Conservation and Diversification of Msx Protein in Metazoan Evolution

Hirokazu Takahashi, Akiko Kamiya,* Akira Ishiguro,* Atsushi C. Suzuki,† Naruya Saitou,‡ Atsushi Toyoda,§ and Jun Aruga**

*Laboratory for Comparative Neurogenesis, RIKEN Brain Science Institute, Wako, Saitama 351-0198, Japan; †Department of Biology, Keio University, Hiyoshi, Yokohama 223-8521, Japan; ‡Division of Population Genetics, National Institute of Genetics, Mishima 4511-8540, Japan; and §Sequence Technology Team, RIKEN Genomic Sciences Center, Yokohama 230-0045, Japan

Msx (/msh) family genes encode homeodomain (HD) proteins that control ontogeny in many animal species. We compared the structures of Msx genes from a wide range of Metazoa (Porifera, Cnidaria, Nematoda, Arthropoda, Tardigrada, Platyhelminthes, Mollusca, Brachiopoda, Annelida, Echiura, Echinodermata, Hemichordata, and Chordata) to gain an understanding of the role of these genes in phylogeny. Exon–intron boundary analysis suggested that the position of the intron located N-terminally to the HDs was widely conserved in all the genes examined, including those of cnidarians. Amino acid (aa) sequence comparison revealed 3 new evolutionarily conserved domains, as well as very strong conservation of the HDs. Two of the three domains were associated with Groucho-like protein binding in both a vertebrate and a cnidarian Msx homolog, suggesting that the interaction between Groucho-like proteins and Msx proteins was established in eumetazoan ancestors. Pairwise comparison among the collected HDs and their C-flanking aa sequences revealed that the degree of sequence conservation varied depending on the animal taxa from which the sequences were derived. Highly conserved Msx genes were identified in the Vertebrata, Cephalochordata, Hemichordata, Echinodermata, Mollusca, Brachiopoda, and Anthozoa. The wide distribution of the conserved sequences in the animal phylogenetic tree suggested that metazoan ancestors had already acquired a set of conserved domains of the current Msx family genes. Interestingly, although strongly conserved sequences were recovered from the Vertebrata, Cephalochordata, and Anthozoa, the sequences from the Urochordata and Hydrozoa showed weak conservation. Because the Vertebrata–Cephalochordata–Urochordata and Anthozoa–Hydrozoa represent sister groups in the Chordata and Cnidaria, respectively, Msx sequence diversification may have occurred differentially in the course of evolution. We speculate that selective loss of the conserved domains in Msx family proteins contributed to the diversification of animal body organization.

Introduction

Msx family proteins are critical regulators of metazoan ontogeny, as has been revealed in many studies (reviewed in Davidson 1995; Bendall and Abate-Shen 2000; Ramos and Robert 2005). In the Ecdysozoa, *Drosophila muscle segment homeobox* (*msh*) has a role in the regional specification of neuroectoderm and muscle progenitors (Isshiki et al. 1997; Nose et al. 1998), and *Caenorhabditis Vab-15* has a role in touch receptor neuron development (Du and Chalfie 2001). (Here, we use the term "Msx family" to denote a gene family orthologous to those encoding vertebrate Msx, including msh and vab-15.) In the Echinodermata, *Strongylocentrotus* and *Heliocidaris* Msx homologs may control gastrulation and skeletal patterning (Tan et al. 1998; Wilson et al. 2005). In the Vertebrata, *Msx1* or *Msx2* plays roles in dorsoventral patterning of embryos, regionalization of the neural tube, establishment of neural crest tissue, and the genesis of various types of organs, including the palate, tooth, skull, middle ear, jaw, hair, mammary gland, and limbs, both in mammals (Satokata and Maas 1994; Bendall and Abate-Shen 2000; Satokata et al. 2000; Lallemand et al. 2005; Ramos and Robert 2005) and amphibians (Suzuki et al. 1997; Monsoro-Burq et al. 2005; Khadka et al. 2006). Mutations in human *MSX1* are responsible for selective tooth agenesis (STA) (Vastardis et al. 1996), cleft palate, cleft lip (van den Boogaard et al. 2000), and nail dysplasia (Jumlongras et al. 2001), and those in *MSX2* are responsible for parietal foramina (PFM) (Wuyts et al. 2000) and craniosynostosis (CSO) (Wilkie et al. 2000). These facts enlighten us about not only the importance of Msx family genes in each animal but also the similarities and differences of their roles in different animal groups.

In phylogenetic terms, the structural features and/or expression profiles of Msx homologs have been described in insects (Walldorf et al. 1989), ascidians (Holland 1991), sponges (Seimiya et al. 1994), and leeches (Master et al. 1996). Some researchers have noticed that the role of the Msx family in neuroectodermal patterning is similar in the fruit fly and vertebrates (Isshiki et al. 1997; Arendt and Nubler-Jung 1999); Msx family genes specify lateral longitudinal columns of neuroectoderm in both types of animals, raising the possibility that the bilaterian ancestors had already used Msx family genes in establishing their nervous systems. The same expression pattern is essentially conserved in another insect, *Tribolium castaneum* (Wheeler et al. 2005). However, this hypothesis is still uncertain and awaits verification by examination in other animal species.

In a previous study, we performed a molecular phylogenetic analysis of wide-ranging groups of animals to determine the role of Zic family zinc finger proteins in evolution. We compared both the amino acid (aa) sequence and the exon–intron organization of many interspecies orthologs from major metazoan phyla (Aruga et al. 2006). The study revealed novel evolutionarily conserved domains and gave us a broad understanding of the processes of protein evolution and the traits involved in evolutionary change. We therefore applied the same strategy to Msx family proteins. Some earlier works had pointed out a promising direction for an analysis of the molecular phylogeny of Msx family genes. A pioneering study by Holland (1991) compared vertebrate, ascidian (Ciona), and Drosophila Msx homologs. The relationship between the Msx genes of

Key words: Msx, homeodomain, protein conserved domain, Groucho, protein–protein interaction, exon–intron boundary.

E-mail: jaruga@brain.riken.jp.

Mol. Biol. Evol. 25(1):69–82. 2008
doi:10.1093/molbev/msm228
Advance Access publication October 16, 2007

© 2007 The Authors.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

zebrafish and other vertebrates were investigated in terms of aa sequences and expression patterns (Ekker et al. 1997) and chromosomal synteny (Postlethwait 2006). Perry et al. (2006) focused on phylogenetic comparisons among primate Msx genes. However, these studies are not sufficient for our purpose because they only dealt with the limited evolutionary processes.

We report here the molecular phylogeny of Msx family genes, as determined by a comparison of the Msx sequences from 13 animal phyla. We compared the conserved domains of Msx proteins and the exon–intron organization of Msx genes, and we performed a functional analysis of some of the conserved domains. The results indicated that the eumetazoan ancestor already possessed Msx protein with a full set of conserved domains; these domains have diverged strongly in some animal groups.

Materials and Methods

Database Search

The similarity search against the current databases was done with the Blast algorithm (BlastP, TBlastN, Altschul et al. 1990, 1997) against public databases (http://www.ncbi.nlm.nih.gov/blast/, http://www.ensembl.org/index.html). Presence or absence was estimated by a database search with TBlastN algorithm against selected genome sequence databases that used key *Mus musculus* Msx-1 or *Nematostella* Msx sequences. A decision on authentic homology was based on reverse Blast analysis in which the sequence that showed the highest score in a certain species was used as a key sequence for the homology search by using the BlastX algorithm against National Center for Biotechnology Information (NCBI) eukaryotic databases. If the sequence that showed the highest Blast score among the identified sequences contained the Msx proteins, this labeled a member of Msx family (table 1).

The 18S ribosomal RNA sequences were collected from the NCBI database (accession number, supplementary text, see Supplementary Material online). These sequences were chosen because they were derived from animal species that were identical, or closely related, to those used in the Msx phylogenetic analysis (table 1).

The search for Eh-like sequences was done as follows. A local aa sequence database was constructed by DNASpace (Hitachi Software Engineering, Tokyo, Japan). The database was subjected to a homology search by using the Smith–Waterman algorithm (Smith and Waterman 1981) and the consensus Eh sequences FSV[DE] × [IL][IL] as key sequences. Sequences that gave scores of more than 15 under set parameters (i.e., Matrix, BLOSUM62; initial gap penalty, −5; extension gap penalty, −2) were listed as candidate sequences for the alignment shown in table 2.

Animals

Genomic resources for *Scolionema suvaense*, *Tubifex tubifex*, *Loligo bleekeri*, *Octopus ocellatus*, *Corbicula fluminea*, *Pandinus imperator*, *Artemia franciscana*, and *Asterina pectinifera* have been described (Aruga et al. 2006). The maintenance and recovery of *Milnesium tardigradum* species have been described (Suzuki 2003). *Limulus polyphemus*, *Lingula anatina*, *Urechis caupo*, and *Anemonia erythraea* were purchased from local vendors. *Thysanozoon* sp. was collected from the seashore in Kanagawa Prefecture, Japan. Its picture can be seen at http://lcn.brain.riken.jp/photos.html. Species identification was done by sequencing 18S ribosomal RNA genes (data not shown).

Polymerase Chain Reaction Cloning of Msx cDNA

RNAs were isolated by using TRIzol reagent (Invitrogen, CA) in accordance with the manufacturer's recommendation. cDNAs were generated by using a 3′-Full RACE Core Set (Takara Bio, Shiga, Japan). The homologs were initially identified by nested polymerase chain reaction (PCR) on cDNA or genomic templates. The following primers were used for PCR amplification of the homeodomain (HD) region of Msx family genes. The first PCR was carried out by using MshDF-1, 5′-YTIMGIAARCAYAARACIAAYMG-3,′ and MshDR-1, 5′-TTIGCICKRTTYTGRAACCA-3′, and the second PCR was done by using MshDF-2, 5′-AAYMGIAARCCIMSIACICCITT-3′, and MshDR-2, 5′-CCAIATYTTIAYYTGIGTYTC-3′. Each PCR consisted of 35 cycles of 94 °C for 1 min, 38 °C for 1 min, and 72 °C for 2 min. The PCR was performed with ExTaq DNA polymerase (Takara Bio) in the presence of BD TaqStart anti-Taq antibody (BD Biosciences, San Diego, CA). cDNAs corresponding to HD at their 3′ ends were cloned by using a 3′-Full RACE Core Set (Takara Bio). The aa sequences were deduced from nucleotide sequencing of multiple PCR fragments.

Cloning of Genomic DNA

Isolation of high molecular weight DNA, fosmid library construction, and library screening were done as described (Aruga et al. 2006). Fosmid genomic libraries were prepared with CopyControl pCC1FOS vector (Epicentre, WI). In this study, we isolated 8 fosmid clones from *A. pectinifera* [Ast-1 (36,228 bp), Ast-3 (42,842 bp)], *C. fluminea* [Cor-2 (37,605 bp), Cor-3 (37,272 bp)], *T. tubifex* [Tub-1 (38,378 bp), Tub-3 (35,661 bp)], and *S. suvaense* [Sco-1 (42,994 bp), Sco-6 (38,643 bp)]. Cor-2 and Cor-3 were overlapping with a completely matching 34,763-bp sequence. Sco-1 and Sco-6 were overlapping with a completely matching 24,029-bp sequence. Ast-1 and Ast-3 and Tub-1 and Tub-3 were derived from independent genes. *Nematostella vectensis* genomic DNA was obtained by PCR amplification of *N. vectensis* genomic DNA using following primers that are based on the draft genome sequences (http://www.ncbi.nlm.nih.gov/BLAST/tracemb.shtml): 5′-CCACCATGGAGGCGGATCGCGATTTGCCT-3′ and 5′-ATAAACTAATGCGGGTGCAGAAAACCTG-3′.

DNA Sequencing and Molecular Phylogenic Analysis

Sequencing and data assembly were done as described (Toyoda et al. 2002). Genomic sequences of *Schmidtea*

Table 1
Animals Used in This Study

Phylum Subphylum Class	Species	Abbreviation	Genome	cDNA	Msx-Related Genes (Synonym, Accession Number)
Chordata					
Vertebrata					
Mammalia	*Mus musculus*	Mm	DB	DB	Msx1 (AAH16426), Msx2 (Q03358), Msx3 (P70354)
Aves	*Gallus gallus*	Gg	DB	DB	Msx1 (Hox7, P50223), Msx2 (Hox8, P28362)
Amphibia	*Xenopus laevis*	Xl		DB	Msx1 (AAH81101)
Osteichthyes	*Danio rerio*	Dr	DB	DB	Msh-A (Q03357), Msh-B (Q03356), Msh-C (Q01703), Msh-D (Q01704), MsxE (AAB03273)
Cephalochordata					
Leptocardia	*Branchiostoma floridae*	Bf		DB	Msx (CAA10201)
Urochordata					
Ascidiacea	*Ciona intestinalis*	Ci	DB	DB	CAD56691
	Molgula oculata	Mo		DB	AAA87223
Appendicularia	*Oikopleura dioica*	Od		DB	AAW24005
Hemichordata					
Enteropneusta	*Saccoglossus kowalevskii*	Sk		DB	Msx (ABD97280)
Echinodermata					
Echinoidea	*Strongylocentrotus purpuratus*	Sp	DB	DB	SpMsx (AAB97688)
	Heliocidaris erythrogramma	Her		DB	Msx (AAY86177)
	Heliocidaris tuberculata	Ht		DB	Msx (AAY86178)
Asteroidea	*Asterina pectinifera*	Ap	Fosmid	PCR	MsxA [Ast-1 (AB302953)][a], MsxB [Ast-3 (AB302954)][a]
Arthropoda					
Chelicerata	*Limulus polyphemus*	Lp		PCR	MsxA (AB302959)[a]
	Pandinus imperator	Pi		PCR	MsxA (AB302960)[a], MsxB (AB302961)[a]
Crustacea	*Artemia franciscana*	Af		PCR	Msx (AB302962)[a]
Insecta	*Tribolium castaneum*	Tc	DB		Muscle segment homeoprotein (AAW21975)
	Drosophila melanogaster	Dm	DB	DB	msh (Q03372)
	Anopheles gambiae	Ag	DB	DB	EAA08817
	Apis mellifera	Ame	DB		Msx (AB362784)[b]
Tardigrada	*Milnesium tardigradum*	Mt		PCR	Msx (AB302966)[a]
Nematoda	*Caenorhabditis elegans*	Ce	DB	DB	vab-15 (Q09604)
	Caenorhabditis briggsae	Cb		DB	CAE58718
Mollusca					
Cephalopoda	*Loligo bleekeri*	Lb		PCR	Msx (AB302963)[a]
	Octopus ocellatus	Oo		PCR	Msx (AB302964)[a]
Bivalvia	*Corbicula fluminea*	Cf	Fosmid	PCR	Msx [Cor-2/Cor-3 (AB302955)][a]
Brachiopoda					
Inarticulata	*Lingula anatina*	La		PCR	Msx (AB302965)[a]
Annelida					
Polychaeta	*Platynereis dumerilii*	Pd		DB	Msx (CAJ38810)
Oligochaeta	*Tubifex tubifex*	Tt	Fosmid	PCR	MsxA [Tt-1 (AB302956)][a], MsxB [Tt-3 (AB302957)][a]
Hirudinida	*Helobdella* sp.	Hel		DB	Msx (AAB37254)
Echiura					
	Urechis caupo	Uc		PCR	Msx (AB302967)[a]
Platyhelminthes					
Tubellaria	*Schmidtea mediterranea*	Sm	DB		Msx (AB362785)[b]
	Thysanozoon sp.	Thy		PCR	Msx (AB302968)[a]
Cnidaria					
Hydrozoa	*Hydra viridis*	Hvi		DB	CAA45912
	Hydra vulgaris	Hvu		DB	CAB88390
	Scolionema suvaense	Ss	Fosmid	PCR	Msx [Sco-1/Sco-6 (AB302958)][a]
	Podocoryne carnea	Pc		DB	AAX58756
Anthozoa	*Nematostella vectensis*	Nv	PCR		Msx (AB362783)[a]
	Anemonia erythraea	Ae		PCR	Msx (AB302969)[a]
	Acropora millepora	Ami		DB	msh3 (ABK41269)
Porifera					
Demospongiae	*Ephydatia fluviatilis*	Ef		DB	prox3 (AAA20151)

NOTE.—DB, sequence collected from NCBI databases; fosmid, isolated as fosmid genomic clones in this study; PCR, cloned by PCR amplification of the cDNA.

[a] Sequence newly determined in this study.

[b] Sequences edited from the publicized draft sequences (http://www.ncbi.nlm.nih.gov/BLAST/tracemb.shtml).

mediterranea, *Hydra vulgaris*, *Caenorhabditis elegans*, *Strongylocentrotus purpuratus*, *Ciona intestinalis*, and *N. vectensis* were derived from public databases (http://www.ncbi.nlm.nih.gov/BLAST/tracemb.shtml). Sequence analysis was done with DNASISPro (Hitachi Software Engineering, Tokyo, Japan), Sequencher (Gene Codes, Ann Arbor, MI), and Genetyx (Genetyx, Tokyo, Japan) software. Homology searching was performed against a public database (http://www.ncbi.nlm.nih.gov/BLAST/) by using Blast and discontinuous MEGA Blast.

Table 2
Eh1 Motifs Found in N-terminal Regions of Metazoan Msx Proteins

Species		Eh1N		Eh1C
Gg_1	55	FSVEALMA	106	FPSVGALGK
Mm_1	61	FSVEALMA	118	FSVGGLL-K
Xl_1	67	FSVEALMA	120	YPVGAIM-Q[a]
Dr_E	44	FSVEALMA	68	FSVEVLQLP
Bf	46	FSVASLMA	91	FSVEGILSK
Ci	76	FSIEFLLS		
Od	6	FSVDWIIS		
Sk	73	FSVASLIS	112	FSVEGILSK
Sp	61	FRVESLFS	100	HSVENILAK
Her	51	FRVESFFS	104	HSVENILAK
Ap_A[b]	59	FRVESLVG	108	YTVEGILAS
Dm	134	FSVASLLA		
Ame	43	FSVDSLLS		
Ce	14	FSVESLLT		
Cf	79	FGVDSIIS	151	FSMDEILGK
Pd	132	FSVDSIIS	180	FSVDGILSK
Tt_A[b]	119	FSVAAIMA	170	FTVDGILGG
Ss	53	FSIDYILN		
Pc	30	FSIDYLLN		
Nv	29	FSVESLIS	57	FSVESILEK
Ami	14	FSVESLIS	43	FSVERLLDK
Consensus		FSV[D/E]S[L/I][L/I]S		FSV[D/E]GILXK

[a] The sequence was highly diverged from other Eh1C sequences. In *Xenopus tropicalis* (AAH62514), Eh1C sequence was FPVGGIMK.

[b] Ap_B and Tt_B Eh1N were identical to those in Ap_A and Tt_A, respectively.

The aa and nucleotide sequences were aligned by ClustalW (Thompson et al. 1994). Some of the aligned sequences were corrected by visual inspection. Ancestral sequences were deduced from the present aa sequence data by using ANCESCON, a distance-based program that gives more accurate ancestral sequence reconstruction than do PAML, PHYLIP, and PAUP* at large evolutionary distances (Cai et al. 2004). Phylogenetic tree analysis was done with MEGA3.1 (Neighbor-Joining [NJ] tree and Maximal Parsimony [MP] tree Kumar et al. 2004) and MrBayes 3.1.2 (Bayesian Inference [BI] tree Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). NJ tree was based on the distance calculation with point accepted mutation (PAM) matrix (Dayhoff et al. 1978) after removing position containing gaps (complete deletion option). In the NJ and MP trees, the tree reliability was estimated by bootstrap test (Felsenstein 1985) with 1,000 repetitions. In the BI analysis, we used an empirical model (WAG distances, Whelan and Goldman 2001) with gamma, alpha shape parameter, and aa frequencies estimated from the data. We ran 1,000,000 generations with 1 cold and 3 incrementally heated Markov chains, random starting trees for each chain, and trees sampled every 100 generations. We constructed a 50% major rule consensus tree from the last 15,000 trees that were saved (burnin = 2,500).

The evolutionary distances in figure 3*C* were calculated by MEGA3.1. The distances of aa sequences (Msx) and nucleotide sequences (18S RNA) were determined by using the PAM matrix (Dayhoff et al. 1978) and the Tamura–Nei model (Tamura and Nei 1993), respectively. Measurement was done after removal of any alignment gap-containing sites, assuming different evolutionary rates among sites (gamma distribution, $\alpha = 0.37$ for Msx aa sequences and $\alpha = 0.42$ for 18S ribosomal RNA nucleotide sequences). The α parameters were estimated by Tree-Puzzle program (Schmidt et al. 2002). For the evolutionary distance analysis, we omitted the paralogs and sequences from identical genera (those in *Heliocidaris*, *Caenorhabditis*, and *Hydra* species), except for one representative sequence. The representative sequences were the most strongly conserved in each group. According to these criteria, Mm_2, Gg_2, Dr_E, Ap_A, Pi_B, Tt_A, Her, Ce, and Hvu were used in the distance analysis, whereas Mm_1, Mm_3, Dr_A, Dr_B, Dr_C, Dr_D, Gg_1, Ap_B, Pi_A, Tt_B, Ht, Cb, and Hvi were not used.

Comparison of evolutionary rates among the chordate and cnidarian subgroups (fig. 3*D* and *E*) were done by counting different aa residues between the target aa sequences and a reference sequence (Anc_Msx) by using MEGA3.1. Counting was performed for the sequences from the Chordata and Cnidaria. Subsequently, the numbers of residues nonidentical and identical to the reference sequence were placed in 2×2 tables to compare the differences between 2 sequences from distinct sister subgroups. Fisher's exact test was performed to evaluate the statistical significance of the differences in the ratios of nonidentical to identical residues.

Plasmid Construction and Immunoprecipitation

Xl_Msx1 cDNA (a gift from Dr Atsushi Suzuki) was cloned into pCS2 + Myc tag vector (Turner and Weintraub 1994). Nv_Msx genomic DNA and mouse Grg1 cDNA were cloned into pcDNA3.1/His (Invitrogen) that was modified to have an initiation methionine, and either 2 Flag epitope tags, 3 HA tags at their *Hin*dIII/*Kpn*I sites. The Grg1 cDNA was obtained from Riken FANTOM clones (http://www.gsc.riken.go.jp/e/FANTOM/) (Carninci et al. 2005). Site-directed mutations were introduced into the protein-coding region according to the method of Ito et al. (1991). The primer sequences will be provided upon request. A truncation (stop) mutant of Xl_Msx1 was generated by inserting a stop linker (5′-TGAATATCA-3′) into a unique *Sma*I site in the open reading frame (ORF) of Xl_Msx cDNA.

Immunoprecipitation was performed essentially as described (Ishiguro et al. 2007). Briefly, COS7 cells were transfected with the epitope-tagged Msx and Grg1 expression vectors with Lipofectamine Plus reagent (Invitrogen). The transfected cells were washed and harvested in PBS(−) containing 1 mM phenylmethylsulfonylfluoride (PMSF), and total cell extracts were prepared with a lysis wash buffer consisting of 20 mM HEPES–KOH, pH 7.8, 10% glycerol, 150 mM NaCl, 0.5 mM dithiothreitol, 0.1 mM ethylenediaminetetraacetic acid, 0.5% NP-40, and 1 mM PMSF. The extracts were incubated with anti-FLAG or anti-HA affinity beads (Sigma, St Louis, MO) at 4 °C for 6 h. The beads were subsequently washed with the buffer. The bound proteins were separated by sodium dodecyl sulphate–polyacrylamide gel electrophoresis, transferred onto polyvinylidene difluoride membranes (Nihon Millipore, Tokyo, Japan), and detected by antibodies against the epitope tags using ECL western blotting detection reagent (GE Healthcare, Tokyo, Japan).

Results

Collection of Msx-Related Sequences from Metazoa

A homology search against current sequence databases revealed Msx orthologs could only be detected in metazoans. We did not find definitive Msx orthologs in fungi, algae, plants, and protists. This result is consistent with a previous finding that homeobox genes belonging to the ANTP class, in which Msx is included, are confined to the Metazoa (Holland and Takahashi 2005).

We first collected Msx family gene sequences through a database search. The homology search, which covered both complete and incomplete nucleotides and protein sequence databases, revealed the presence of Msx homologs in the Porifera, Cnidaria, Nematoda, Arthropoda, Annelida, Echinodermata, Hemichordata, and Chordata (table 1). Porifera *prox3* (Seimiya et al. 1994) was an Msx ortholog, as suggested by the phylogenetic tree analysis in Galle et al. (2005). Genomic information on exon–intron boundaries was available for vertebrates, insects, and a urochordate. To improve the comprehensiveness of our analysis, we newly cloned partial cDNA fragments of Msx homologs from a wide range of animals in the Platyhelminthes, Mollusca, Echiura, Tardigrada, Brachiopoda, Cnidaria, Annelida, Arthropoda, and Echinodermata. In addition, fosmid genomic clones were isolated from starfish (*A. pectinifera*), sludgeworm (*T. tubifex*), bivalve (*C. fluminea*), and jellyfish (*S. suvaense*), and their entire nucleotide sequence was determined. The number of Msx genes collected in a species varied from 1 to 5. We observed closely related, but significantly diverged, genes (paralogs) in *T. tubifex*, *A. pectinifera*, and *P. imperator*, besides those described earlier in vertebrates (Ekker et al. 1997). In total, we obtained 17 additional Msx sequences from 14 animal species from 9 animal phyla (table 1).

The aa Residues Functionally Important in Mammalian Msx HDs are Strongly Conserved in Metazoan Orthologs

After sequencing the collected cDNA and genomic clones, the aa sequences were deduced. We first compared the aa sequences of the HD and its C-terminally flanking (CF) region (fig. 1). The alignment revealed that the HDs were strongly conserved. The Msx HD has DNA-binding activity. The residues that are responsible for the molecular interaction with DNA bases (R2, R5, K46, I47, Q50, N51, and R58) or DNA phosphoribosyl backbones (K3, T6, F8, Y25, R31, W48, R53, K55, and K57) (Hovde et al. 2001) were absolutely conserved among the collected sequences. Some residues (F8 and R58) were conserved in Msx, but not in Dlx family proteins that contained HDs most similar to Msx (fig. 1) (Gauchat et al. 2000). The Msx HD is also known to physically interact with proteins that are essential for the molecular function of Msx. K3, R5, and F8 are required for interactions and transcriptional repression by the general transcription factor TFIIF (Zhang et al. 1996). Mutations in human MSX HDs, including R31P (STA, Hu et al. 1998), L13P (corresponding to MSX2 L154P in PFM, Wuyts et al. 2000), RK18-19del (corresponding to MSX2 RK159-160del in PFM, Wuyts et al. 2000), P7H (corresponding to MSX2 P148H in CSO, Jabs et al. 1993) are missense mutations that cause genetic disorders. These sites were conserved among the metazoan Msx proteins, except that RK18-19 had different residues in urochordates, insects, and cnidarian species (fig. 1). These results indicate that functionally important residues in mammalian Msx proteins are strongly conserved in all metazoan Msx HDs.

The Degree of HD Sequence Conservation Varies among Taxonomic Groups

To determine the structural features of the Msx proteins in each taxon, we first drew a phylogenetic tree, based on the alignment shown in figure 1, by using the NJ (fig. 2), BI (supplementary fig. 1, Supplementary Material online), and MP (supplementary fig. 2, Supplementary Material online) methods. Monophyly of the Hydrozoa (Hvu, Ss, and Pc) was recovered by the NJ/BI/MP trees, and those of the Diptera (Dm and Ag) and Cephalopoda (Oo and Lb) were recovered by the BI/MP trees. However, with the exception of paralogs, there were no other strong groupings supported by multiple trees. Instead, there was a high level of similarity between several evolutionarily distant animal species [e.g., only 2 changes in the 69 aa (9–77) sequences between Mm_2 (Chordata) and La (Brachiopoda) (fig. 1)]. We therefore speculated that the metazoan ancestral Msx sequence was strongly retained in some animal groups but not in others.

To test this idea, we determined the mean distances of the HD + CF sequence between all pairs of 12 taxa (fig. 3*A*). In this grouping, the Tardigrada were grouped with the Arthropoda in light of their sister-group relationship, as determined by morphological and molecular analyses (Brusca and Brusca 2003); the Echiura were grouped with the Annelida (McHugh 1997; Bleidorn et al. 2003); and the Hemichordata were grouped with the Echinodermata on the basis of recent molecular phylogenetic analyses (Cameron et al. 2000; Peterson 2004). Two classes in the Cnidaria—-Hydrozoa and Anthozoa—were tested separately because analysis of the phylogenetic trees suggested that there was a strong difference in tree length between these two taxa. For reference sequence comparison, we deduced the metazoan ancestor Msx sequence (Anc_Msx in fig. 1) by using ANCESCON, a program for the reconstruction of ancestral protein sequences that takes into account the observed variations in evolutionary rates between positions (Cai et al. 2004). We first determined the aa number where substitutions occurred in the Anc_Msx sequence for each taxon (fig. 3*B*). The sequences from the Vertebrata–Cephalochordata (*1* [group no. in fig. 3]), Echinodermata–Hemichordata (*3*), Mollusca (*6*), Brachiopoda (*7*), and Anthozoa (*11*) had low values (no greater than 6 aa substitutions), whereas those from the Urochordata (*2*), Hydrozoa (*10*), and Porifera (*12*) had large values (22–27 aa substitutions). The mean distances between all pairs of 2 taxa were then determined concerning the Msx and 18S ribosomal RNA sequences (fig. 3*C*, Supplementary table 1, Supplementary Material online). There were strongly deviated pairs along the distances of the Msx sequences. The lowest group (pairs under the oblique line in fig. 3*C*)

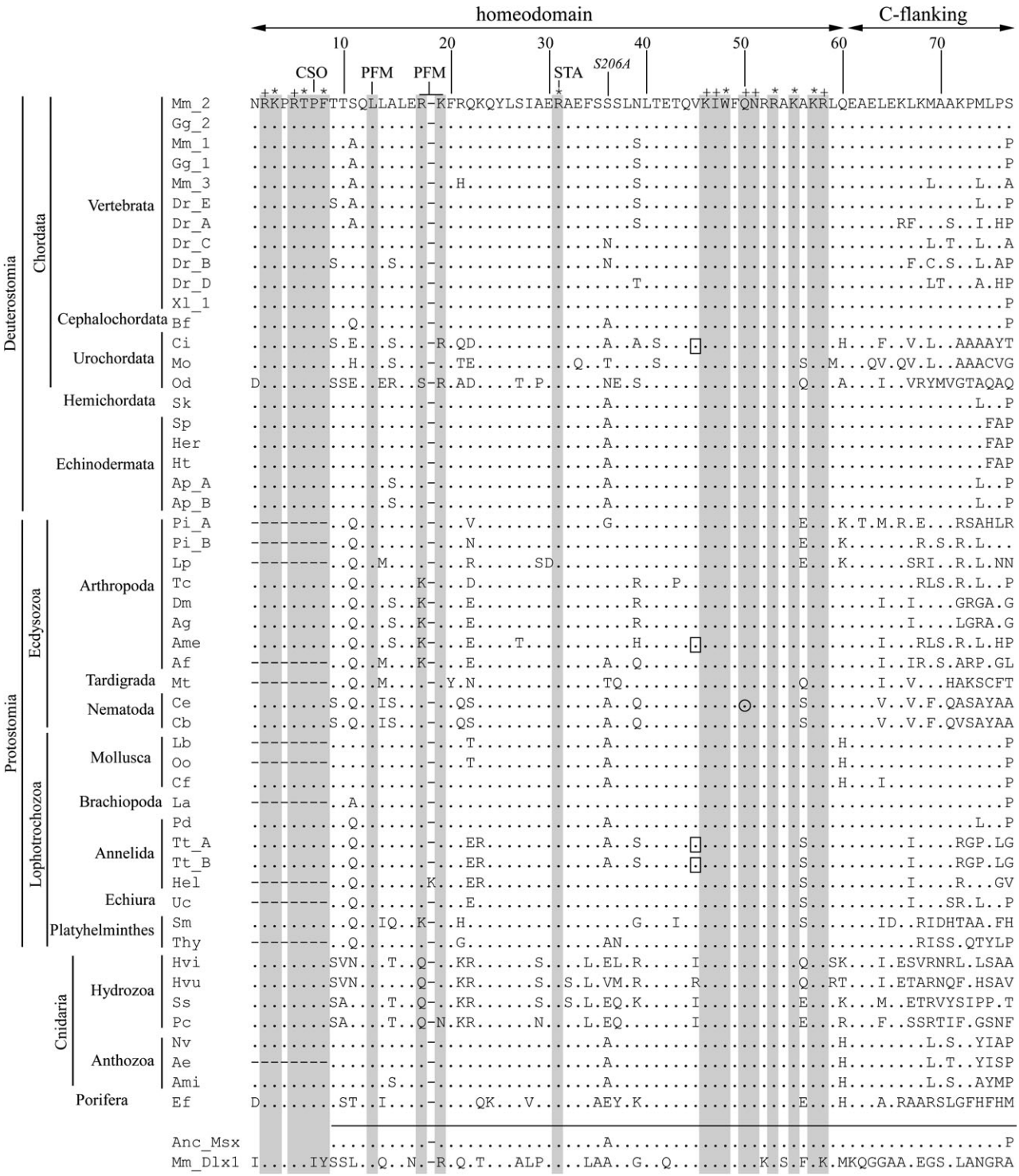


FIG. 1.—The aa alignment of HD and a C-terminal flanking region (HD + CF). "Dots" indicate identical aa residues to those in the top sequence (Mm_2), and "hyphens" indicate the spaces artificially inserted in the multiple alignment program. The abbreviations for the animal species names are given in table 1. *Anc_Msx* indicates the metazoan ancestral Msx sequence deduced by ANCESCON. "Plus" indicates the aa residues that are responsible for molecular interaction with DNA bases, and "asterisk" indicates DNA phosphoribosyl backbones. The locations of residues responsible for human genetic disorders (CSO, PFM, and STA; see text) are indicated at the top of the alignment. *S206A* indicates an artificial missense mutation introduced into Xl-Msx1 in figure 6. "Shading" indicates the locations of DNA-associated or disease-causative aa residues in the alignment. "Open boxes" and "open circle" indicate the presence of phase-0 and phase-2 introns, respectively. The genes subjected for the exon–intron boundary analysis are shown in figure 4. Bottom horizontal line indicates the region used in the calculation of evolutionary distances.

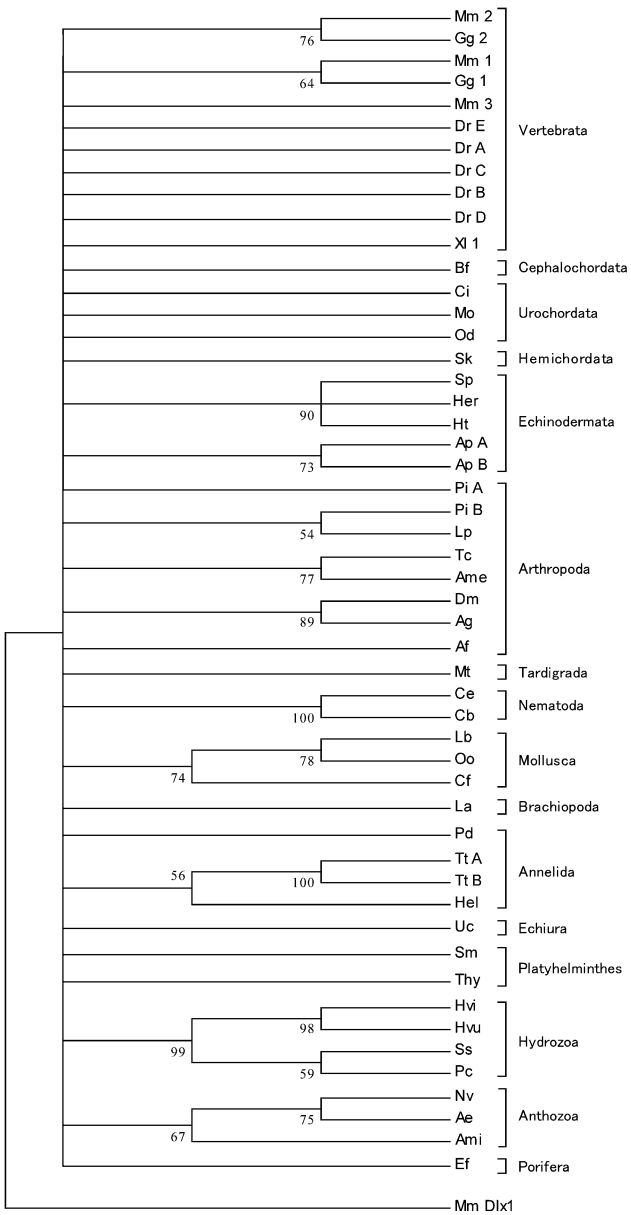


FIG. 2.—NJ tree used for the metazoan Msx HD + CF aa sequences. The tree was rooted with mouse Dlx1 (*Mm Dlx1*) as an outgroup. Internal labels indicate bootstrap values (1,000 replicates). PAM matrix was used for distance calculation by using the alignment shown in figure 1. Branches with less than 50% bootstrap values were condensed. Abbreviations for animal species names are indicated in table 1.

included pairs among the Vertebrata–Cephalochordata–Hemichordata, Echinodermata, Mollusca, Brachiopoda, and Anthozoa groups, whereas the highest one (pairs above the broken line in fig. 3*C*) contained pairs among mainly the Urochordata, Hydrozoa, and Porifera. From these results, we postulated that the Vertebrata, Cephalochordata, Hemichordata, Echinodermata, Mollusca, Brachiopoda, and Anthozoa Msx sequences shared highly conserved features among all the metazoan Msx sequences. On the other hand, the sequences belonging to the Urochordata, Hydrozoa, and Porifera diverged strongly.

Evolutionary Rates Differ between Two Independent Pairs of Sister Groups

The evolutionary rates of 2 representative phyla, the Cnidaria and Chordata, were compared. After comparison between the ancestor sequence and each sequence from the Cnidaria, we compared the diversification rates in any pair of sequences between the Anthozoa and Hydrozoa (fig. 3*D*, supplementary table 2, Supplementary Material online). In the Chordata, we compared the diversification rates between the Vertebrata, Cephalochordata, and Urochordata (fig. 3*E*, supplementary table 2, Supplementary Material online). Fisher's exact test of the results rejected the null hypothesis that evolutionary rates were equal between the 2 tested sequences in both the Anthozoa–Hydrozoa and the Vertebrata–Cephalochordata–Urochordata comparisons (all P values were less than 0.01). The analysis indicated clear differences in the evolutionary rates among sister groups in the Cnidaria and Chordata.

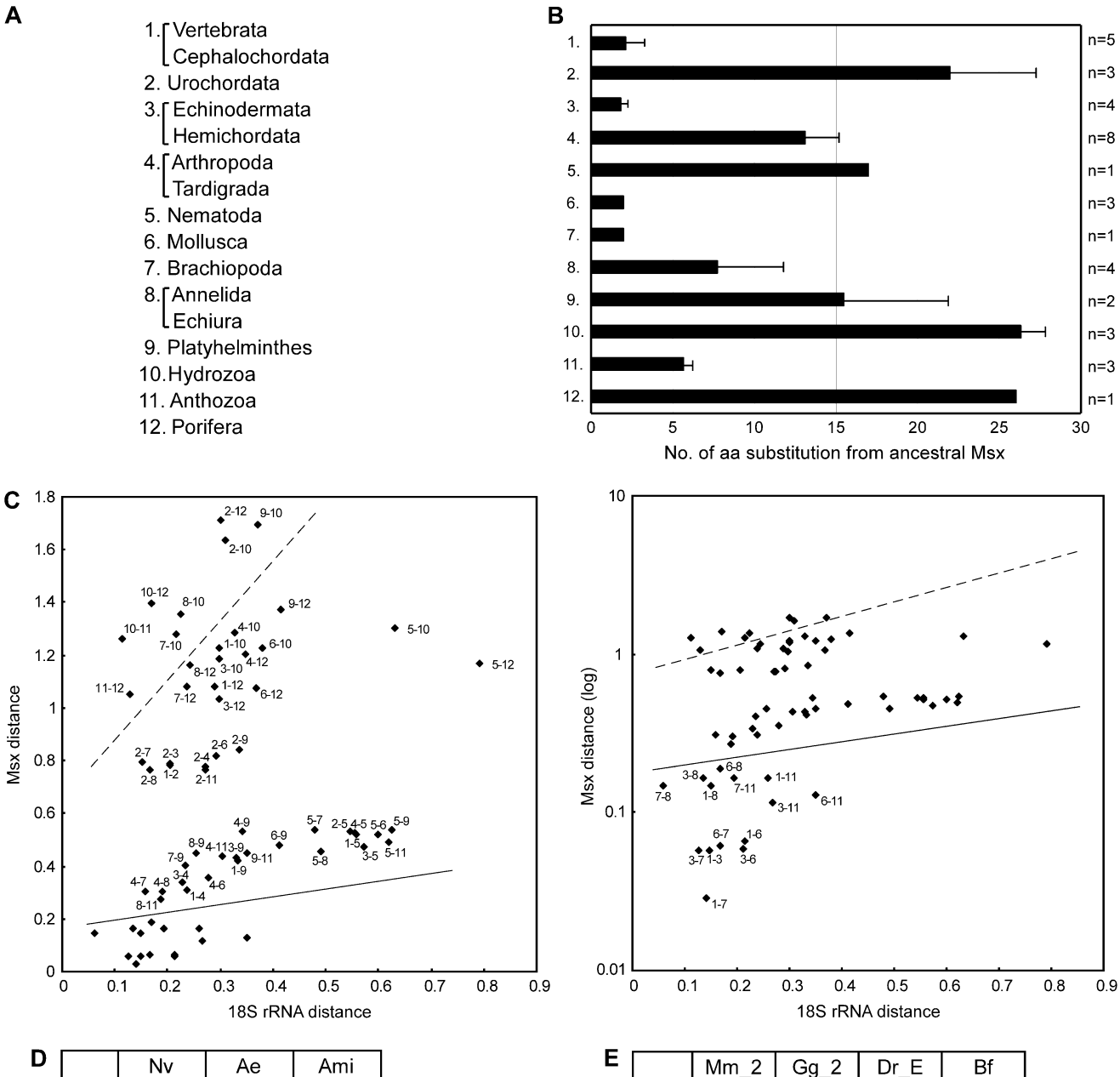
Comparison of Exon–Intron Organization of Msx Family Proteins

We next compared the exon–intron organization of Msx genes (fig. 4). All the Msx genes examined contained at least 1 intron in the protein-coding region. All of the exon–intron boundaries followed the GT/AG rule. Eighteen Msx genes possessed only 1 intron, whereas the other 5 (*Ame-Msx*, *Ci-Msh*, *Ce-VAB15*, *Tt-MsxA*, and *Tt-MsxB*) possessed additional introns. The one in the N-terminal region flanking the HD was located in a region with little sequence similarity (fig. 5). However, marked similarities were observed in the distance from the most N-terminal end of the HD (11–28 aa), and the phases of the intron insertion site in the ORF were "1" without exception. Furthermore, Seimiya et al. (1994) putatively assigned a "phase-1" splicing acceptor site in the 17 aa from the N-terminal end of the Ef-Msx (prox3) HD, although the preceding exon was not identified. When we aligned the N-terminal region aa sequence by adjusting the intron position as the cardinal point, there was a weakly conserved sequence near the intron position (fig. 5). The conserved sequence was summarized as FPWMQ, where tryptophan was strongly conserved (hereafter, we call the conserved sequence "PWM").

On the basis of these observations, we speculated that the intron was acquired in a common ancestor of a eumetazoa, or possibly a metazoan, Msx, and has been retained in the all Msx family genes; we call this intron the AC (absolutely conserved) intron. The other conserved intron position was located near the C-terminal end of the HD at V45 (phase 0) in 4 genes [*Tt-MsxA*, *Tt-MsxB* (Annelida), *Ame-Msx* (Arthropoda), and *Ci-Msh* (Urochordata)] (fig. 1).

Comparison of Non–HD Region aa Sequences

The N-terminal region aa sequences have been evolutionarily conserved in a wide range of metazoan animals (table 2). The most conserved ones were located at the N-terminal end, and the consensus sequence can be summarized as FSV[D/E]S[L/I][L/I]S, an Engrailed Homology 1 motif (Eh1), termed Eh1N in this paper. The other one



D

	Nv	Ae	Ami
Hvu	2.0E-05	2.0E-04	6.3E-05
Ss	8.6E-05	8.6E-05	2.5E-04
Pc	1.7E-04	1.7E-04	4.8E-04

E

	Mm_2	Gg_2	Dr_E	Bf
Ci	1.4E-04	1.4E-04	5.3E-03	2.4E-05
Mo	3.2E-05	3.2E-05	1.6E-03	4.8E-06
Od	4.2E-08	4.2E-08	5.7E-06	4.4E-09
Bf	1.0E+00	1.0E+00	2.1E-01	

FIG. 3.—Evolutionary distances computed from metazoan Msx aa sequences. (*A*) Taxonomic groups used in the analysis. Numbers correspond to those in (*B* and *C*). (*B*) aa substations in comparison with ancestral Msx (Anc_Msx in fig. 1). The numbers of substituted aa residues were counted for each Msx HD + CF sequence. Solid bars indicate mean numbers of substituted aa in each taxon shown in (*A*). Error bars indicate SDs. Error bars are shown if there is more than one sequence; the number of sequences in each group (*n*) are indicated in the right column. The within-group distances were as follows: Vertebrata–Cephalochordata–Hemichordata, 0.044; Urochordata, 1.106; Echinodermata, 0.032; Arthropoda–Tardigrada, 0.464; Mollusca, 0.021; Annelida–Echiura, 0.152; Platyhelminthes, 0.657; Hydrozoa, 0.889; and Anthozoa, 0.066. The values for the Vertebrata–Cephalochordata–Hemichordata, Echinodermata, Mollusca, and Anthozoa were clustered below 0.05. (*C*) Between-group means of evolutionary distances of all pairs of the 12 taxa in (*A*). Scatter graph indicates distances based on the Msx aa sequences [*y* axis, normal scale (left graph); log scale (right graph)] and those based on the 18S RNA gene nucleotide sequences (*x* axis). The 'oblique lines" are arbitrarily placed for the indication of the pairs with the low (below the continuous line) and high (above the broken line) values along the distance of Msx sequences. The sequences subjected to the analysis in (*B*) and (*C*) were selected according to the criteria in Materials and Methods. (*D* and *E*) Probabilities of equal molecular evolutionary rate, as determined by Fisher's exact test. The tests were done between pairs of sequence from Cnidaria sister groups (*D*) or Chordata sister groups (*E*). The abbreviation of the sequence name can be seen in table 1. In both analyses, the presumptive ancestral Msx sequence (Anc_Msx) was used as a reference sequence. Deuterostomia ancestral sequence was identical to Anc_Msx in the ANCESCON analysis.

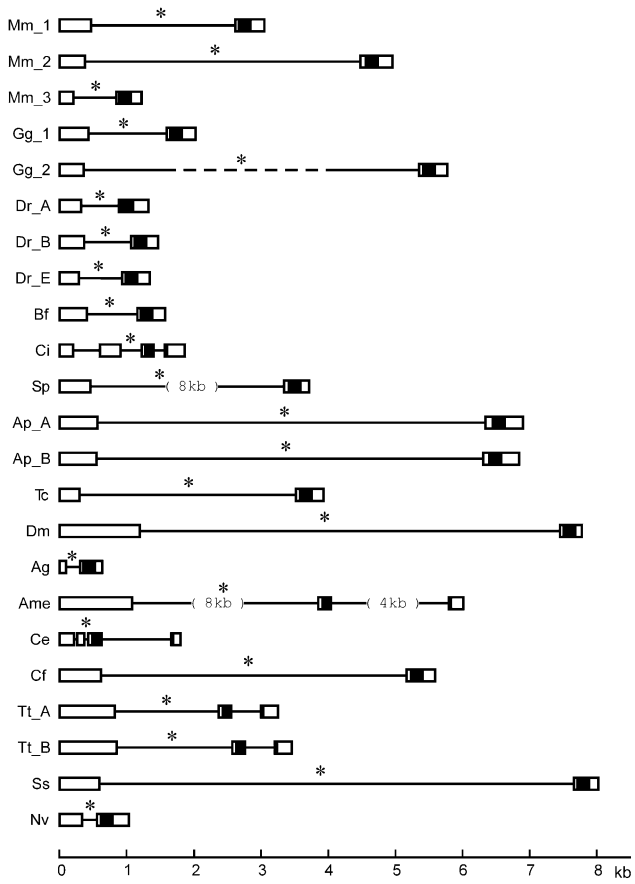


FIG. 4.—Schematic drawing of exon–intron structures of Msx genes from various animals. "Open boxes" and "closed boxes" indicate non–HD and HD areas, respectively, of the Msx protein-coding region. "Horizontal bars" indicate introns, and "asterisks" above the horizontal bars indicate AC introns.

was located between the Eh1N and PWM sequences. The generalized sequence motif of this domain was FSV[D/E]GILXK; it was thus similar to Eh1N and was termed Eh1C. Whereas Eh1N was conserved in all species, Eh1C showed scattered distribution among the invertebrate Msx proteins examined. When we mapped the presence of the Eh1C sequence and PWM sequence in the animal groups, it became clear that the Eh1C sequence was retained in the Vertebrata, Cephalochordata, Hemichordata, Echinodermata, Mollusca, and Anthozoa (Cnidaria) but not in the Urochordata and Hydrozoa (Cnidaria) Msx homologs isolated so far. In the Vertebrata, 2–5 paralogues existed in the examined species, and Eh1C-related sequence could be observed at least in 1 paralogue, but not always in the others. An examination of the current NCBI database revealed that, in the Vertebrata, all the examined Msx1 homologs in *Xenopus*, *Ambyostoma*, *Notophthalamus*, *Gallus*, *Mus*, *Rattus*, *Bos*, and all primate species kept the same Eh1N sequence, LPFSVEALMAD. All the Msx2 homologs in *Gallus*, *Mus*, *Rattus*, *Canis*, *Bos*, and all primates had LPFSVEALMSD (data not shown).

In addition to the Eh1-related motifs, short conserved sequence stretches were found in the N-terminal region (supplementary fig. 3*A*, Supplementary Material online) and in C-terminal flanking of the region shown in fig. 1 (supplementary fig. 3*B*, Supplementary Material online). The conservation of the 2 regions was limited to the sequences from Vertebrata, Cephalochordata, Echinodermata, Hemichordata, Mollusca, and Annelida. However, their functional significance was not clear at this point.

Eh1N and Eh1C can Act as Binding Sites for Groucho-Related Proteins

Eh1N and Eh1C were very close to the sequences for binding of the transcriptional corepressor Groucho (Paroush et al. 1994; Fisher et al. 1996; Tolkunova et al. 1998). To evaluate the functional significance of the conserved domains in the N-terminal regions, we generated mutant Msx proteins that had alanine substitutions in the conserved Eh1N, Eh1C, and PWM regions and analyzed their Groucho-related protein-binding abilities. For this purpose, expression constructs for FLAG or HA epitope-tagged mouse Groucho-related-gene1 (Grg1), Myc-tagged Xl-Msx1, and FLAG-tagged Nv-Msx were generated and used for a coimmunoprecipitation assay. When these expression vectors were transfected into COS7 cells, we detected each epitope-tagged protein with the expected molecular weight in an immunoblot assay (fig. 6 and data not shown). FLAG- or HA-Grg1 expression vector was then cotransfected with Myc-Xl-Msx1 or HA-Nv-Msx, respectively, and the Grg1 proteins were immunoprecipitated with anti-epitope tag antibodies. Both Myc-Xl-Msx1 and FLAG-Nv-Msx were coprecipitated with Grg1 (fig. 6*B* and *C*).

We also tested the physical interaction between Grg1 and mutant Msx1 proteins that contained substitution mutations in their conserved domains (fig. 6*A*). We prepared 5 expression vectors with substituted mutant Xl-*Msx1* (Stop, Eh1N, PWM, Eh1N&PWM, HD-S206A) (fig. 6*A*). The mutants lacking Eh1N (Eh1N, Eh1N&PWM) lost binding to Grg1 protein, whereas the PWM and HD substitution mutants retained Grg1-binding abilities comparable to those of wild-type Xl-Msx1 (fig. 6*B*).

Because Nv-Msx contained 2 Eh1-like sequences (Eh1N and Eh1C), we generated expression vectors for 4 Nv-Msx mutant proteins, including either single or combined mutants of the Eh1 domains (fig. 6*A*). Immunoprecipitation experiments showed that a mutation in Eh1N or Eh1C gave a strong or weak decrement, respectively, in Grg1-binding ability (fig. 6*C*), whereas the PWM mutants were unaffected. The Eh1N and Eh1C combined mutant (NvMsxEh1N&C) completely lacked the ability to bind to Grg1, suggesting that both domains bind to Grg1. Collectively, these results indicated that the Eh1-like domains can be bound by Grg1 protein in both vertebrate and anthozoan animals.

Discussion

Functional Significance of Msx Conserved Domains

Msx1 HD binds specific DNA target sequences (Hovde et al. 2001). In addition, many proteins bind mammalian Msx1 or Msx2 proteins [Lhx1 to HD (Bendall et al. 1998); Dlx to HD (Zhang et al. 1997); Grg to the N-terminal region (Rave-Harel et al. 2005); PIAS to the C-terminal region (Lee et al. 2006); Pax3 (Bendall et al. 1999), Pax9

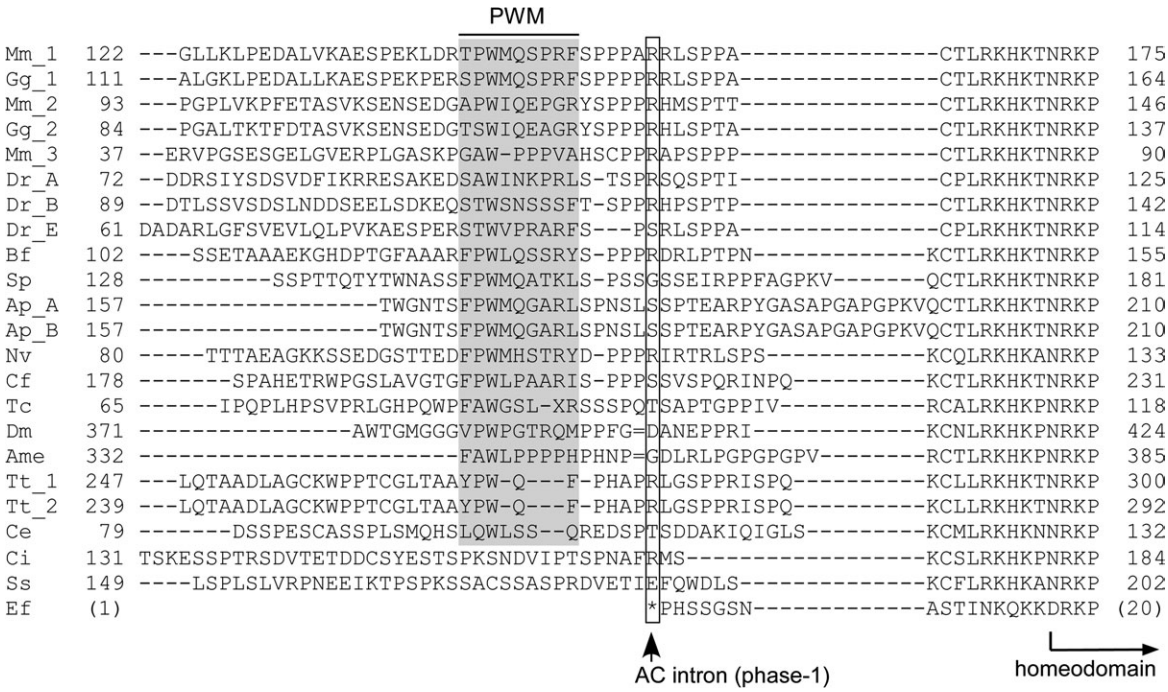


FIG. 5.—The aa alignment of the HD N-terminal flanking region. PWM indicates the newly identified conserved domain (shaded, PWM). All the AC introns were located in phase 1 of the "boxed" residues. "Long arrow" indicates the HD region. "Hyphens" and "equals marks" indicate the spaces and deletions, respectively, that were introduced for the alignment. The equals marks in Dm and Ame represent deletions of PPGMFPGAGFGG and HGHLYTPHGGPTSPN, respectively. The "asterisk" in the Ef line indicates the presence of a putative splicing acceptor site in phase-1 of the following HD-containing exon.

(Ogawa et al. 2006), and Histone H1b to the N-terminal region (Lee et al. 2004); and Dlxin-1 (Masuda et al. 2001) and TFIIF (RAP74 and RAP30) to the N-terminal region of the HD (Newberry et al. 1997)]. The physical interaction between these molecules may have acted as a form of evolutionary constraint. The overall conservation of the Msx HD suggests that interaction with these molecules is critical for the molecular functions of Msx. At the aa residue level, all the DNA-interacting and TFIIF-interacting residues in the Msx HD were conserved without exception, suggesting that interaction with these molecules, in particular, is essential for many metazoans to survive against natural selection.

Other structure–function information can be obtained through an analysis of the point mutations in human *MSX1* and *MSX2*. Some missense mutations in the HD are associated with STA, PFM, and CSO (fig. 1). Besides these HD mutations, it is noteworthy that a mutation in the methionine of Eh1N (L61K in human *MSX1*) is associated with oligodontia, a congenital form of tooth agenesis (Lidral and Reising 2002). Our results revealed that these residues are strongly conserved not only in primates, as revealed by Perry et al. (2006), but also in the Msx of many metazoans. The characteristic signs of these congenital anomalies suggest that the disease-associated residues are essential, at least for cranial and tooth development in humans. However, the strong conservation of the residues in invertebrate animals suggests that the same sequences can be used in various developmental and/or survival contexts of invertebrates. This view supports an idea that, in many metazoans, the gene encoding Msx protein is among the most fundamental and versatile.

Our results revealed that the Eh1N and Eh1C domains are required for the interaction of Msx with Grg1 in mammalian cells. The Eh1 domain was originally identified in the domain mediating the transcriptional repression of Drosophila engrailed protein through interaction with the transcriptional corepressor, Groucho. Although the Eh1-like domains have been identified in various transcription factors (Copley 2005), a recent study showed that Grg1 can physically interact with Msx1 and can regulate an Msx target gene (Rave-Harel et al. 2005). Our study revealed that Grg1 binding was mediated by an Eh1-like domain in Xl-Msx1 protein and that the 2 Eh1-like sequences in anthozoan Nv-Msx were able to bind to Grg1 in mammalian cells. These results suggest that Groucho-like–protein-binding activity is retained in metazoan Msx proteins. Interestingly, the Nematostella genome contains sequences highly similar to Grg1 (H.T. and J.A., unpublished data). Furthermore, vertebrate, Drosophila, and nematoda Groucho-related genes play critical roles in animal development (Pflugrad et al. 1997; Gasperowicz and Otto 2005). It would be interesting to see how putative Msx-binding domains in Groucho family proteins are conserved in metazoans. Collectively, these results suggest that the physical interaction between Msx and Groucho-like protein was established in the eumetazoan ancestor and has been utilized in various ways in the course of evolution.

Variation in the Degree of Protein Structure Conservation in the Msx Family

We focused here on the conserved domains in Msx proteins based on the phylogeny of species. Phylogenetic

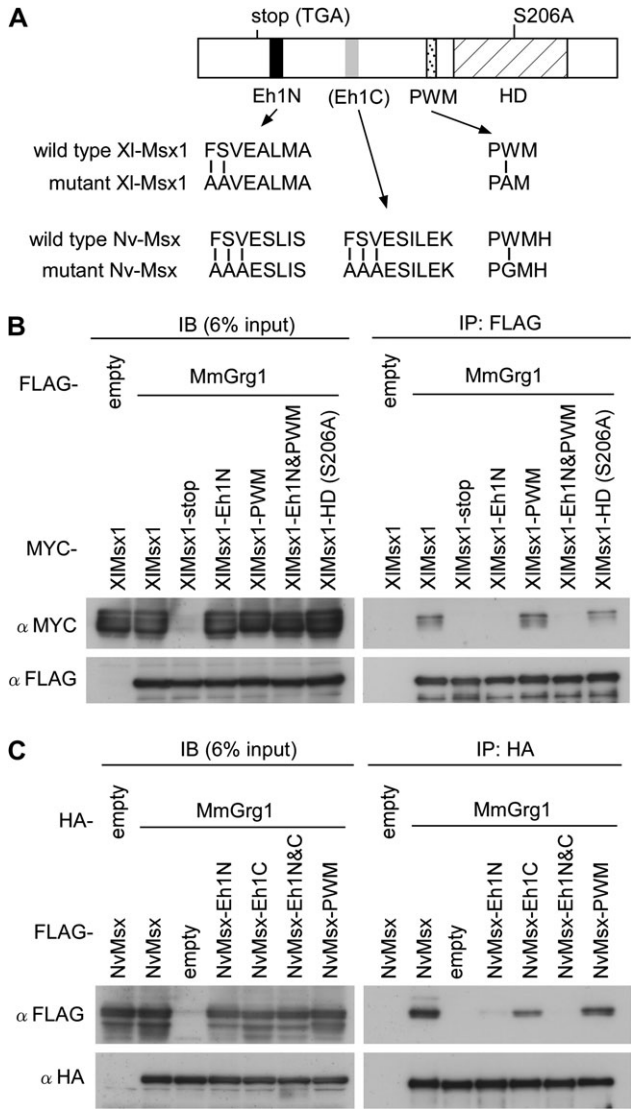


FIG. 6.—Functional characterization of the conserved domains. (*A*) Schematic drawing of conserved domains in the Msx proteins and the aa sequences in the mutant and wild-type Msx proteins. *Stop(TGA)*, a truncation mutant generated by inserting a stop codon just after proline 62 in Xl_Msx1 aa sequence (AAH81101); *S206A*, a substitution mutation in which serine 206 in the Xl_Msx1 was changed into alanine (position of the targeted serine residue is indicated in fig. 1). Substitution mutations in *Eh1N*, *Eh1C*, and *PWM* sequences of Xl_Msx1 or Nv_Msx are indicated in below the diagram. The highly diverged Eh1C sequence Xl_Msx1 (table 2) was not changed. (*B*) Immunoprecipitation assay between FLAG-Mm-Grg1– and Myc-Xl-Msx1–derived proteins. (*C*) Immunoprecipitation assay between HA-Mm-Grg1– and FLAG-Nv-Msx–derived proteins. In (*B*) and (*C*), the immunoprecipitation experiments were carried out by using cell extracts from COS7 cells cotransfected with Grg1 and Msx expression vectors. The results of immunoblot using the antibodies indicated on the left side of the panels are shown. The combinations of transfected vectors in each cell lysate are indicated at the tops of the pictures. Immunoblot analyses were performed both on the 6% of cell lysates subjected to the immunoprecipitation experiments (IB, 6% input) and the immunoprecipitated (IP) proteins.

trees of the 4 conserved domains that we characterized (Eh1N, Eh1C, PWM, and HD) showed a correlation between the degree of HD sequence divergence and the absence of Eh1C and PWM sequences (fig. 7). In the Hydrozoa and Urochordata, which showed HD sequence divergence in Msx, conservation of the Eh1C or PWM sequences was not evident. In contrast, the HD sequence of Msx was conserved in the Anthozoa, Echinodermata, Cephalochordata, Vertebrata, and Mollusca. HD in the Annelida, Arthropoda, and Nematoda showed an intermediate degree of sequence diversification; these taxa lacked Eh1C and had incomplete PWM sequences.

These results allow us to speculate on the overall Msx evolutionary process (fig. 7). Ancestral Msx appeared in the metazoan ancestor, and this gene probably possessed the AC intron. Because of the absence of the N-terminal region of Ef-Msx, we do not know whether the metazoan ancestor was already equipped with all the conserved domains. However, the eumetazoan ancestor, at the latest, may have already contained the Eh1N, Eh1C, PWM, and HD sequences. The 4 conserved sequences diverged differentially in the course of evolution. The degree of divergence was relatively small in the Anthozoa, Echinodermata, Cephalochordata, Vertebrata, and Mollusca but large in the Porifera, Hydrozoa, and Urochordata.

There are several potential pitfalls in this hypothesis. First, we are assuming that the diverged-type Msx genes were generated from conserved-type Msx genes. As an opposing idea, convergence to the one prototype sequence cannot be ruled out at this point. Second, roles of possible phylum-specific or class-specific conserved domains remains unclear due to insufficient numbers of sequences for the within-phyla comparison. In this regard, increasing-complexity-style evolution is still possible in Msx family evolution. Third, we might have missed highly diverged, but functionally equivalent, sequences because conservation of the Eh1N, Eh1C, and PWM sequences relies on short sequence stretches. Fourth, we cannot yet conclude that the variable divergence rates reflect general acceleration (or deceleration) of the molecular evolutionary rate or are limited to particular molecular species. Although we utilized the available 18S ribosomal RNA sequences as references, a comprehensive evaluation should be done with additional references. These points may be readily addressed by an extended analysis of additional molecular species.

Our hypothesis could also be evaluated by examining whether or not other genes fundamental to the organization of the animal body [so called "tool-kit" genes (Carroll et al. 2001)] show similar evolutionary tendencies. In a previous study, we found that selective loss of the conserved domain in the Zic family is found in certain animal taxa proteins (Aruga et al. 2006). In the case of Zic family proteins, conserved Zic can be seen in the Arthropoda, Mollusca, Annelida, Echinodermata, and Chordata (vertebrates and cephalochordates), whereas diverged Zic can be seen in the Platyhelminthes, Cnidaria, Nematoda, and Chordata (urochordates). On the basis of the phylogenetic distribution of the conserved Zic proteins, we proposed that the bilaterian ancestors had already acquired the full set of conserved domains that is found in currently living animals. Thus, in the cases of both the Msx and the Zic family proteins, the ancestral genes may already have possessed a set of conserved domains that were selectively lost in the course of evolution.

This differential divergence rate among the taxa could be involved in diversification of the organization of the

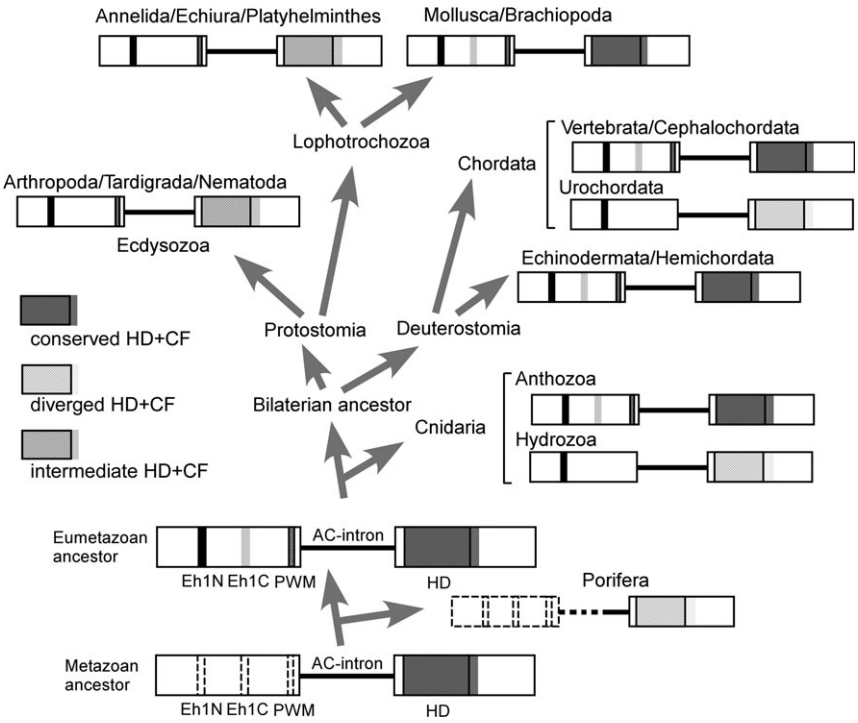


FIG. 7.—Evolutionary process of Msx genes. Hypothetical model that explains the diversity of Msx genes in the Eumetazoa. "Thick horizontal bars" indicate AC (absolutely conserved) introns. Dark, light, and intermediate gray boxes indicate conserved type, diverged type, and intermediate type HD + CF regions, respectively; black thick vertical lines indicate Eh1N; gray thick vertical lines indicate Eh1C; striped boxes indicate PWM; and hatched boxes indicate HD.

animal body. However, it is premature to conclude this because our understanding of the role of the Msx genes in each animal taxa is very limited. If we consider the Cnidaria sister groups, Anthozoa and Hydrozoa, hydrozoan Msx is expressed in regenerating muscle tissues (Yanze et al. 1999; Galle et al. 2005) and anthozoan Msx in the planula larval ectoderm (de Jong et al. 2006). However, because these reports dealt with expression profiles in only limited stages of the animals' life cycles, we are deterred from considering further the similarities and differences between Msx usage in the two classes of animals. If we consider the chordate Msx homologs, ascidian Msx genes appear to be commonly expressed in muscle, notochord, neural plate precursor cells, and the folding neural plate (Ma et al. 1996; Aniello et al. 1999). Some of these expression sites overlap with those of Msx gene expression in vertebrates (Davidson 1995), suggesting that some of the roles of Msx are shared between ascidians and vertebrates. However, vertebrate Msx genes are expressed in the limb bud, mandibular process, tooth, uterus, and other many organs that are not apparently found in ascidians. It is possible that the negative pressures for structural diversification differ between ascidians and vertebrates.

Finally, our results revealed the differential diversification of the Msx protein functional domain that became established in the eumetazoan ancestor in the course of evolution. The similarity of this evolution to that of Zic family proteins raises the possibility that a group of "tool-kit" genes shared the same feature, that is, differential diversification of the conserved domain. We await further extension of the metazoan-wide phylogenetic analysis of developmentally critical genes.

Supplementary Material

Supplementary text, figures 1–3, and tables 1 and 2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/). Database deposition: The sequences reported in the paper have been deposited in the GenBank/EMBL/DDBJ database accession numbers. AB302953–AB302969, AB362783–AB362785.

Acknowledgments

We thank Yayoi Nozaki and all the technical staff of the Sequence Technology Team of RIKEN GSC at the Research Resource Center, RIKEN BSI, for their technical assistance. We also thank Atsushi Suzuki (Hiroshima University) for the plasmid. This study was done as a collaborative research project of the Strategic Research Program at RIKEN. It was supported by grants-in-aid from the Japanese Ministry of Education, Science, Sports, and Technology.

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Aniello F, Locascio A, Villani MG, Di Gregorio A, Fucci L, Branno M. 1999. Identification and developmental expression of Ci-msxb: a novel homologue of *Drosophila* msh gene in Ciona intestinalis. Mech Dev. 88:123–126.

Arendt D, Nubler-Jung K. 1999. Comparison of early nerve cord development in insects and vertebrates. Development. 126: 2309–2325.

Aruga J, Kamiya A, Takahashi H, et al. (15 co-authors). 2006. A wide-range phylogenetic analysis of Zic proteins: implications for correlations between protein structure conservation and body plan complexity. Genomics. 87:783–792.

Bendall AJ, Abate-Shen C. 2000. Roles for Msx and Dlx homeoproteins in vertebrate development. Gene. 247: 17–31.

Bendall AJ, Ding J, Hu G, Shen MM, Abate-Shen C. 1999. Msx1 antagonizes the myogenic activity of Pax3 in migrating limb muscle precursors. Development. 126:4965–4976.

Bendall AJ, Rincon-Limas DE, Botas J, Abate-Shen C. 1998. Protein complex formation between Msx1 and Lhx2 homeoproteins is incompatible with DNA binding activity. Differentiation. 63:151–157.

Bleidorn C, Vogt L, Bartolomaeus T. 2003. New insights into polychaete phylogeny (Annelida) inferred from 18S rDNA sequences. Mol Phylogenet Evol. 29:279–288.

Brusca RC, Brusca GJ. 2003. The emergence of the Arthropods, Onychophorans, Tardigrades, Trilobites, and the Arthropod bauplan. Invertebrates. Sunderland (MA): Sinauer. p. 461–510.

Cai W, Pei J, Grishin NV. 2004. Reconstruction of ancestral protein sequences and its applications. BMC Evol Biol. 4:33.

Cameron CB, Garey JR, Swalla BJ. 2000. Evolution of the chordate body plan: new insights from phylogenetic analyses of deuterostome phyla. Proc Natl Acad Sci USA. 97:4469–4474.

Carninci P, Kasukawa T, Katayama S, et al. (194 co-authors). 2005. The transcriptional landscape of the mammalian genome. Science. 309:1559–1563.

Carroll SB, Grenier JK, Weatherbee SD. 2001. From DNA to diversity, molecular genetics and the evolution of animal design. Oxford: Blackwell.

Copley RR. 2005. The EH1 motif in metazoan transcription factors. BMC Genomics. 6:169.

Davidson D. 1995. The function and evolution of Msx genes: pointers and paradoxes. Trends Genet. 11:405–411.

Dayhoff MO, Schwarz RM, Orcutt BC. 1978. A model of evolutionary change in proteins. In: Dayhoff MO, editor. Atlas of protein sequence and structure. Silver Spring (MD): National Biomedical Research Foundation. p. 342–352.

de Jong DM, Hislop NR, Hayward DC, Reece-Hoyes JS, Pontynen PC, Ball EE, Miller DJ. 2006. Components of both major axial patterning systems of the Bilateria are differentially expressed along the primary axis of a 'radiate' animal, the anthozoan cnidarian *Acropora millepora*. Dev Biol. 298:632–643.

Du H, Chalfie M. 2001. Genes regulating touch cell development in *Caenorhabditis elegans*. Genetics. 158:197–207.

Ekker M, Akimenko MA, Allende ML, Smith R, Drouin G, Langille RM, Weinberg ES, Westerfield M. 1997. Relationships among msx gene structure and function in zebrafish and other vertebrates. Mol Biol Evol. 14:1008–1022.

Felsenstein J. 1985. Confidence limits on phylogenies, an approach using the bootstrap. Evolution. 39:783–791.

Fisher AL, Ohsako S, Caudy M. 1996. The WRPW motif of the hairy-related basic helix-loop-helix repressor proteins acts as a 4-amino-acid transcription repression and protein-protein interaction domain. Mol Cell Biol. 16:2670–2677.

Galle S, Yanze N, Seipel K. 2005. The homeobox gene Msx in development and transdifferentiation of jellyfish striated muscle. Int J Dev Biol. 49:961–967.

Gasperowicz M, Otto F. 2005. Mammalian Groucho homologs: redundancy or specificity? J Cell Biochem. 95:670–687.

Gauchat D, Mazet F, Berney C, Schummer M, Kreger S, Pawlowski J, Galliot B. 2000. Evolution of Antp-class genes and differential expression of Hydra Hox/paraHox genes in anterior patterning. Proc Natl Acad Sci USA. 97:4493–4498.

Holland PW. 1991. Cloning and evolutionary analysis of msh-like homeobox genes from mouse, zebrafish and ascidian. Gene. 98:253–257.

Holland PW, Takahashi T. 2005. The evolution of homeobox genes: implications for the study of brain development. Brain Res Bull. 66:484–490.

Hovde S, Abate-Shen C, Geiger JH. 2001. Crystal structure of the Msx-1 homeodomain/DNA complex. Biochemistry. 40:12013–12021.

Hu G, Vastardis H, Bendall AJ, et al. (12 co-authors). 1998. Haploinsufficiency of MSX1: a mechanism for selective tooth agenesis. Mol Cell Biol. 18:6044–6051.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 17:754–755.

Ishiguro A, Ideta M, Mikoshiba K, Chen DJ, Aruga J. 2007. Zic2-dependent transcriptional regulation is mediated by DNA-dependent protein kinase, poly-ADP ribose polymerase, and RNA helicase A. J Biol Chem. 282:9983–9995.

Isshiki T, Takeichi M, Nose A. 1997. The role of the msh homeobox gene during Drosophila neurogenesis: implication for the dorsoventral specification of the neuroectoderm. Development. 124:3099–3109.

Ito W, Ishiguro H, Kurosawa Y. 1991. A general method for introducing a series of mutations into cloned DNA using the polymerase chain reaction. Gene. 102:67–70.

Jabs EW, Muller U, Li X, et al. (12 co-authors). 1993. A mutation in the homeodomain of the human MSX2 gene in a family affected with autosomal dominant craniosynostosis. Cell. 75:443–450.

Jumlongras D, Bei M, Stimson JM, Wang WF, DePalma SR, Seidman CE, Felbor U, Maas R, Seidman JG, Olsen BR. 2001. A nonsense mutation in MSX1 causes Witkop syndrome. Am J Hum Genet. 69:67–74.

Khadka D, Luo T, Sargent TD. 2006. Msx1 and Msx2 have shared essential functions in neural crest but may be dispensable in epidermis and axis formation in *Xenopus*. Int J Dev Biol. 50:499–502.

Kumar S, Tamura K, Nei M. 2004. MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform. 5:150–163.

Lallemand Y, Nicola MA, Ramos C, Bach A, Cloment CS, Robert B. 2005. Analysis of Msx1; Msx2 double mutants reveals multiple roles for Msx genes in limb development. Development. 132:3003–3014.

Lee H, Habas R, Abate-Shen C. 2004. MSX1 cooperates with histone H1b for inhibition of transcription and myogenesis. Science. 304:1675–1678.

Lee H, Quinn JC, Prasanth KV, Swiss VA, Economides KD, Camacho MM, Spector DL, Abate-Shen C. 2006. PIAS1 confers DNA-binding specificity on the Msx1 homeoprotein. Genes Dev. 20:784–794.

Lidral AC, Reising BC. 2002. The role of MSX1 in human tooth agenesis. J Dent Res. 81:274–278.

Ma L, Swalla BJ, Zhou J, Dobias SL, Bell JR, Chen J, Maxson RE, Jeffery WR. 1996. Expression of an Msx homeobox gene in ascidians: insights into the archetypal chordate expression pattern. Dev Dyn. 205:308–318.

Master VA, Kourakis MJ, Martindale MQ. 1996. Isolation, characterization, and expression of Le-msx, a maternally expressed member of the msx gene family from the glossiphoniid leech, Helobdella. Dev Dyn. 207:404–419.

Masuda Y, Sasaki A, Shibuya H, Ueno N, Ikeda K, Watanabe K. 2001. Dlxin-1, a novel protein that binds Dlx5 and regulates its transcriptional function. J Biol Chem. 276:5331–5338.

McHugh D. 1997. Molecular evidence that echiurans and pogonophorans are derived annelids. Proc Natl Acad Sci USA. 94:8006–8009.

Monsoro-Burq AH, Wang E, Harland R. 2005. Msx1 and Pax3 cooperate to mediate FGF8 and WNT signals during Xenopus neural crest induction. Dev Cell. 8:167–178.

Nose A, Isshiki T, Takeichi M. 1998. Regional specification of muscle progenitors in Drosophila: the role of the msh homeobox gene. Development. 125:215–223.

Ogawa T, Kapadia H, Feng JQ, Raghow R, Peters H, D'Souza RN. 2006. Functional consequences of interactions between Pax9 and Msx1 genes in normal and abnormal tooth development. J Biol Chem. 281:18363–18369.

Paroush Z, Finley RL Jr, Kidd T, Wainwright SM, Ingham PW, Brent R, Ish-Horowicz D. 1994. Groucho is required for Drosophila neurogenesis, segmentation, and sex determination and interacts directly with hairy-related bHLH proteins. Cell. 79:805–815.

Perry GH, Verrelli BC, Stone AC. 2006. Molecular evolution of the primate developmental genes MSX1 and PAX9. Mol Biol Evol. 23:644–654.

Peterson KJ. 2004. Isolation of Hox and Parahox genes in the hemichordate Ptychodera flava and the evolution of deuterostome Hox genes. Mol Phylogenet Evol. 31:1208–1215.

Pflugrad A, Meir JY, Barnes TM, Miller DM 3rd. 1997. The Groucho-like transcription factor UNC-37 functions with the neural specificity gene unc-4 to govern motor neuron identity in C. elegans. Development. 124:1699–1709.

Postlethwait JH. 2006. The zebrafish genome: a review and case study of msx genes. Genome Dyn. 2:183–197.

Ramos C, Robert B. 2005. msh/Msx gene family in neural development. Trends Genet. 21:624–632.

Rave-Harel N, Miller NL, Givens ML, Mellon PL. 2005. The Groucho-related gene family regulates the gonadotropin-releasing hormone gene through interaction with the homeodomain proteins MSX1 and OCT1. J Biol Chem. 280: 30975–30983.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19:1572–1574.

Satokata I, Ma L, Ohshima H, et al. (14 co-authors). 2000. Msx2 deficiency in mice causes pleiotropic defects in bone growth and ectodermal organ formation. Nat Genet. 24:391–395.

Satokata I, Maas R. 1994. Msx1 deficient mice exhibit cleft palate and abnormalities of craniofacial and tooth development. Nat Genet. 6:348–356.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics. 18:502–504.

Seimiya M, Ishiguro H, Miura K, Watanabe Y, Kurosawa Y. 1994. Homeobox-containing genes in the most primitive metazoa, the sponges. Eur J Biochem. 221:219–225.

Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. J Mol Biol. 147:195–197.

Suzuki A, Ueno N, Hemmati-Brivanlou A. 1997. *Xenopus* msx1 mediates epidermal induction and neural inhibition by BMP4. Development. 124:3037–3044.

Suzuki AC. 2003. Life history of *Milnesium tardigradum* Doyére (tardigrada) under a rearing environment. Zool Sci. 20:49–57.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol. 10:512–526.

Tan H, Ransick A, Wu H, Dobias S, Liu YH, Maxson R. 1998. Disruption of primary mesenchyme cell patterning by misregulated ectodermal expression of SpMsx in sea urchin embryos. Dev Biol. 201:230–246.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Tolkunova EN, Fujioka M, Kobayashi M, Deka D, Jaynes JB. 1998. Two distinct types of repression domain in engrailed: one interacts with the groucho corepressor and is preferentially active on integrated target genes. Mol Cell Biol. 18:2804–2814.

Toyoda A, Noguchi H, Taylor TD, Ito T, Pletcher MT, Sakaki Y, Reeves RH, Hattori M. 2002. Comparative genomic sequence analysis of the human chromosome 21 down syndrome critical region. Genome Res. 12:1323–1332.

Turner DL, Weintraub H. 1994. Expression of achaete-scute homolog 3 in *Xenopus* embryos converts ectodermal cells to a neural fate. Genes Dev. 8:1434–1447.

van den Boogaard MJ, Dorland M, Beemer FA, van Amstel HK. 2000. MSX1 mutation is associated with orofacial clefting and tooth agenesis in humans. Nat Genet. 24:342–343.

Vastardis H, Karimbux N, Guthua SW, Seidman JG, Seidman CE. 1996. A human MSX1 homeodomain missense mutation causes selective tooth agenesis. Nat Genet. 13:417–421.

Walldorf U, Fleig R, Gehring WJ. 1989. Comparison of homeobox-containing genes of the honeybee and *Drosophila*. Proc Natl Acad Sci USA. 86:9971–9975.

Wheeler SR, Carrico ML, Wilson BA, Skeath JB. 2005. The Tribolium columnar genes reveal conservation and plasticity in neural precursor patterning along the embryonic dorsal-ventral axis. Dev Biol. 279:491–500.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. 18: 691–699.

Wilkie AO, Tang Z, Elanko N, Walsh S, Twigg SR, Hurst JA, Wall SA, Chrzanowska KH, Maxson RE Jr. 2000. Functional haploinsufficiency of the human homeobox gene MSX2 causes defects in skull ossification. Nat Genet. 24:387–390.

Wilson KA, Andrews ME, Rudolf Turner F, Raff RA. 2005. Major regulatory factors in the evolution of development: the roles of goosecoid and Msx in the evolution of the direct-developing sea urchin *Heliocidaris erythrogramma*. Evol Dev. 7:416–428.

Wuyts W, Reardon W, Preis S, Homfray T, Rasore-Quartino A, Christians H, Willems PJ, Van Hul W. 2000. Identification of mutations in the MSX2 homeobox gene in families affected with foramina parietalia permagna. Hum Mol Genet. 9:1251–1255.

Yanze N, Groger H, Muller P, Schmid V. 1999. Reversible inactivation of cell-type-specific regulatory and structural genes in migrating isolated striated muscle cells of jellyfish. Dev Biol. 213:194–201.

Zhang H, Catron KM, Abate-Shen C. 1996. A role for the Msx-1 homeodomain in transcriptional regulation: residues in the N-terminal arm mediate TATA binding protein interaction and transcriptional repression. Proc Natl Acad Sci USA. 93:1764–1769.

Zhang H, Hu G, Wang H, Sciavolino P, Iler N, Shen MM, Abate-Shen C. 1997. Heterodimerization of Msx and Dlx homeoproteins results in functional antagonism. Mol Cell Biol. 17:2920–2932.

William Martin, Associate Editor

Accepted October 11, 2007