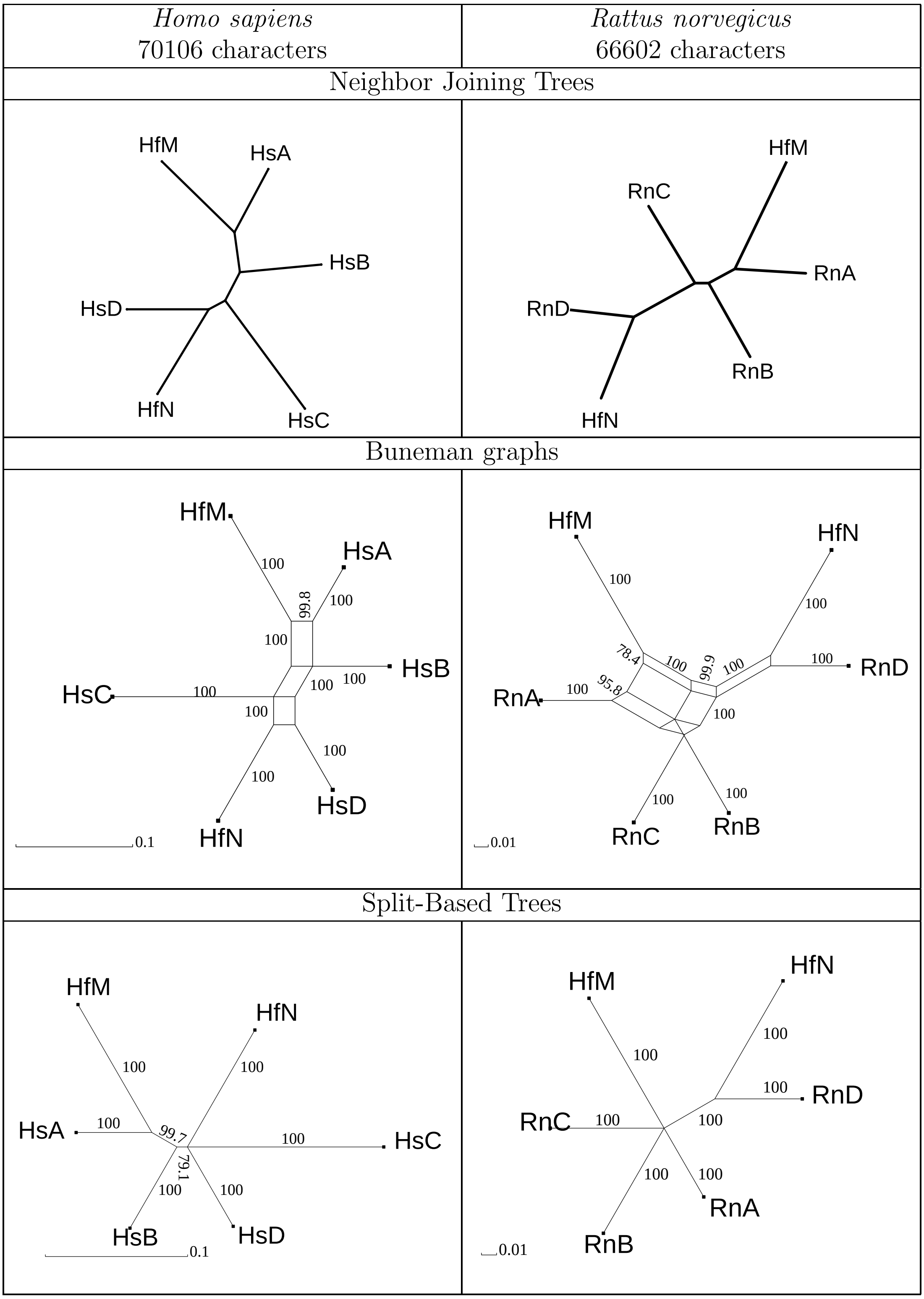

Figure 2. Distance-based phylogenies of shark and mammalian Hox clusters. Neighbor joining trees (Saitou and Nei, 1987) are computed using Felsenstein's `phylip` package (version 3.6). Buneman graphs representing the canonical decomposition of the distance function and the split-based Buneman trees are computed using Daniel Huson's `splitstree` package, version 3.1, (Huson, 1998).

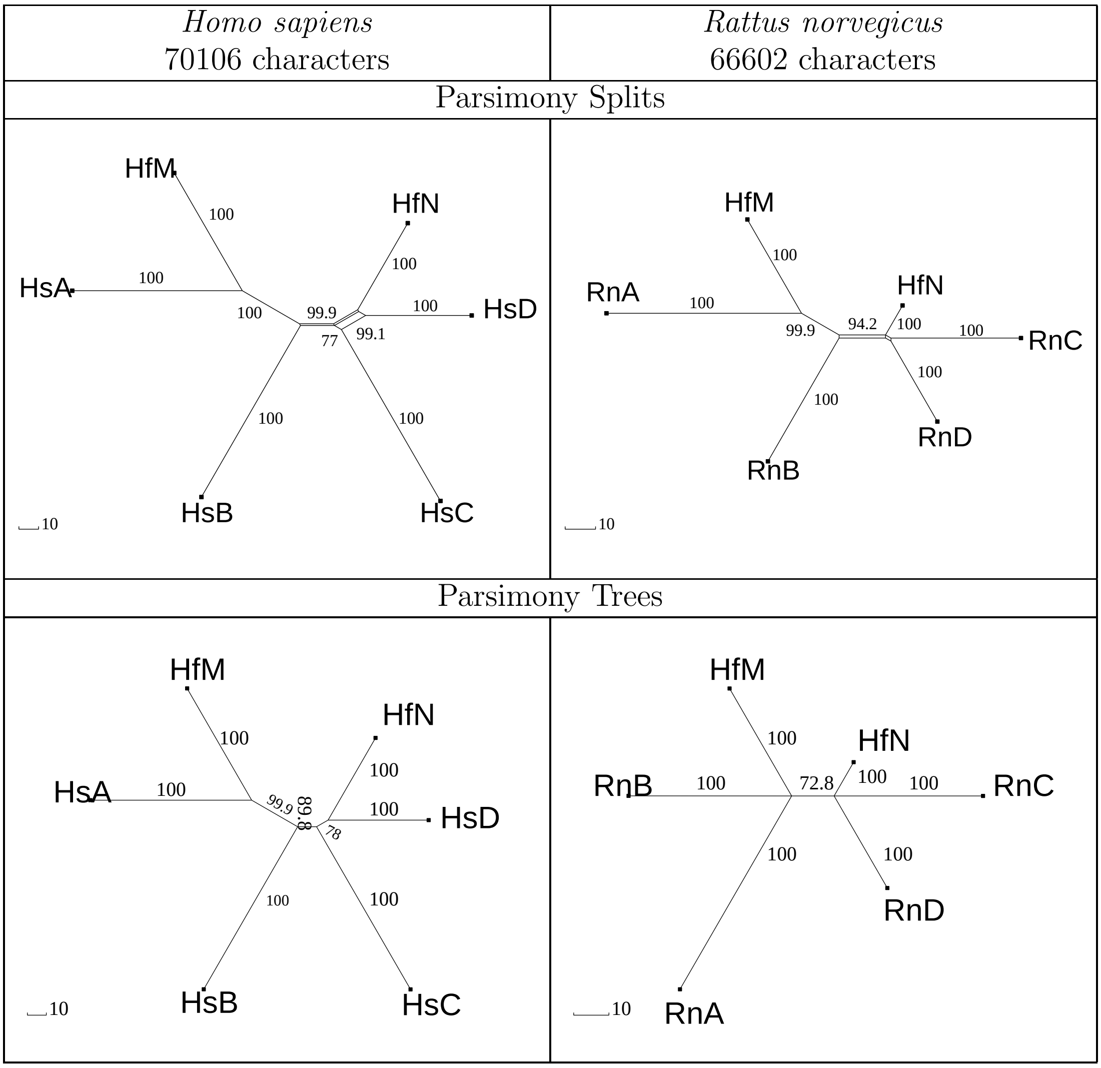

Figure 3. Parsimony-based phylogenies of shark and mammalian Hox clusters computed using `splitstree`, version 3.1, (Huson, 1998).

Holland, 1994; Holland *et al.*, 1994). Of course this result does not guarantee that all gnathostome lineages in fact have at least four Hox clusters since clusters can be lost. This can happen in particular soon after the duplication, which might have occurred shortly before the split between the shark and mammalian lineages.

Acknowledgments. Funding for this research is gratefully acknowledged: DFG Bioinformatics Initiative BIZ-6/1-2 to SJP, CF, and PFS; NSF IBN-9905403 to FR and GPW;

References

Altschul S.F., Gish W., Miller W., Myers E.W., and Lipman D.J., 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**:403–410.

Amores A., Force A., Yan Y.L., Joly L., Amemiya C., Fritz A., Ho R.K., Langeland J., Prince V., Wang Y.L., *et al.*, 1998. Zebrafish hox clusters and vertebrate genome evolution. *Science* **282**:1711–1714.

Aparicio S., Chapman J., Stupka E., Putnam N., Chia J.M., Dehal P., Christoffels A., Rash S., Hoon S., Smit A., *et al.*, 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**:1301–1310.

Bailey W.J., Kim J., Wagner G., and Ruddle F.H., 1997. Phylogenetic reconstruction of vertebrate Hox cluster duplications. *Mol. Biol. Evol.* **14**:843–853.

Bandelt H.J. and Dress A.W.M., 1992. A canonical decomposition theory for metrics on a finite set. *Adv. math.* **92**:47.

Bandelt H.J. and Dress A.W.M., 1993. A relational approach to split decomposition. In *Information and Classification*, editors Opitz O., Lausen B., and Klar R., pages 123–131. Springer-Verlag, Berlin.

Buneman P., 1971. The recovery of trees from measures of dissimilarity. In *Mathematics and the Archeological and Historical Sciences*, editors Hodson F.R., Kendall D.G., and Tautu P., pages 387–395. Edinburgh University Press, Edinburgh, UK.

Chiu C.h., Amemiya C., Dewar K., Kim C.B., Ruddle F.H., and Wagner G.P., 2002. Molecular evolution of the HoxA cluster in the three major gnathostome lineages. *Proc. Natl. Acad. Sci. USA* **99**:5492–5497.

DOE Joint Genome Institute, 2002. Fugu genome database.
version 2.0: `http://genome.jgi-psf.org/fugu3/fugu3.home.html`,
version 3.0: `http://genome.jgi-psf.org/fugu6/fugu6.home.html`.

Duret L. and Bucher P., 1997. Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.* **7**:399–406.

Felsenstein J., 1989. Phylip – phylogeny inference package (version 3.2). *Cladistics* **5**:164–166.

Garcia-Fernández J. and Holland P.W., 1994. Archetypal organization of the amphioxus hox gene cluster. *Nature* **370**:563–566.

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

Holland P.W.H., Garcia-Fernández J., Williams N.A., and Sidow A., 1994. Gene duplication and the origins of vertebrate development. *Development* **(Suppl.)**:125–133.

Huson D.H., 1998. Splitstree: analyzing and visualizing evolutionary data. *Bioinformatics* **14**:68–73.

Kappen C., Schughart K., and Ruddle F.H., 1989. Two steps in the evolution of antennapedia-class vertebrate homeobox genes. *Proc. Natl. Acad. Sci. USA* **86**:5459–5463.

Kim C.B., Amemiya C., Bailey W., Kawasaki K., Mezey J., Miller W., Minosima S., Shimizu N., P. W.G., and Ruddle F., 2000. Hox cluster genomics in the horn shark, *heterodontus francisci*. *Proc. Natl. Acad. Sci. USA* **97**:1655–1660.

Maddison D.R. and Maddison W.P., 2000. *MacClade 4: Analysis of Phylogeny and Character Evolution*. Sinauer Associates, Sunderland, Massachusetts. E-book and computer program.

Málaga-Trillo E. and Meyer A., 2001. Genome duplications and accelerated evolution of *Hox* genes and cluster architecture in teleost fishes. *Amer. Zool.* **41**:676–686.

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

Morgenstern B., 1999. `DIALIGN 2`: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* **15**:211–218.

Prohaska S., Fried C., Flamm C., Wagner G., and Stadler P.F., 2003. Surveying phylogenetic footprints in large gene clusters: Applications to Hox cluster duplications. *Mol. Phyl. Evol.* Submitted; SFI preprint #03-02-011.

Saitou N. and Nei M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol. Evol.* **4**:406–425.

Semple C. and Steel M., 2003. *Phylogenetics.* Oxford University Press, Oxford UK.

Sumiyama K., Kim C., and Ruddle F.H., 2001. An efficient cis-element discovery method using multiple sequence comparisons based on evolutionary relationships. *Genomics* **71**:260–262.

Tagle D.A., Koop B.F., Goodman M., Slightom J.L., Hess D.L., and Jones R.T., 1988. Embryonic epsilon and gamma globin genes of a prosimian primate (galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.* **203**:439–455.

Thompson J.D., Higgs D.G., and Gibson T.J., 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties, and weight matrix choice. *Nucl. Acids Res.* **22**:4673–4680.