Chapter 4.

Evolution of novel protein property after gene duplication by weak selection on many amino acid sites

4.1. Introduction

In the previous chapter, I found that relatively weak selection on a large number of amino acid sites has played an important role in the evolution of novel structural properties of a protein. This perspective was obtained from comparative evolutionary analysis of phosphoglucose isomerase genes (*Pgi*; EC. 5.3.1.9) by focusing on the electric charges of the proteins.

To examine whether or not the mode of adaptive molecular evolution drawn from this analysis is relevant to other proteins as well, I analyzed another enzyme involved in glycolysis and gluconeogenesis, fructose-1,6-bisphosphate aldolase (ALD; EC 4.1.2.13) in this chapter. The gene encoding this enzyme (*Ald*) is estimated to have undergone two duplication events during the evolution of jawed vertebrates, but after the separation of jawless fish (Merritt and Quattro, 2002; Steinke et al., 2006). Probably because of these duplications, most of the jawed vertebrates possess three ALD isoforms, ALD-A, ALD-B, and ALD-C. These isoforms are known to show differential expression in tissue-specific manner and catalytic preference for substrates (Gamblin et al., 1991). ALD-A is efficient in glycolysis and strongly expressed in skeletal muscle. Otherwise, ALD-B is efficient in gluconeogenesis and expressed only in liver and kidney. ALD-C has intermediate catalytic properties and is predominant in brain. All this information implies that the duplicated *Ald* genes in jawed vertebrates underwent sub-neofunctionalization (Force et al., 1999; Lynch and Force, 2000A; He and Zhang, 2005; Rastogi and Liberles, 2005) with respect to their expression and catalytic function. Therefore, the *Ald* gene family in jawed vertebrates would be an

appropriate model to investigate the molecular evolution after subfunctionalization.

It is also known that each of the three ALD-A, ALD-B, and ALD-C isoforms have a specific electric charge (Merritt and Quattro, 2002). Among them, particularly interesting is the electric charge of the ALD-C, which is remarkably negative (estimated isoelectric point [pI]: 6.2 to 6.7) as compared to the ALD-A and ALD-B (estimated pI: 8.2 to 8.4; Merritt and Quattro, 2002). Therefore, analyzing the evolutionary trajectory of electric charge following separation between ALD-C and the other isoforms will provide a new opportunity for testing the generality of the finding in Chapter 3.

In this chapter, therefore, I analyze the ALD protein family in jawed vertebrates by focusing on the evolutionary changes in the electric charge. Firstly, in order to confirm the origins of the three ALD isoforms, a molecular phylogeny was estimated for *Ald* genes obtained from divergent lineages of jawed vertebrates. Next, a comparative evolutionary approach used in the Chapter 3 was also applied for the analysis of ALD, and the result confirmed the role of weak selection on many amino acid sites during the evolution of proteins. Moreover, to compare with ALD and PGI, I focused on other proteins, in which the evolution of electric charge is supposed to be driven by strong selection on particular amino acid sites, triose-phosphate isomerase (TPI; EC 5.3.1.1; Merritt and Quattro, 2001) and pancreatic ribonuclease (RNASE1; EC 3.1.27.5; Zhang, et al., 2002; Zhang, 2006). From examinations of tertiary structures, amino acid compositions, and evolutionary patterns of the proteins, I suggested a possible relationship between a mode of molecular evolution and a number of amino acid sites available for evolutionary modification.

4.2. Materials and Methods

4.2.1. Species and *Ald* genes analyzed in this study

I chose 12 representative species from divergent lineages of jawed vertebrates for which whole-genome sequence data is available, as follows—tetrapods: *Homo sapiens* (human), *Mus musculus* (mouse), *Canis familiaris* (dog), *Bos taurus* (cow), *Ornithorhynchus anatinus* (platypus), *Gallus gallus* (chicken), *Xenopus tropicalis* (clawed frog); teleost fish: *Tetraodon nigroviridis* (pufferfish), *Fugu rubripes* (Fugu), *Gasterosteus aculeatus* (stickleback), *Oryzias latipes* (medaka), *Danio rerio* (zebrafish). Protein and cDNA sequences of *Ald* genes from these vertebrates were gathered via queries to the Ensembl genome database (Birney et al., 2006) and the "orthologue prediction" section in the database. Putative orthologues described in this section are the genes identified through a reciprocal BLAST approach among genome sequence data from multiple species. I further confirmed that there were no extra *Ald* genes in the genome sequences by my own BLAST searches (*E*-value cut-off of $<10^{-3}$). As outgroups for subsequent phylogenetic analysis, *Ald* genes from agnathan *Lethenteron japonicum* (lamprey) were chosen. The list of species, *Ald* genes identified, and their Ensembl Gene IDs or GenBank accession numbers were shown in Table 4-1.

4.2.2. Phylogenetic analysis

Amino acid sequences of ALD proteins from 13 vertebrates (Table 4-1) were aligned using ClustalW (Thompson et al., 1994) and phylogenetically analyzed with the maximum likelihood methods (ML) by using the program TREEFINDER software package (version of June 2007; Jobb, 2007). The program ProtTest 1.4 (Abascal et al., 2005) was used to select the best-fitting model of amino acid substitution. The resultant tree topology was evaluated by approximate bootstrap tests (LR-ELW edge support: the Expected-Likelihood Weights applied to Local Rearrangements of tree topology; Strimmer and Rambaut, 2002; Jobb, 2007) with 1000 replications.

Table 4-1.

List of vertebrate species and *Ald* genes analyzed in this study, with Ensembl Gene IDs or GenBank accession numbers.

Common name	Species	Ensembl Gene ID or GenBank Accession No.
Human	*Homo sapiens*	ENSG00000149925 (*Ald-A*)
		ENSG00000136872 (*Ald-B*)
		ENSG00000109107 (*Ald-C*)
Mouse	*Mus musculus*	ENSMUSG00000030695 (*Ald-A*)
		ENSMUSG00000028307 (*Ald-B*)
		ENSMUSG00000017390 (*Ald-C*)
Dog	*Canis familiaris*	ENSCAFG00000008221 (*Ald-A*)
		ENSCAFG00000002598 (*Ald-B*)
		ENSCAFG00000018719 (*Ald-C*)
Cow	*Bos taurus*	ENSBTAG00000012927 (*Ald-A*)
		ENSBTAG00000015358 (*Ald-B*)
		ENSBTAG00000013099 (*Ald-C*)
Platypus	*Ornithorhynchus anatinus*	ENSOANG00000008135 (*Ald-A*)
		ENSOANG00000009964 (*Ald-B*)
		ENSOANG00000014180 (*Ald-C*)
Chicken	*Gallus gallus*	ENSGALG00000015544 (*Ald-B*)
Clawed frog	*Xenopus tropicalis*	ENSXETG00000020674 (*Ald-A*)
		ENSXETG00000024058 (*Ald-B*)
		ENSXETG00000027677 (*Ald-C*)
Pufferfish	*Tetraodon nigroviridis*	GSTENG00028308001 (*Ald-A*)
		GSTENG00026570001 (*Ald-B*)
		GSTENG00018960001 (*Ald-C1*)
		GSTENG00030103001 (*Ald-C2*)
Fugu	*Fugu rubripes*	SINFRUG00000136330 (*Ald-A*)
		SINFRUG00000160637 (*Ald-B*)
		SINFRUG00000145351 (*Ald-C1*)
		SINFRUG00000129245 (*Ald-C2*)
Stickleback	*Gasterosteus aculeatus*	ENSGACG00000007727 (*Ald-A1*)
		ENSGACG00000004178 (*Ald-A2*)
		ENSGACG00000017982 (*Ald-B*)
		ENSGACG00000020619 (*Ald-C1*)
		ENSGACG00000011018 (*Ald-C2*)

Medaka	*Oryzias latipes*	ENSORLG00000006367 (*Ald-A*)
		ENSORLG00000004742 (*Ald-B*)
		ENSORLG00000005349 (*Ald-C1*)
		ENSORLG00000004617 (*Ald-C2*)
Zebrafish	*Danio rerio*	ENSDARG00000011665 (*Ald-A1*)
		ENSDARG00000034470 (*Ald-A2*)
		ENSDARG00000053684 (*Ald-B*)
		ENSDARG00000019702 (*Ald-C1*)
		ENSDARG00000057661 (*Ald-C2*)
Lamprey	*Lethenteron japonicum*	D38620 (*Ald-1*)
		D38619 (*Ald-2*)

4.2.3. Charge evolution analysis

The ML inference of the ancestral sequences of *Ald* genes was performed under methods that have been described previously in Chapter 3 (*3.2.6. Charge evolution analysis*), on the basis of the *Ald* phylogeny estimated in this study, and the known phylogeny of mammals (Murphy et al., 2004) and teleost fish (Yamanoue et al., 2006). A nucleotide sequence alignment of the coding region of *Ald* genes (1089 base pairs) from 12 jawed vertebrates plus lamprey was used. Because the program BASEML (Yang, 1997) cannot perform ancestral state inference for nucleotide sites with gaps, gapped regions in the alignment (136–138 in *Ald-C1* of pufferfish, 328–330 in *Ald-A* of platypus, 376–378 in *Ald-A* of pufferfish, 1000–1011 in *Ald-A* of pufferfish, 1000–1089 in *Ald-C2* of pufferfish and Fugu, 1039–1041 in *Ald-C* of mouse and *Ald-1* of lamprey, 1048–1050 in *Ald-C1* of teleosts, 1054–1056 in *Ald-A1* of pufferfish, Fugu, stickleback and medaka, and *Ald-A2* of stickleback, and 1072–1074 of *Ald-C2* of stickleback) were filled by a relevant sequence of the most closely related gene, respectively. The filled regions were subsequently re-deleted from the estimated ancestral sequences that parsimoniously considered to have undergone the deletion of the relevant region.

For maximum likelihood inference of ancestral nucleotide sequences, the GTR + Γ (Yang, 1994) model was selected as the best fitting model by the hLRTs (Table 4-2). The average overall accuracy of the reconstructed sequences (#1–#32) was 0.9560 ± 0.0052 SE. The solvent-accessible surface area (SASA) of each amino acid residue in ALD protein molecule was estimated with GETAREA 1.1 (Fraczkiewicz and Braun, 1998) using the X-ray refined crystal structure of human ALD-A (Dalby et al., 1999) as a reference data. A solvent radius of 1.4 Å (approximately the size of a water molecule) was used. The structural portion of the ALD composed of amino acid residues with more than 20 $Å^2$ SASA was considered "molecular surface." This boundary mostly agreed with other criteria based on the ratio of

Table 4-2.

Hierarchical likelihood ratio tests (hLRTs) among nested models of nucleotide substitution.

Models	Likelihood scores (lnL)	Likelihood ratio tests[1]			
		K80 + Γ [4]	HKY85 + Γ [7]	TN93 + Γ [8]	GTR + Γ [11]
JC69 + Γ [3]	-18883	824.0* (1)	852.7* (4)	962.3* (5)	990.3* (8)
K80 + Γ [4]	-18471	NA	28.7* (3)	138.3* (4)	166.3* (7)
HKY85 + Γ [7]	-18457	NA	NA	109.6* (1)	137.6* (4)
TN93 + Γ [8]	-18402	NA	NA	NA	28.0* (3)
GTR + Γ [11]	-18388	NA	NA	NA	NA

*$P < 0.01$

[1]The likelihood ratio test statistic ($2\Delta L$) is approximated using the χ^2 distribution with degrees of freedom (in parentheses) equal to the difference in the number of parameters (in brackets) between the comparing pairs of nucleotide substitution models. The best-fitting model is indicated in bold.

side-chain surface area to random coil value per residue (Fraczkiewicz and Braun, 1998). A three-dimensional graphical model of the ALD molecule was constructed using RasMol (Sayle and Milner-White, 1995).

4.2.4. Calculation of the expected spatial distribution of amino acid substitutions

To determine which model of amino acid substitution provided the best fit to the data (363-amino-acid sequence of ALDs from 13 vertebrates and the supposed *Ald* phylogeny mentioned above), likelihood ratio tests were conducted among pairs of five models mounted in PAML 3.13d (Yang, 1997). Parameters F and Γ were incorporated in this analysis. As a result, the amino acid substitution matrix JTT (Jones et al., 1992) gave the highest likelihood score (lnL = -6176.35); the second-best matrix was WAG (Whelan and Goldman, 2001) (lnL = -6188.42). Using the JTT matrix, theoretical ratio of the charge-changing substitutions to charge-neutral substitutions was calculated under methods that have been described previously in Chapter 3 (*3.2.7. Calculation of the expected spatial distribution of amino acid substitutions*). The normalized frequencies of each amino acid residue were estimated separately for the surface and interior portions of the inferred common ancestral protein of ALD-A and ALD-C2 to consider differential amino acid composition in different parts of the protein (Table 4-3). Using the theoretical ratio $\Sigma P_{\text{charge-changing}}$:$\Sigma P_{\text{charge-neutral}}$ of the surface (r_1:r_2) and interior (r_3:r_4) portions of the ALD protein molecule under the assumption of random mutation, and the ratio of the numbers of inferred amino acid substitution sites at the surface (190 sites) to the interior (173 sites) in the ALD-A and ALD-C proteins since their duplication, an expected spatial distribution of amino acid substitutions was estimated based on the ratio of charge-changing substitutions in the molecular surface:charge-neutral substitutions in the molecular surface:charge-changing substitutions in the molecular interior:charge-neutral substitutions in the molecular interior = $190r_1$:$190r_2$:$173r_3$:$173r_4$.

Table 4-3.

Inferred normalized amino acid frequencies of the surface and interior portions of the common ancestral protein of ALD-A and ALD-C.

Residue	Normalized frequency	
	Surface	Interior
Ala (A)	0.08421	0.15607
Arg (R)	0.06316	0.01156
Asn (N)	0.05789	0.02890
Asp (D)	0.07895	0.01156
Cys (C)	0.00000	0.04046
Gln (Q)	0.05263	0.03468
Glu (E)	0.11053	0.02890
Gly (G)	0.07368	0.09249
His (H)	0.03158	0.00578
Ile (I)	0.03158	0.10983
Leu (L)	0.03684	0.13295
Lys (K)	0.13684	0.01156
Met (M)	0.01053	0.01156
Phe (F)	0.01053	0.04046
Pro (P)	0.06316	0.02890
Ser (S)	0.04737	0.04046
Thr (T)	0.05263	0.06358
Trp (W)	0.00000	0.01734
Tyr (Y)	0.03158	0.03468
Val (V)	0.02632	0.09827

4.3. Results

4.3.1. Gene duplications and evolutionary relationships of the *Ald* genes in vertebrates

The inferred molecular phylogeny showed that *Ald* genes in jawed vertebrates were group into three major clades, *Ald-A*, *Ald-B*, and *Ald-C* (Fig. 4-1A), suggesting that these *Ald* genes were generated by two gene-duplication events that occurred before the separation of tetrapods and teleost fish, but after the split with agnathans. This supports the idea mentioned in *4.1.Introduction* that the *Ald-A*, *Ald-B*, and *Ald-C* genes of jawed vertebrates, which are mainly expressed in skeletal muscle, liver, and brain, respectively (Gamblin et al., 1991), underwent subfunctionalization with respect to the patterns of gene expression after gene duplication. Thus, the *Ald* gene family in jawed vertebrates is an appropriate model for studying molecular evolution after subfunctionalization.

The molecular phylogeny of *Ald* genes shown in Fig. 4-1A also indicates that the *Ald-A* and *Ald-C* were duplicated specifically in the teleost lineage to give rise to *Ald-A1* and *Ald-A2*, and *Ald-C1* and *Ald-C2*, respectively. These duplications seem to have occurred in common ancestor of the teleosts examined in this study, however, a precise relationship between *Ald-A1* and *Ald-A2* genes remained unclear in the phylogenetic tree presented in Fig. 4-1A. To clarify this point, I exclusively analyzed the molecular phylogeny of the *Ald-A* genes using human *Ald-B* and *Ald-C* as outgroups. The resultant tree presented in Fig. 4-2B clearly showed that the duplication of the *Ald-A* gene has occurred in the common ancestor of the teleosts. Through these analyses, I gained an overview of the relationships and timings of gene duplications of *Ald* genes in jawed vertebrates; before the split between tetrapods and teleosts, the *Ald-B* gene had firstly diverged from the common ancestral gene of the *Ald-A* and *Ald-C*, and subsequently, the latter two genes diverged. Furthermore, after the separation between tetrapods and teleosts, *Ald-A* and *Ald-C* genes underwent specific duplications in the

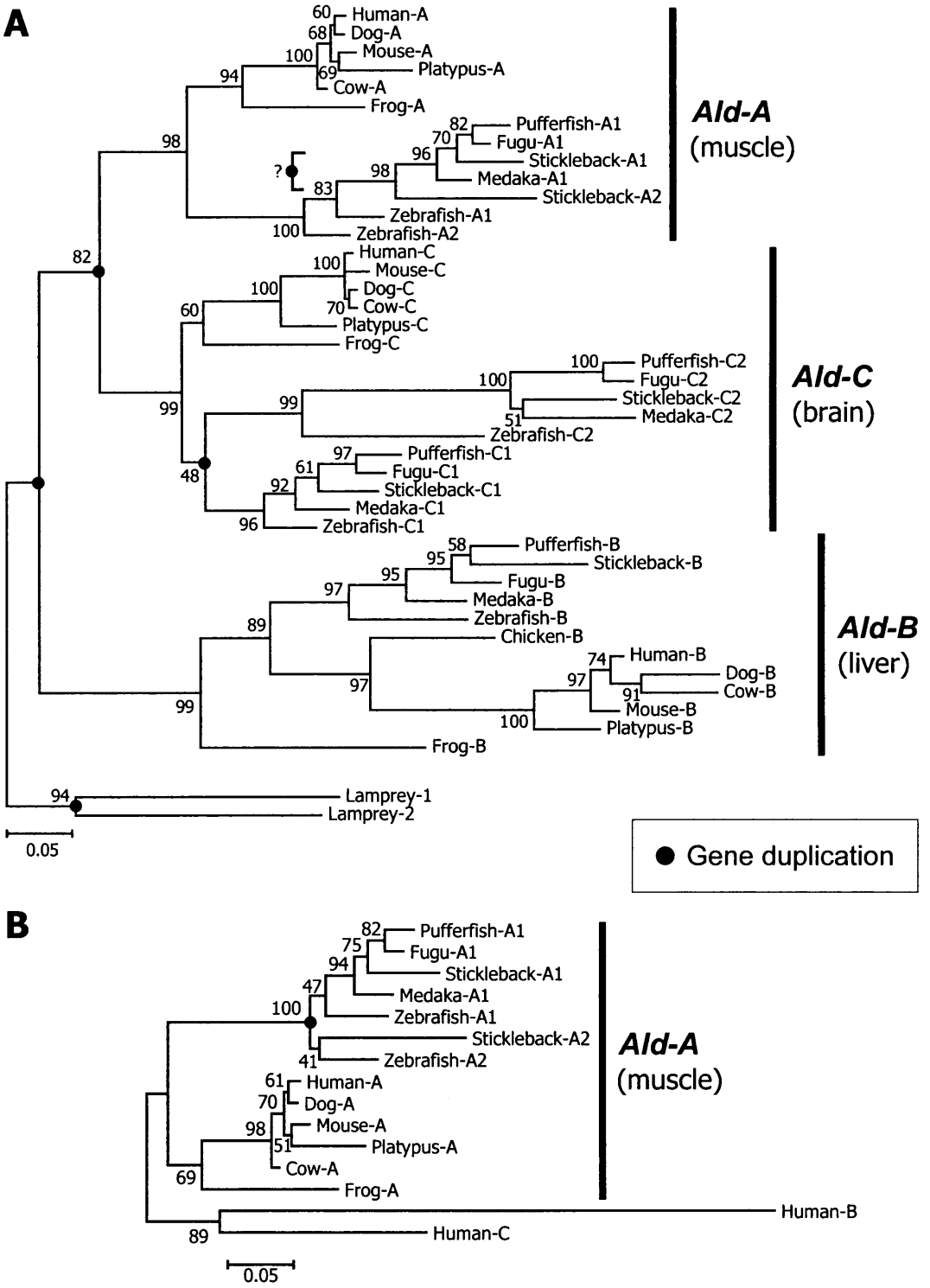


Fig. 4-1. Maximum likelihood tree of *Ald* genes derived from 13 vertebrates. Numbers indicate approximate bootstrap values from 1,000 LR-ELW (the Expected-Likelihood Weights applied to Local Rearrangements of tree topology) tests that support for the nodes. Circles denote inferred gene duplication event. **(A)** Molecular phylogeny of *Ald-A*, *Ald-B*, and *Ald-C*. **(B)** Molecular phylogeny of *Ald-A* by using human *Ald-B* and *Ald-C* as outgroups to root the phylogeny.

teleost lineage, respectively. These results confirmed the origins of each isoform of ALDs in jawed vertebrates.

4.3.2. Evolution of the enzyme active sites and electric charges in ALD isoforms of jawed vertebrates

To clarify differences in the conformation of active enzyme sites among the ALD-A, ALD-B, and ALD-C, which are differentiated by catalytic preferences and expression organs, I compared their amino acid sequences. Among them, most of the proposed amino acid residues involved in the reaction center of ALD enzyme (Dalby et al., 2001; Maurady et al., 2002; St-Jean et al., 2005) were conserved; that are, the Asp33, Glu34, Ser35, Arg42, Lys107, Lys146, Arg148, Glu187, Glu189, Lys229, Leu270, Ser271, Ser300, and Arg303 (Table 4-4). I identified, however, an active enzyme site of ALD that have a unique residue specific to an isoform; all of the ALD-B isoform possess Thr38 whereas most of the ALD-A and ALD-C possess Ser38, showing the existence of an isoform-specific active enzyme residue (Table 4-4). Moreover, it was shown the presence of species-specific active site residues on positions 38 and 356 (usually Ser38 and Leu356, respectively); Frog ALD-A, His38; Frog ALD-C, Asp38; Pufferfish ALD-C2, missing; Fugu ALD-C2, missing; Stickleback-ALD-C2, Tyr356; Medaka ALD-C2, Tyr356; Zebrafish ALD-C2, His356 (Table 4-4).

Next, to characterize the difference in electric charge among the ALD isoforms, biochemical properties were examined using the ProtParam tool (Gasteiger et al., 2005; Table 4-5). The estimated pI values of ALD-A were 8.03–8.61 (average, 8.37), and those of ALD-C were 6.04–6.72 (average, 6.33; excluding teleost-specific ALD-C2), with no overlap ($P < 0.0000$, Mann–Whitney U test, $n_1 = 13$, $n_2 = 11$). The values of the ALD-B (7.52–8.93; average, 8.51) were close to those of ALD-A. In peptide length or predicted overall hydrophobicity, no significant difference was observed among the ALD isoforms (Table 4-5).

Table 4-4.

Variety of enzyme active sites among ALD isoforms from various species.

Reference	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
St-Jean et al. (2005)	D33		S35	S38	R42	K107	K146	R148	E187	E189	K229	L270	S271	S300	R303	
Maurady et al. (2002)	D33	E34			R42	K107	K146	R148	E187	E189	K229	L270	S271	S300	R303	
Dalby et al. (1999)	D33		S35	S38		K107	K146	R148	E187	E189	K229		S271		R303	L356
Conserved?	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N
ALD-B				T38												
Frog ALD-A				H38												
Frog ALD-C				D38												
Pufferfish ALD-C2																Missing
Fugu ALD-C2																Missing
Stickleback ALD-C2																Y356
Medaka ALD-C2																Y356
Zebrafish ALD-C2																H356

Table 4-5.

Biochemical parameters of vertebrate ALD proteins.

	Biochemical characters[a]			No. of hydrophilic charged residues		
	No. of amino acids	Hydro-phobicity (GRAVY)	Isoelectric point (pI)	Positively charged (Arg+Lys)	Negatively charged (Asp+Glu)	Difference[b]
ALD-A						
Human	364	-0.262	8.30	41	38	3
Mouse	364	-0.270	8.30	41	38	3
Dog	364	-0.278	8.30	41	38	3
Cow	364	-0.272	8.45	42	38	4
Platypus	364	-0.299	8.30	40	37	3
Clawed frog	364	-0.286	8.48	43	39	4
Pufferfish	358	-0.313	8.45	42	38	4
Fugu	363	-0.300	8.59	41	36	5
Stickleback-1	363	-0.316	8.61	41	36	5
Stickleback-2	363	-0.270	8.03	42	40	2
Medaka	363	-0.293	8.31	40	37	3
Zebrafish-1	364	-0.281	8.26	43	40	3
Zebrafish-2	364	-0.238	8.49	42	38	4
ALD-B						
Human	364	-0.227	8.01	38	36	2
Mouse	364	-0.262	8.52	39	35	4
Dog	364	-0.234	8.93	41	33	8
Cow	364	-0.270	8.69	40	35	5
Platypus	364	-0.321	8.81	42	36	6
Chicken	364	-0.253	8.81	40	34	6
Clawed frog	364	-0.174	7.52	39	38	1
Pufferfish	361	-0.254	8.48	41	37	4
Fugu	364	-0.241	8.58	40	35	5
Stickleback	364	-0.309	8.76	42	36	6
Medaka	364	-0.324	8.57	41	36	5
Zebrafish	364	-0.246	8.48	42	38	4

ALD-C						
Human	364	-0.241	6.41	41	43	-2
Mouse	363	-0.233	6.67	41	42	-1
Dog	364	-0.232	6.21	40	43	-3
Cow	364	-0.220	6.21	40	43	-3
Platypus	364	-0.248	6.28	40	43	-3
Clawed frog	364	-0.231	6.04	39	43	-4
Pufferfish-1	362	-0.249	6.72	38	39	-1
Pufferfish-2	334	-0.207	7.64	38	37	1
Fugu-1	363	-0.266	6.41	39	41	-2
Fugu-2	334	-0.177	6.80	37	38	-1
Stickleback-1	363	-0.338	5.88	38	43	-5
Stickleback-2	363	-0.318	8.09	41	39	2
Medaka-1	363	-0.247	6.67	39	40	-1
Medaka-2	364	-0.275	8.03	42	40	2
Zebrafish-1	363	-0.284	6.21	38	41	-3
Zebrafish-2	364	-0.252	6.76	40	41	-1
ALD						
Lamprey-1	363	-0.219	8.84	42	37	5
Lamprey-2	364	-0.167	8.63	42	38	4

[a]The predicted overall hydrophobicity (GRAVY; grand average of hydropathicity) and pI values of the ALD proteins were estimated based on the amino acid sequences translated from the cDNA sequences of *Ald* genes using the ProtParam tool (Gasteiger et al., 2005).

[b]Differences in number of positively and negatively charged residues in each ALD protein.

Between ALD-A and ALD-C, which were remarkably different in terms of electric charge, 64 amino acid sites differed by the presence or absence of hydrophilic charged residues [Lys (K), Arg (R), Asp (D), and Glu (E)], which mainly contribute to net protein charge (Fig. 4-2B). I found only two amino acid sites, which were fixed for a unique charge state in the examined ALD-A or ALD-C proteins; position 314 (Gly in ALD-A and Lys or Arg in ALD-C) and 332 (Leu or Val in ALD-A and Glu in ALD-C). Furthermore, only a small number of unique charged sites were shared among two or more genealogically related isoforms: four in mammalian ALD-C (87, 98, 119, and 156), two in ALD-C1 (positions 57 and 318), and three in ALD-C2 (positions 9, 10, and 143). These observations imply that very few amino acid residues were acquired specifically to the ancestral proteins of ALD-A and/or ALD-C, and involved in differences in electric charge between current ALD-A and ALD-C.

Then, I estimated the underlying process of the electric charge evolution of the ALD proteins by using ancestral sequence reconstructions based on maximum-likelihood inference. This method using the program BASEML (Yang, 1997) was applied on the basis of the inferred phylogeny of *Ald* genes (Fig. 4-1) and known phylogeny of mammals (Murphy et al., 2004) and teleosts (Yamanoue et al., 2006), and a nucleotide substitution model that best fit the data (GTR + Γ; Yang, 1994; see Table 4-2). The result was shown in Fig. 4-2A; this suggests that pI value (electric charge) of a common ancestral protein of ALD-A and ALD-C was relatively negative (6.65) similar to that of current ALD-C, and pI values of ALD-A increased independently in both the tetrapod and teleost fish lineages.

To clarify the detailed process of the electric charge evolution in ALD isoforms, I assigned charge-changing substitutions (between Lys/Arg or Asp/Glu and other residues) to the tree branches based on pairwise sequence comparisons among the inferred ancestral sequences or between the extant and ancestral sequences along the tree topology (Fig. 4-2C). This result of assignment showed that the charge-changing substitutions were inferred to have occurred in excess (10 to 22 substitutions assigned from duplication between *Ald-A* and *Ald-*

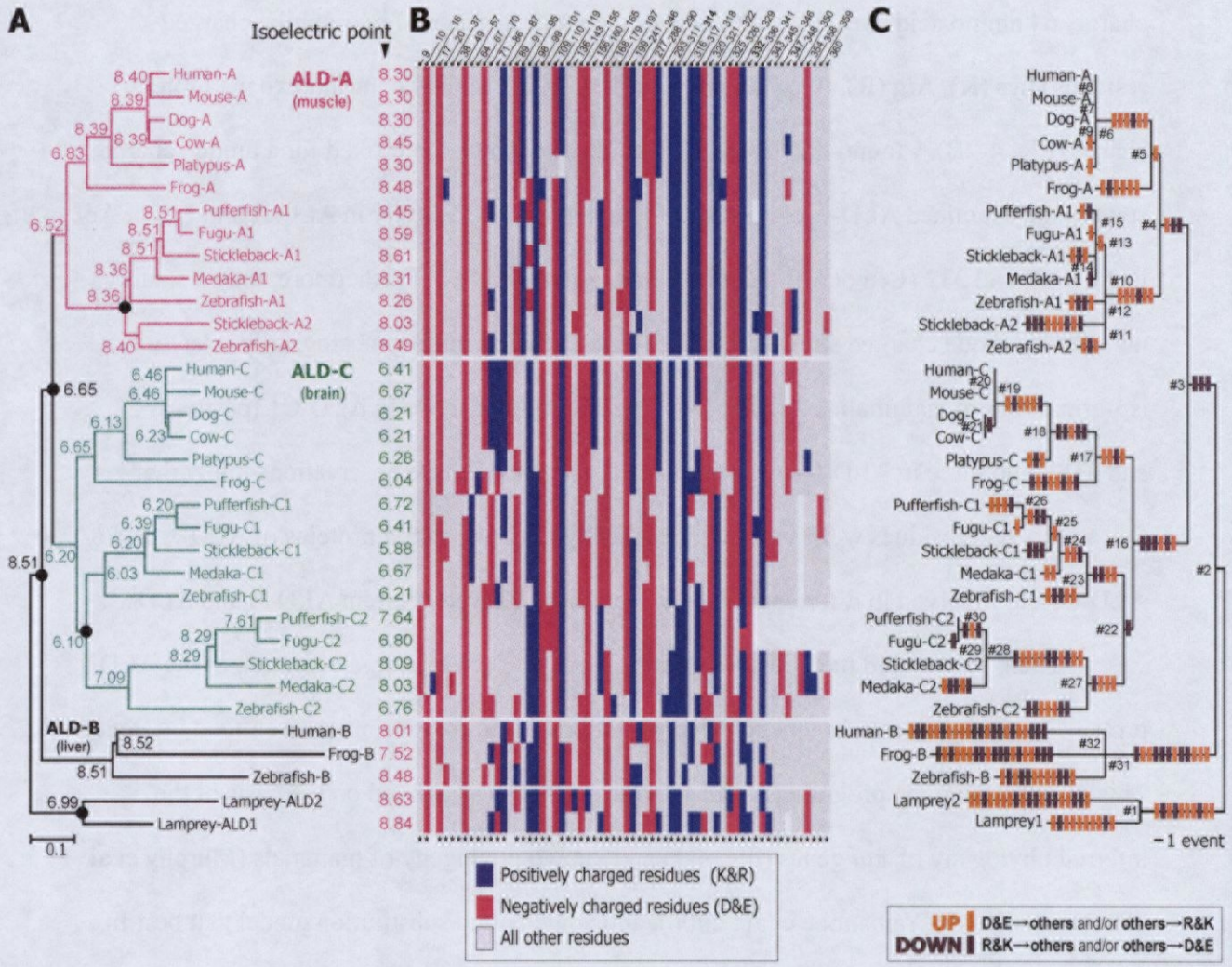


Fig. 4-2. Current states and inferred evolutionary process of electric charge of ALD isoforms. **(A)** Maximum likelihood tree of *Ald* genes in ray-finned fishes inferred by BASEML (Yang, 1997) with known phylogeny (Murphy et al., 2004; Yamanoue et al., 2006). Numbers indicate estimated pI. Circles denote gene duplication. **(B)** Amino acid sites that differ by the presence or absence of hydrophilic charged residues between current ALD-A and ALD-C. Positively charged residues are colored blue; negatively charged residues, red; other residues, light grey; gapped sites, white. The numbers above refer to the amino acid positions of human ALD-A (Dalby et al., 1999). The stars below indicate sites located on the molecular surface. **(C)** Inferred charge-changing substitution events mapped over the ALD phylogeny. Orange and brown bars denote upward and downward direction of charge change, respectively.

C [#3] to tips of the branches; except for ALD-C2) of that expected for parsimonious evolution in electric charge differences between ALD-A and ALD-C (average 5.1 substitutions correspond to minimum evolution of the charge differences; Fig. 4-2C; see also Table 4-5).

4.3.3. Statistical analyses of the spatial clustering of inferred amino acid substitutions

To confirm whether the inferred charge-changing substitutions shown in Fig. 4-2C were actually involved in the evolution of electric charge, further analyses were performed based on 3-D structural information on the ALD protein molecule (Dalby et al., 1999). The spatial locations of the charge-changing and charge-neutral substitution sites mapped on the 3D protein structure were shown in Fig. 4-3A. This shows that the inferred charge-changing substitution sites since the duplication between *Ald-A* and *Ald-C* genes (colored in magenta) were concentrated at the surface of the ALD molecule ($P<0.0001$, two-tailed Fisher's exact test), in contrast to the inferred charge-neutral substitution sites (colored in black; $P=0.0242$, two-tailed Fisher's exact test). The inferred number of charge-changing and charge-neutral substitutions that can potentially occur at identical site classes also followed the same trend (Fig. 4-3B); the proportion of charge-changing substitutions per all substitutions significantly increases with SASA ($Å^2$) ($P<0.0001$, Cochran-Armitage trend test, $n=677$).

However, as described in Chapter 3 (*3.3.3. Statistical analyses of the spatial clustering of inferred amino acid substitutions*), because water-soluble proteins are generally surrounded by a hydrophilic shell containing a high density of polar residues, it is natural to expect that random mutations cause the charge-changing substitutions to occur more frequently on the protein surface without any selection. Considering this expected mutation bias, further analysis was performed (Table 4-6; for details, see section 4.2.4. of Materials and Methods). This comparison of theoretically expected and ML-inferred numbers of amino acid

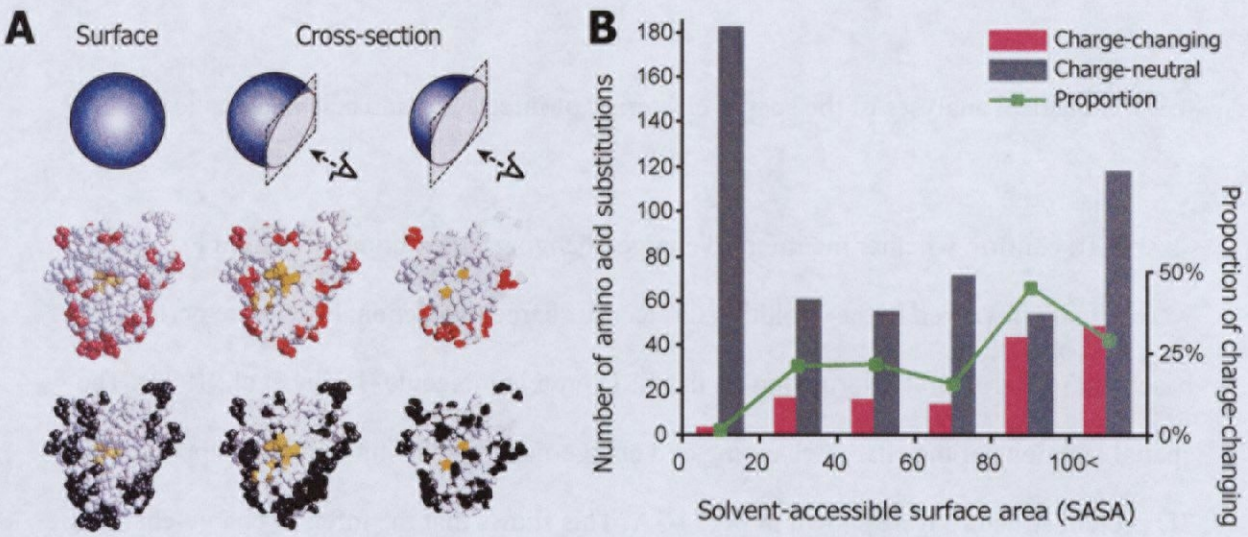


Fig. 4-3. Spatial locations of inferred amino acid substitutions in ALD structure. **(A)** ML-inferred charge-changing (CC) substitution sites after the duplication between *Ald-A and Ald-C* genes are colored magenta; charge-neutral (CN) substitution sites, dark grey; enzyme active sites, yellow. Full molecular models are shown on the left, and two cross-sections are shown center and right. The inferred CC sites localize to the surface of the ALD molecule (61 CC sites / 190 total surface sites, 3 CC sites / 173 total interior sites; $P<0.0001$, two-tailed Fisher's exact test), in contrast to the inferred CN sites (105 CN sites / 190 total surface sites, 62 CN sites / 173 total interior sites; $P=0.0242$, two-tailed Fisher's exact test). **(B)** Histograms of the inferred numbers of the CC and CN substitutions after the duplication between *Ald-A* and *Ald-C* genes. The solid green line denotes the proportion of the CC substitutions per all substitutions within the site classes based on solvent accessibility (horizontal axis): this proportion significantly increases with solvent-accessible surface area ($P<0.0001$, Cochran-Armitage trend test, $n=677$).

substitutions imply that charge-neutral substitutions have occurred more frequently than expected at the molecular surface [ML-inferred value, 66.2% = 357/(182 + 357); expected value, 46.2% = 233.88/(272.87 + 233.88)], consistent with the general observation that molecular evolutionary rates are faster at the surface than in the interior portions of globular proteins (Bustamante et al., 2000; Choi et al., 2006). However, the more important point in this table is that almost all of the charge-changing substitutions were occurred at the surface of the ALD molecule [ML-inferred value, 97.8% = 135/(3 + 135)] greater than expected by chance [expected value, 70.8% = 120.47/(49.78 + 120.47)].

Table 4-6.

Analytically inferred and theoretically predicted numbers of charge-changing and charge-neutral substitutions.

	Charge-changing		Charge-neutral		Sum
	Interior	Surface	Interior	Surface	
Maximum likelihood-inferred numbers	3	135	182	357	677
Theoretical prediction	49.78	120.47	272.87	233.88	677.00
P value*	0.00000		0.00000		

* *P* values are from two-tailed exact tests.

4.4. Discussion

4.4.1. Importance of weak selection on many amino acid sites in adaptive protein evolution

The analysis of molecular phylogeny (Fig. 4-1) and biochemical properties (Table 4-5) of ALD proteins in jawed vertebrates suggest that since the origination of ALD-A and ALD-C by gene duplication in the ancestor of jawed vertebrates, electric charges of the two isoforms diverged. Considering that ALD-A and ALD-C are mainly expressed in skeletal muscle and brain, respectively (Gamblin et al., 1991), the divergent evolution of the electric charges of the ALD-A and ALD-C seems to be a consequence of specialization for distinct cellular environment of tissues where each isoform is predominantly expressed, as suggested for PGI proteins (Chapter 3) and several other proteins (Merritt and Quattro, 2001; Zhang, et al., 2002; Zhang, 2006). Therefore, the charge evolution of ALD-A and ALD-C can also be understood under the sub-neofunctionalization model of duplicate gene evolution (Force et al., 1999; Lynch and Force, 2000A; He and Zhang, 2005; Rastogi and Liberles, 2005), which proposes that the partitioning of function between the duplicated genes differentiates selection pressure on each gene locus, resulting in structural and/or functional specialization of the encoded proteins by natural selection.

The inferred evolutionary process of electric charge of ALD-A and ALD-C suggests that relatively weak selection on a large number of amino acid sites has driven the evolution of novel charge-state of ALD isoforms, similar to the PGI analyzed in the Chapter 3. Because only a few charged amino acid sites were specific to ALD-A or ALD-C (Fig. 4-2B), and the electric charge of ALD-A was estimated to have increased independently among lineages, that is, tetrapods and teleosts (Fig. 4-2A). And since the duplication of *Ald-A* and *Ald-C* genes, charge-changing substitutions have occurred in both directions of charge change and more frequently than expected for parsimonious evolution in electric charge differences (Fig. 4-2C),

whereas such charge-changing substitutions concentrated at the surface of ALD molecule (Fig. 4-3). In addition, no significant acceleration of the rate of nonsynonymous substitution was detected during the evolution of these ALD isoforms (Merritt and Quattro, 2002). These data, altogether, suggest that the electric charge of ALD isoforms in jawed vertebrates has been changed, not by strong selection on particular amino acid sites, but by weak selection on many amino acid sites, and thus, the generality of this mode of protein evolution was confirmed.

4.4.2. A possible relationship between the number of modifiable amino acid sites and strength of selection pressure on individual amino acid sites

Besides the mode of protein evolution by weak selection on many amino acid sites as presented above, there also exists a well-defined mode of protein evolution by strong selection on particular amino acid sites, as presented in many previous studies. The latter mode of selection is also estimated to have driven the adaptive evolution of protein charges by some previous studies. One famous example is the evolution of RNASE1 in primates (Zhang et al., 2002; Zhang, 2006). The genes encoding this protein were independently duplicated in Asian and African leaf-eating monkeys (the douc langur *Pygathrix nemaeus* and the guereza *Colobus guereza*, respectively), generating the RNASE1B, RNASE1β, and RNASE1γ (Fig. 4-4A). These extra RNASE1 proteins are involved in the digestion of foregut-fermenters in a ruminant-like alimentary system characteristic to leaf-eating monkeys, and exhibit remarkably low values of pI (see Fig. 4-4A). Such negative charges appear to be attributed to amino acid sites specific to the RNASE1B, RNASE1β, and RNASE1γ (position 4, 6, and 39; indicated in bold face in Fig. 4-4A). Another example is TPI-A (triose-phosphate isomerase A), which is a brain type isoform specific to teleost fish. TPI-A also has distinctively lower pI value than TPI-B in teleosts (see Fig. 4-4B), and this negative charge appears to be attributed to amino

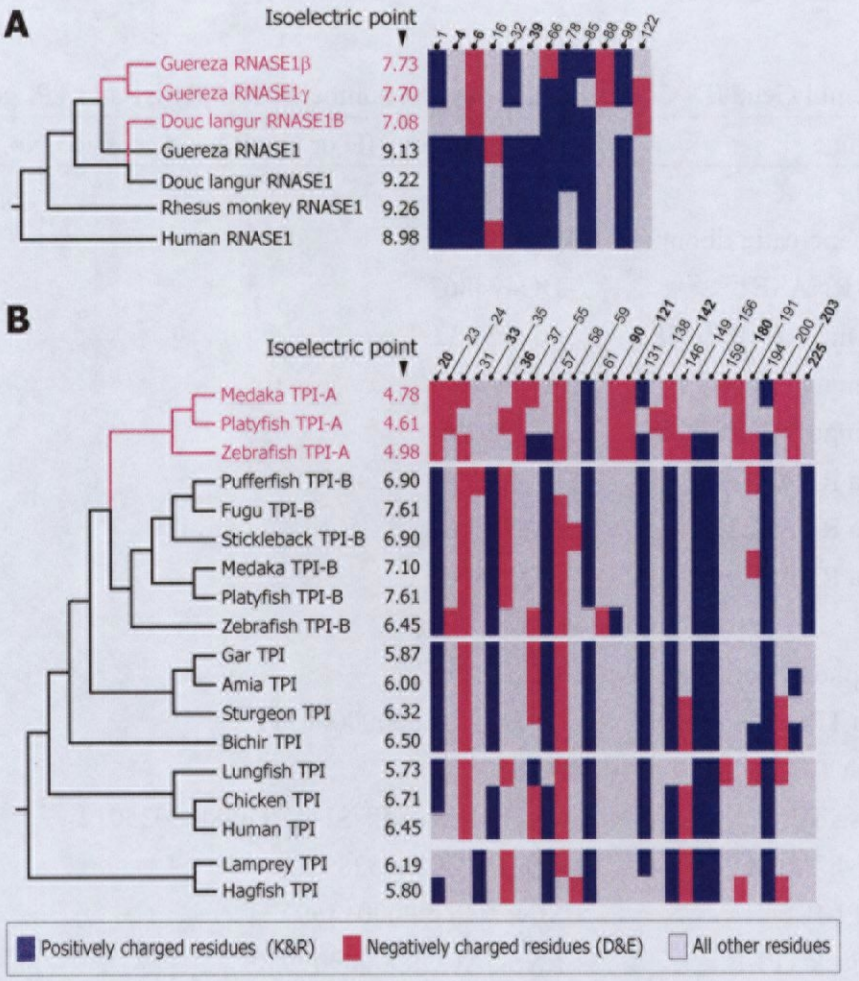


Fig. 4-4. Amino acid sites that differ by the presence or absence of hydrophilic charged residues between differentially charged protein isoforms. Positively charged residues are colored blue; negatively charged residues, red; other residues, light grey. The numbers above refer to the amino acid positions; numerals in bold refer to the amino acid sites that have a charge state specific to negatively-charged isoforms (lettered in magenta). **(A)** Primate RNASE1 (pancreatic ribonuclease) and digestive RNASE1B/RNASE1β/RNASE1γ, which are specific to leaf-eating monkeys. The topology of the gene phylogeny shown in the left side of the figure was estimated by Zhang (2006). **(B)** TPI (triose-phosphate isomerase)-A and TPI-B of teleost fish. The gene phylogeny shown in the left side of the figure is according to TPI phylogeny (Merritt and Quattro, 2001) and known phylogeny of ray-finned and lobefin fish (Inoue et al., 2003; Kikugawa et al., 2004). GenBank accession numbers of the sequences are listed in Table 4-7.

Table 4-7.

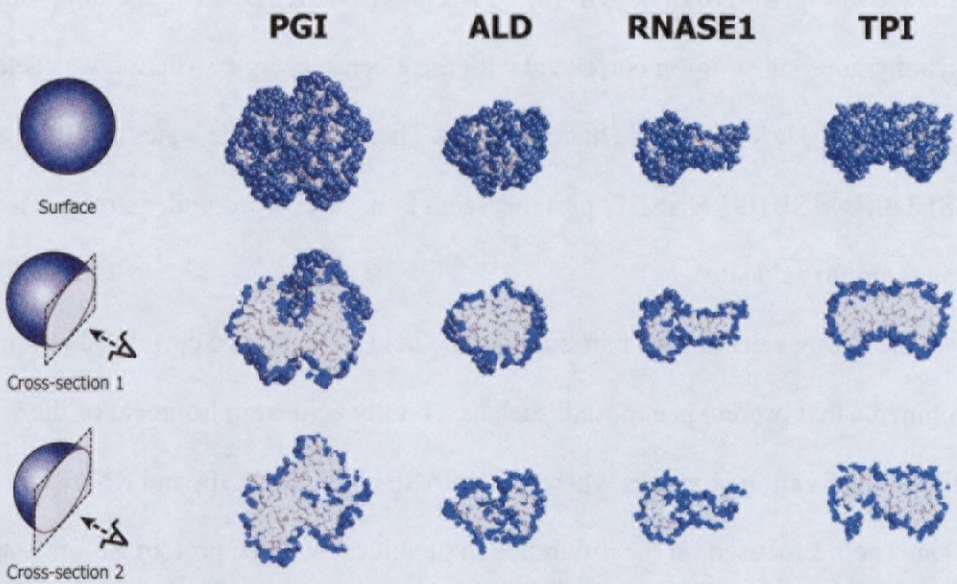
List of Ensembl Gene IDs or GenBank accession numbers of RNASE1 and TPI genes.

Gene name	Ensembl Gene ID or GenBank Accession No.
RNASE1 (pancreatic ribonuclease, EC 3.1.27.5)	
Human RNASE1	DQ494867
Rhesus monkey RNASE1	AF449632
Douc langur RNASE1	AF449642
Douc langur RNASE1B	AF449643
Guereza RNASE1	DQ516063
Guereza RNASE1β	DQ516064
Guereza RNASE1γ	DQ516065
TPI (triose-phosphate isomerase, EC 5.3.1.1)	
Medaka TPI-A	ENSORLG00000006391
Platyfish TPI-A	AF387822
Zebrafish TPI-A	AF387820 or ENSDARG00000025012
Pufferfish TPI-B	GSTENG00005289001
Fugu TPI-B	SINFRUG00000146972
Stickleback TPI-B	ENSGACG00000009950
Medaka TPI-B	ENSORLG00000012599
Platyfish TPI-B	AF387821
Zebrafish TPI-B	AF387819 or ENSDARG00000040988
Sturgeon TPI	AF387818
Gar TPI	AB111394
Amia TPI	AB111401
Bichir TPI	AB111416
Lungfish TPI	AB111366
Human TPI	M10036 or ENSG00000111669
Chicken TPI	M11941 or ENSGALG00000014526
Lamprey TPI	AB025327
Hagfish TPI	AB025322

acid residues specific to TPI-A (position 20, 33, 36, 90, 121, 142, 180, 203, and 225; indicated in bold face in Fig. 4-4B). In both of the TPI-A and RNASE1B, an acceleration of the rate of non-synonymous substitution correlated with the electric charge evolution was detected (Zhang et al., 2002; Merritt and Quattro, 2001). Therefore, charge states of TPI-A and RNASE1B/RNASE1β/RNASE1γ proteins seem to have evolved under strong selection on particular amino acid sites.

The strong selection on particular amino acid sites may be correlated to immoderate environments that protein is exposed, such as a highly acidic environment of the small intestine of leaf-eating monkeys where the RNASE1B, RNASE1β, and RNASE1γ proteins function. Then, I focused on the difference in pI values between protein isoforms as an indicator of selection pressure on protein charges. The difference in pI, however, was not remarkably large in RNASE1 and TPI protein families, in which particular charged residues were strongly selected during evolution; the difference in estimated pI (in average) was 1.64 between RNASE1B/RNASE1β/RNASE1γ and RNASE1, 2.31 between TPI-A and TPI-B, 1.87 between ALD-A and ALD-C, and 0.83 between PGI-1 and PGI-2. These observations imply that there is little association between strength of selection on net protein charge and that on individual amino acid sites.

Further analysis suggested that absolute number of amino acid sites available for evolutionary modification determines the strength of selection pressure on individual amino acid sites. When few amino acid sites can be changed, then selection pressure per site becomes strong, and when many amino acid sites can be changed, then selection pressure per site becomes weak. As shown in Fig. 4-5, the PGI and ALD, which evolved through weak selection on many amino acid sites, seem to harbor larger number of amino acid sites in their surface portion of a protein than RNASE1 and TPI, which evolved through strong selection on particular amino acid sites. Furthermore, the surface portion of the PGI and ALD appears to possess more amino acid residues that have higher probability to be involved in charge-



	PGI	ALD	RNASE1	TPI
Formation:	Dimer	Monomer	Dimer	Dimer
No. of surface sites:	234	190	87	133
No. of interior sites:	316	173	35	115
Total no. of sites:	550	363	122	248

Fig. 4-5. Spatial locations and numbers of amino acid sites that comprise the surface portion of the protein molecules (colored in blue). Amino acid residues with more than 20 angstrom-square of solvent-accessible surface area (SASA) were considered to comprise the molecular surface. SASA values of each amino acid residue were estimated by using GETAREA 1.1 (Fraczkiewicz and Braun, 1998) with a solvent radius of 1.4 angstrom (approximately the size of a water molecule). Reference protein structures used are as follows: PGI (phosphoglucose isomerase): Protein Data Bank (PDB) ID code 1XTB (Lee and Jeffery, 2005); ALD (fructose-1,6-biphosphate aldolase): PDB ID code 4ALD (Dalby et al., 1999); RNASE1 (pancreatic ribonuclease): PDB ID code 1DZA (Pous et al., 2000); TPI (Triosephosphate isomerase): PDB ID code 1WYI (Kinoshita et al., 2005).

Table 4-8.

Probabilities of involvement in charge-changing substitutions of each amino acid residue, and amino acid compositions of surface portion of the respective protein molecules.

Residue	$\Sigma_{(charge\text{-}changing)}$[a]	No. of amino acids in surface portion of the protein molecule			
		PGI	ALD	RNASE1	TPI
Arg	0.1510	16	13	9	7
Lys	0.1317	32	24	7	19
Glu	0.1241	26	19	7	13
Asp	0.1239	18	12	2	9
Gln	0.0989	17	11	6	9
Asn	0.0900	13	9	9	6
Gly	0.0552	12	14	6	12
His	0.0403	14	8	4	2
Ala	0.0389	9	15	4	11
Ser	0.0305	17	13	11	9
Thr	0.0285	18	11	4	8
Pro	0.0140	13	14	4	6
Val	0.0136	4	3	6	8
Met	0.0118	2	2	1	0
Leu	0.0105	7	9	2	4
Trp	0.0104	7	0	0	4
Cys	0.0101	1	0	1	0
Ile	0.0074	3	3	0	2
Tyr	0.0072	2	6	3	0
Phe	0.0019	3	4	1	4
Total no. of amino acid sites		234	190	87	133

[a]Relative probabilities of involvement in charge-changing substitutions estimated on the basis of transition rates between pairs of amino acids and relative mutabilities of each amino acid residue described in the amino acid substitution matrix JTT (Jones et al., 1992).

changing substitutions (Table 4-8). This suggests that, during the evolution of novel protein properties, the total number of amino acid sites available for evolutionary change governs the mode of adaptive molecular evolution; that is, whether weak selection on many amino acid sites or strong selection on particular amino acid sites predominates. This concept presented here will provide a broader understanding of the principles of the evolution of new genes and protein properties.

Chapter 5.

General discussion

5.1. Contributions of this study to understanding evolution of fish genomes and novel proteins

To elucidate how novel genes and proteins can arise during evolution, the genes duplicated through fish-specific genome doubling, or 3R-WGD, have been analyzed from various viewpoints in this study, and the results provided several new findings. Firstly, in Chapter 2, more than 100 gene families were systematically analyzed based on the whole-genome sequences from multiple fish species and advanced maximum-likelihood methods for phylogenetic inference. This revealed that many of duplicated genes were rapidly lost after generation by the 3R-WGD, while the genes remained comprise about 40% of gene loci ingenomes of extant teleost fishes. This temporal process of gene loss was interpreted that the 3R-WGD-derived duplicated genes which persisted for a long evolutionary periods of time were maintained in the genome primarily through sunfunctionalization, and this interpretation was supported by comparative analyses of several characteristics of the genes, such as a number of interaction partners in the above organic systems. To my knowledge, this is the first assessment of the fate of a large number of gene families following the 3R-WGD event based on the multiple genome sequences of teleosts. Because most previous studies on genome evolution of fishes, especially that focusing on the 3R-WGD, have investigated only a particular gene family or a single genome of a fish species (e.g., Hoegg and Meyer, 2007; Woolfe and Elgin, 2007; Yu et al., 2007). Through the analysis of this study, therefore, it was firstly determined the process of loss of duplicated genes after the 3R-WGD. The inferred process appeared to be compatible with the sub-neofunctionalization model of duplicate gene evolution, and thus be a new evidence for the concept.

In the next step in Chapter 3, comparative evolutionary analysis of fish-specific duplicate *Pgi* genes derived by the 3R-WGD has highlighted the role of weak selection on many amino acid sites in adaptive evolution of genes and proteins. On the other hand, the main concern in the study of adaptive molecular evolution was to seek positive selection on individual amino acid substitutions or particular amino acid sites (e.g., Gu and Vander Velden, 2002; Bielawski and Yang, 2005; Zhang et al., 2005). Therefore, the above finding presented in this study should extend our understanding of molecular evolution. The generality of this newly identified mode of adaptive molecular evolution was confirmed by the analysis of other duplicate genes in vertebrates, *Ald*, in Chapter 4. Furthermore in this chapter, examinations of tertiary structures, amino acid compositions, and evolutionary patterns of several proteins have led to the concept that the total number of amino acid sites available for evolutionary modification in a protein governs strength of selection pressure on amino acid sites; that is, whether weak selection on many sites or strong selection on a few particular sites predominates. This concept can explain both previously known and newly found mode of adaptive molecular evolution presented in this study. Altogether, these findings will contribute to our knowledge about the evolution of fish genomes, as well as the origin and mechanisms of evolution of novel genes and proteins.

5.2. Evolutionary changes are not always caused by major mutations in a particular gene

The mode of adaptive evolution of genes and proteins proposed in this study was difficult to find by existing methods for molecular evolutionary studies, which seek to find a mutation that is expected to play a major role in evolutionary change. These existing methods are based on the neutral theory of evolution, in which mutations are generally divided into three groups based on their supposed or observed effects: adaptive, neutral, and deleterious

mutations. In the case shown here, however, individual amino acid substitutions were not clearly divided into neutral and adaptive mutations, because various amino acid sequences were potentially relevant to a certain selective pressure on a property of protein. In other words, an evolutionary change in protein property was not always attributed to particular amino acid substitutions. Probably for this reason, analytical methods that seek signals of adaptive evolution by focusing on only nucleotide and amino acid sequences were sometimes incapable of resolving the relationships between the evolution of amino acid sequences and that of protein properties, as shown in this study. The findings of this study thus may give us a new insight into molecular evolution. The analytical approach for investigating adaptation at a molecular level, focusing on both "higher hierarchical" and "lower hierarchical" information such as protein properties and amino acid sequences, will be important to study molecular evolution, and also the evolution of molecular interaction networks and macrophenotypic characters.

5.3. Relationship between diversity of fishes and the fish-specific genome doubling

It has been hypothesized that the fish-specific genome doubling or 3R-WGD, which contributed largely to the formation of teleost genomes as shown by this study, has promoted speciation among teleosts, leading to an increase in species diversity of teleosts (Vogel, 1998; Meyer and Malaga-Trillo, 1999; Meyer and Schartl, 1999; Taylor et al., 2001; Taylor et al., 2003). This hypothesis is mainly based on the theory that differential evolution of duplicated genes among populations can cause reproductive incompatibility by the reduction in fitness of hybrids, resulting in speciation (Ferris et al., 1979; Werth and Windham, 1991; Lynch and Force, 2000B; Lynch, 2002). Extending this model, it can be assumed that a whole genome doubling, which produces an enormous number of gene duplicates, provides numerous opportunities for divergent evolution of duplicated genes and thus facilitates speciation events.

It is also proposed that a genome doubling may reduce risk of extinction via functional redundancy and mutational robustness of genes, and accordingly, increases species diversity (Crow and Wagner, 2006). These notions imply a causal relationship between the 3R-WGD and species diversity in teleosts.

However, the number of species of Teleostei (about 27,000) is almost the same as that of Sarcopterygii (about 27,000 including tetrapods), which is a sister group to ray-finned fish including teleosts and basal non-teleosts (see Fig. 1-1B; Nelson, 2006). Therefore, we cannot conclude that the number of species of teleosts, which have experienced the 3R-WGD, is much larger than that of sarcopterygians including tetrapods, which have not experienced the 3R-WGD. On the other hand, some authors focused on only ray-finned fishes, and pointed out that basal non-teleosts (about 50 species including Polypteriformes, Acipenseriformes, Semionotiformes, and Amiiformes), which may not have experienced the 3R-WGD, are poor in number of species compared with teleosts, and thus the authors have proposed a causal relationship between the 3R-WGD and species diversity in teleosts (Hoegg and Meyer, 2005; Meyer and Van de Peer, 2005). Arguing against this notion, a study based on the fossil record cautioned that teleosts are not necessarily more "diverse" or "species-rich" than non-teleosts, when extinct lineages are taken into account (Donoghue and Purnell, 2005). Taken together, the 3R-WGD may not be associated with species diversity in teleosts.

On the other hand, my study suggested that the 3R-WGD has contributed to genetic differences among teleosts via lineage-specific loss or sub-neofunctionalization of a large number of duplicated genes. This implies that the 3R-WGD contributed not merely to morphological diversity, but rather to genetic and physiological diversity of teleosts. Such effect of the 3R-WGD may underlie adaptive radiation of teleosts that inhabit a broad range of ecological zones—from marine to fresh water, from deep sea to the shallow sea environments, and from equatorial to polar regions. On the basis of this view, to characterize the difference between teleosts and non-teleosts in relation to genetic effect of the 3R-WGD will be

particularly valuable for understanding the evolution of teleosts, and consequently, the evolution of vertebrates that have experienced WGD events (see Fig. 1-1A). This concern needs to be addressed for future studies of fish genomes. Such genomic studies will also provide useful information on sound conservation and management of fish stocks. For example, this study suggested that the sticklebacks have a specific, possibly adaptive feature in olfactory transduction system (see chapter 2). Such knowledge will promote to develop a proper conservation program for sticklebacks, avoiding disturbance of the natural chemical environment of inhabitats of sticklebacks.

Acknowledgements

I am grateful to Prof. Mutsumi Nishida for generous encouragement and advice throughout this study, and Assoc. Prof. Susumu Hyodo, Assoc. Prof. Shoji Kawamura, Prof. Hirohisa Kishino, and Prof. Kazuhiro Kogure (in alphabetical order) for review and helpful comments on this dissertation. I thank lab members, particularly Drs. Yoichiro Azuma, Jun G. Inoue, Yusuke Yamanoue, Takashi P. Satoh, Kaoru Kuriiwa, Hirohiko Takeshima, Yasuyuki Hashiguchi for aid and information. I also thank Prof. Toshiaki Miyadai of Fukui Prefectural University for a specimen of Fugu; Dr. Akihiro Okamura of IRAGO Institute Co., Ltd., for a European eel specimen; and Atsunobu Murase Tokyo University of Marine Science and Technology for a mullet specimen.

Literature Cited

Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104–2105.

Abi-Rached L, Gilles A, Shiina T, Pontarotti P, Inoko H. 2002. Evidence of *en bloc* duplication in vertebrate genomes. Nat. Genet. 31:100–105.

Adams KL, Wendel JF. 2005. Polyploidy and genome evolution in plants. Curr. Opin. Plant Biol. 8:135–141.

Alsop E, Silver M, Livesay DR. 2003. Optimized electrostatic surfaces parallel increased thermostability: a structural bioinformatic analysis. Protein Eng. 16:871–874.

Altschmied J, Delfgaauw J, Wilde B, Duschl J, Bouneau L, Volff JN, Schartl M. 2002. Subfunctionalization of duplicate *mitf* genes associated with differential degeneration of alternative exons in fish. Genetics 161:259–267.

Amores A, Suzuki T, Yan YL, Pomeroy J, Singer A, Amemiya C, Postlethwait JH. 1998. Zebrafish *hox* clusters and vertebrate genome evolution. Science 282:1711–1714.

Amores A, Suzuki T, Yan YL, Pomeroy J, Singer A, Amemiya C, Postlethwait JH. 2004. Developmental roles of pufferfish *Hox* clusters and genome evolution in ray-fin fish. Genome Res. 14:1–10.

Aparicio S, Hawker K, Cottage A, Mikawa Y, Zuo L, Venketesh B, Chen E, Krumlauf R, Brenner S. 1997. Organization of the *Fugu rubripes Hox* clusters: evidence for continuing evolution of vertebrate *Hox* complexes. Nat. Genet. 16:79–83.

Avise JC, Kitto GB. 1973. Phosphoglucose isomerase gene duplication in the bony fishes: an evolutionary history. Biochem. Genet. 8:113–132.

Bielawski JP, Yang Z. 2005. Maximum likelihood methods for detecting adaptive protein evolution. In: Nielsen R, editors. Statistical Methods in Molecular Evolution. New York: Springer-Verlag. p. 103–124.

Birney E, Clamp M, Durbin R. GeneWise and Genomewise. 2004. Genome Res. 14:988–995.

Birney E, Andrews D, Caccamo M, et al. (51 co-authors). 2006. Ensembl 2006. Nucleic Acids Res. 34(Database issue):D556–D561.

Blanc G, Wolfe KH. 2004. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. Plant Cell. 16:1679–1691.

Borenstein E, Shlomi T, Ruppin E, Sharan R. 2007. Gene loss rate: a probabilistic measure for the conservation of eukaryotic genes. Nucleic Acids Res. 35:e7.

Bustamante CD, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD, Hartl DL. 2000. Solvent accessibility and purifying selection within proteins of *Escherichia coli* and *Salmonella enterica*. Mol. Biol. Evol. 17:301–308.

Choi SS, Vallender EJ, Lahn BT. 2006. Systematically assessing the influence of 3-dimensional structural context on the molecular evolution of mammalian proteomes. Mol. Biol. Evol. 23:2131–2133.

Chiu CH, Amemiya C, Dewar K, Kim CB, Ruddle FH, Wagner GP. 2002. Molecular evolution of the *HoxA* cluster in the three major gnathostome lineages. Proc. Natl. Acad. Sci. U.S.A. 99:5492–5497.

Chiu CH, Dewar K, Wagner GP, et al. (13 co-authors). 2004. Bichir *hox*A cluster sequence reveals surprising trends in ray-finned fish genomic evolution. Genome Res. 14:11–17.

Christoffels A, Koh EG, Chia JM, Brenner S, Aparicio S, Venkatesh B. 2004. Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. Mol. Biol. Evol. 21:1146–1151.

Christoffels A, Brenner S, Venkatesh B. 2006. *Tetraodon* genome analysis provides further evidence for whole-genome duplication in the ray-finned fish. Comp. Biochem. Physiol. Part D Genomics Proteomics 1:13–19.

Crow KD, Wagner GP. 2006. What is the role of genome duplication in the evolution of complexity and diversity? Mol. Biol. Evol. 23:887–892.

Dalby AR, Tolan DR, Littlechild JA. 2001. The structure of human liver fructose-1,6-bisphosphate aldolase. Acta Crystallogr. D Biol. Crystallogr. 57:1526–1533.

Dando PR. 1980. Duplication of the glucosephosphate isomerase locus in vertebrates. Comp. Biochem. Physiol. 66:373–378.

Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary change in proteins. In: Dayhoff MO, editors. Atlas of Protein Sequence and Structure 5, Supp. 3. Silver Springs, Maryland: National Biomedical Research Foundation. p. 345–352.

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. 3:e314.

de Souza FS, Bumaschny VF, Low MJ, Rubinstein M. 2005. Subfunctionalization of expression and peptide domains following the ancient duplication of the proopiomelanocortin gene in teleost fishes. Mol. Biol. Evol. 22:2417–2427.

Donoghue PC, Purnell MA. 2005. Genome duplication, extinction and vertebrate evolution. Trends Ecol. Evol. 20:312–319.

Ferris SD, Portnoy SL, Whitt GS. 1979. The roles of speciation and divergence time in the loss of duplicate gene expression. Theor. Popul. Biol. 15:114–139.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531-1545.

Fraczkiewicz R, Braun W. 1998. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. J. Comput. Chem. 19:319–333.

Friedberg F, Taliaferro L. 2005. Calmodulin genes in zebrafish (revisited). Mol. Biol. Rep. 32:55–60.

Frolow F, Harel M, Sussman JL, Mevarech M, Shoham M. 1996. Insights into protein adaptation to a saturated salt environment from the crystal structure of a halophilic 2Fe-2S ferredoxin. Nature Struct. Biol. 3:452–458.

Froschauer A, Braasch I, Volff JN. 2006. Fish genomes, comparative genomics and vertebrate evolution. Curr. Genomics 7:43–57.

Gamblin SJ, Davies GJ, Grimes JM, Jackson RM, Littlechild JA, Watson HC. 1991. Activity and specificity of human aldolases. J. Mol. Biol. 219:573–576.

Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, and Bairoch A. 2005. Protein identification and analysis tools on the ExPASy server. In: Walker JM, editors. The Proteomics Protocols Handbook. Totowa, NJ: Humana Press. p. 571–607.

Gilad Y, Man O, Glusman G. 2005. A comparison of the human and chimpanzee olfactory receptor gene repertoires. Genome Res. 15:224–230.

Gregory TR, Nicol JA, Tamm H, Kullman B, Kullman K, Leitch IJ, Murray BG, Kapraun DF, Greilhuber J, Bennett MD. 2007. Eukaryotic genome size databases. Nucleic Acids Res. 35:D332–D338.

Gu X, Vander Velden K. 2002. DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. Bioinformatics 18:500–501.

Hashiguchi Y, Furuta Y, Kawahara R, Nishida M. 2007. Diversification and adaptive evolution of putative sweet taste receptors in threespine stickleback. Gene 396:170–179.

He X, Zhang J. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics 169:1157–1164.

Hoegg S, Brinkmann H, Taylor JS, Meyer A. 2004. Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. J. Mol. Evol. 59:190–203.

Hoegg S, Meyer A. 2005. Hox clusters as models for vertebrate genome evolution. Trends Genet. 21:421–424.

Hoegg S, Meyer A. 2007. Phylogenomic analyses of KCNA gene clusters in vertebrates: why do gene clusters stay intact? BMC Evol. Biol. 7:139.

Hoekstra HE, Coyne JA. 2007. The locus of evolution: evo devo and the genetics of adaptation. Evolution 61:995–1016.

Holland PWH, Garcia-Fernandez J, Williams NA, Sidow A. 1994. Gene duplication and the origins of vertebrate development. Development (Suppl.):125–133.

Hughes AL. 2005. Gene duplication and the origin of novel proteins. Proc. Natl. Acad. Sci. U.S.A. 102:8791–8792.

Hughes AL. 2007. Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. Heredity 99:364–373.

Inoue JG, Miya M, Tsukamoto K, Nishida M. 2003. Basal actinopterygian relationships: a mitogenomic perspective on the phylogeny of the ''ancient fish''. Mol. Phylogenet. Evol. 26:110–120.

Inoue JG, Miya M, Venkatesh B, Nishida M. 2005. The mitochondrial genome of Indonesian coelacanth *Latimeria menadoensis* (Sarcopterygii: Coelacanthiformes) and divergence time estimation between the two coelacanths. Gene 349:227–235.

Ishimaru Y, Okada S, Naito H, Nagai T, Yasuoka A, Matsumoto I, Abe K. 2005. Two families of candidate taste receptors in fishes. Mech. Dev. 122:1310–1321.

Jaillon O, Aury JM, Brunet F, et al. (61 co-authors). Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature 431:946–957.

Jeon YH, Heo YS, Kim CM, Hyun YL, Lee TG, Ro S, Cho JM. 2005. Phosphodiesterase: overview of protein structures, potential therapeutic applications and recent progress in drug development. Cell. Mol. Life Sci. 62:1198–1220.

Jobb G. 2007. TREEFINDER version of June 2007. Distributed by the author at www.treefinder.de. Munich, Germany.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8:275–282.

Kao HW, Lee SC. 2002. Phosphoglucose isomerases of hagfish, zebrafish, gray mullet, toad, and snake, with reference to the evolution of the genes in vertebrates. Mol. Biol. Evol. 19:367–374.

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. 2004. The KEGG resource for deciphering the genome. Nucleic Acids Res. 32(Database issue):D277–D280.

Kasahara M, Nakaya J, Satta Y, Takahata N. 1997. Chromosomal duplication and the emergence of the adaptive immune system. Trends Genet. 13:90–92.

Kasahara M, et al. (38 co-authors). 2007. The medaka draft genome and insights into vertebrate genome evolution. Nature 447:714–719.

Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature 428:617–624.

Kikugawa K, Katoh K, Kuraku S, Sakurai H, Ishida O, Iwabe N, Miyata T. 2004. Basal jawed vertebrate phylogeny inferred from multiple nuclear DNA-coded genes. BMC Biol. 2:3.

Kim CB, Amemiya C, Bailey W, Kawasaki K, Mezey J, Miller W, Minoshima S, Shimizu N, Wagner G, Ruddle F. 2000. Hox cluster genomics in the horn shark, *Heterodontus francisci*. Proc. Natl. Acad. Sci. U.S.A. 97:1655–1660.

Kinoshita T, Maruki R, Warizaya M, Nakajima H, Nishimura S. 2005. Structure of a high-resolution crystal form of human triosephosphate isomerase: improvement of crystals using the gel-tube method. Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun. 61:346–349.

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

Kurosawa G, Takamatsu N, Takahashi M, Sumitomo M, Sanaka E, Yamada K, Nishii K, Matsuda M, Asakawa S, Ishiguro H, Miura K, Kurosawa Y, Shimizu N, Kohara Y, Hori H. 2006. Organization and structure of *hox* gene loci in medaka genome and comparison with those of pufferfish and zebrafish genomes. Gene 370:75–82.

Lavoué S, Miya M, Inoue JG, Saitoh K, Ishiguro NB, Nishida, M. 2005. Molecular systematics of the gonorynchiform fishes (Teleostei) based on whole mitogenome sequences: implications for higher-level relationships within the Otocephala. Mol. Phylogenet. Evol. 37:165–177.

Lee JH, Jeffery CJ. 2005. The crystal structure of rabbit phosphoglucose isomerase complexed with D-sorbitol-6-phosphate, an analog of the open chain form of D-glucose-6-phosphate. Protein Sci. 14:727–734.

Lee JS. 2000. The internally self-fertilizing hermaphroditic teleost *Rivulus marmoratus* (Cyprinodontiformes, Rivulidae) beta-actin gene: amplification and sequence analysis with conserved primers. Mar. Biotechnol. 2:161–166.

Longhurst TJ, Joss JMP. 1999. Homeobox genes in the australian lungfish *Neoceratodus forsteri.* J. Exp. Zool. B. Mol. Dev. Evol. 285:140–145.

Lundin LG. 1993. Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. Genomics 16:1–19.

Lynch M, Walsh JB. 1998. Genetics and Analysis of Quantitative Traits. Sinauer Associates, Sunderland, MA.

Lynch M, Force A. 2000A. The probability of duplicate gene preservation by subfunctionalization. Genetics 154:459–473.

Lynch M, Force A. 2000B. The origin of interspecific genomic incompatibility via gene duplication. Am. Nat. 156:590–605.

Lynch M. 2002. Gene duplication and evolution. Science 297:945–947.

Lynch M, Katju V. 2004. The altered evolutionary trajectories of gene duplicates. Trends Genet. 20:544–549.

Lynch M. 2005. Simple evolutionary pathways to complex proteins. Protein Sci. 14:2217–2225.

Maddison WP, Maddison DR. 2003. MacClade, ver. 4.06. OSX. Sinauer, Sunderland, Massachusetts.

Maurady A, Zdanov A, de Moissac D, Beaudry D, Sygusch J. 2002. A conserved glutamate residue exhibits multifunctional catalytic roles in D-fructose-1,6-bisphosphate aldolases. J. Biol. Chem. 277:9474–9483.

Merritt TJS, Quattro JM. 2001. Evidence for a period of directional selection following gene duplication in a neurally expressed locus of triosephosphate isomerase. Genetics 159:689–697.

Merritt TJS, Quattro JM. 2002. Negative charge correlates with neural expression in vertebrate aldolase isozymes. J. Mol. Evol. 55:674–683.

Meyer A, Malaga-Trillo E. 1999. Vertebrate genomics: More fishy tales about *Hox* genes. Curr. Biol. 9:R210–R213.

Meyer A, Schartl M. 1999. Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. Curr. Opin. Cell Biol. 11:699–704.

Meyer A, Van de Peer Y. 2005. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). Bioessays. 27:937–945.

Miya M, Takeshima H, Endo H, Ishiguro NB, Inoue JG, Mukai T, Satoh TP, Yamaguchi M, Kawaguchi A, Mabuchi K, Shirai SM, Nishida M. 2003. Major patterns of higher teleostean phylogenies: a new perspective based on 100 complete mitochondrial DNA sequences. Mol. Phylogenet. Evol. 26:121–138.

Miya M, Satoh TP, Nishida M. 2005. The phylogenetic position of toadfishes (order Batrachoidiformes) in the higher ray-finned fish as inferred from partitioned Bayesian analysis of 102 whole mitochondrial genome sequences. Biol. J. Linn. Soc. Lond. 85:289–306.

Murphy WJ, Pevzner PA, O'Brien SJ. 2004. Mammalian phylogenomics comes of age. Trends Genet. 20:631–639.

Naruse K, Fukamachi S, Mitani H, et al. (20 co-authors). 2000. A detailed linkage map of medaka, *Oryzias latipes*: comparative genomics and genome evolution. Genetics 154:1773–1784.

Nei M, Roychoudhury KA. 1973. Probability of fixation of nonfunctional genes at duplicate loci. Am. Nat. 107:362–372.

Nelson JS. 2006. Fishes of the World, 4th Edition. New Jersey: John Wiley & Sons Inc.

Niimura Y, Nei M. 2005. Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods. Proc. Natl. Acad. Sci. U.S.A. 102:6039–6044.

Nylander JAA. 2004. MrModeltest v2. Program distributed by the author. Uppsala: Evolutionary Biology Centre, Uppsala University.

Ohno S. 1970. Evolution by gene duplication. New York: Springer-Verlag.

Ohno S. 1999. Gene duplication and the uniqueness of vertebrate genomes circa 1970–1999. Semin. Cell Dev. Biol. 10:517–522.

Panopoulou G, Hennig S, Groth D, Krause A, Poustka AJ, Herwig R, Vingron M, Lehrach H. 2003. New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. Genome Res. 13:1056–1066.

Panopoulou G, Poustka AJ. 2005. Timing and mechanism of ancient vertebrate genome duplications – the adventure of a hypothesis. Trends Genet. 21:559–567.

Posada D, Crandall KA. 1998. MODEL TEST: testing the model of DNA substitution. Bioinformatics Appl. Note 14:817–818.

Pous J, Canals A, Terzyan SS, Guasch A, Benito A, Ribó M, Vilanova M, Coll M. 2000. Three-dimensional structure of a human pancreatic ribonuclease variant, a step forward in the design of cytotoxic ribonucleases. J. Mol. Biol. 303:49–60.

Prince VE, Pickett FB. 2002. Splitting pairs: the diverging fates of duplicated genes. Nat. Rev. Genet. 3:827–837.

Prohaska SJ, Fried C, Amemiya CT, Ruddle FH, Wagner GP, Stadler PF. 2004. The shark HoxN cluster is homologous to the human HoxD cluster. J. Mol. Evol. 58:212–217.

Rastogi S, Liberles DA. 2005. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evol. Biol. 5:28.

Robinson-Rechavi M, Boussau B, Laudet V. 2004. Phylogenetic dating and characterization of gene duplications in vertebrates: the cartilaginous fish reference. Mol. Biol. Evol. 21:580–586.

Robinson-Rechavi M, Alibés A, Godzik A. 2006. Contribution of electrostatic interactions, compactness and quaternary structure to protein thermostability: lessons from structural genomics of *Thermotoga maritima*. J. Mol. Biol. 356:547–557.

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

Sato Y, Nishida M. 2007. Post-duplication charge evolution of phosphoglucose isomerases in teleost fishes through weak selection on many amino acid sites. BMC Evol. Biol. 7:204.

Sayle RA, Milner-White EJ. 1995. RASMOL: biomolecular graphics for all. Trends Biochem. Sci. 20:374–376.

Shi P, Zhang J. 2006. Contrasting modes of evolution between vertebrate sweet/umami receptor genes and bitter receptor genes. Mol Biol Evol. 23:292–300.

Stadler PF, Fried C, Prohaska SJ, Bailey WJ, Misof BY, Ruddle FH, Wagner GP. 2004. Evidence for independent Hox gene duplications in the hagfish lineage: a PCR-based gene inventory of *Eptatretus stoutii*. Mol. Phylogenet. Evol. 32:686–694.

Steinke D, Hoegg S, Brinkmann H, Meyer A. 2006. Three rounds (1R/2R/3R) of genome duplications and the evolution of the glycolytic pathway in vertebrates. BMC Biol. 4:16.

St-Jean M, Lafrance-Vanasse J, Liotard B, Sygusch J. 2005. High resolution reaction intermediates of rabbit muscle fructose-1,6-bisphosphate aldolase: substrate cleavage and induced fit. J. Biol. Chem. 280:27262–27270.

Strimmer K, Rambaut A. 2002. Inferring confidence sets of possibly misspecified gene trees. Proc. Roy. Soc. B. 269:137–142.

Swofford DL. 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods), version 4.0b10 (Alvitec). Sinauer, Sunderland, Massachusetts.

Taylor JS, Van de Peer Y, Meyer A. 2001. Genome duplication, divergent resolution and speciation. Trends Genet. 17:299–301.

Taylor JS, Braasch I, Frickey T, Meyer A, Van de Peer Y. 2003. Genome duplication, a trait shared by 22000 species of ray-finned fish. Genome Res. 13:382–390.

Tocchini-Valentini GD, Fruscoloni P, Tocchini-Valentini GP. 2005. Structure, function, and evolution of the tRNA endonucleases of Archaea: an example of subfunctionalization. Proc. Natl. Acad. Sci. U.S.A. 102:8933–8938.

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.

Van der Hoeven F, Sordino P, Fraudeau N, Izpisua-Belmonte JC, Duboule D. 1996. Teleost *HoxD* and *HoxA* genes: comparison with tetrapods and functional evolution of the *HoxD* complex. Mech. Dev. 4:9–21.

Vandepoele K, De Vos W, Taylor JS, Meyer A, Van de Peer Y. 2004. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. Proc. Natl. Acad. Sci. U.S.A. 101:1638–1643.

Venkatesh B, Ning Y, Brenner S. 1999. Late changes in spliceosomal introns define clades in vertebrate evolution. Proc. Natl. Acad. Sci. U.S.A. 96:10267–10271.

Venkatesh B, Erdmann MV, Brenner S. 2001. Molecular synapomorphies resolve evolutionary relationships of extant jawed vertebrates. Proc. Natl. Acad. Sci. U.S.A. 98:11382–1387.

Venter JC, Adams MD, Myers EW, et al. (272 co-authors). 2001. The sequence of the human genome. Science 291:1304–1351.

Vogel G. 1998. Doubled genes may explain fish diversity. Science. 281:1119–1121.

Waterston RH, et al. (222 co-authors). 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562.

Werth CR, Windham MD. 1991. A model for divergent, allopatric speciation of polyploid pteridophytes resulting from silencing of duplicate-gene expression. Am. Nat. 137:515–526.

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.

Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. Nat. Rev. Genet. 2:333–341.

Woolfe A, Elgar G. 2007. Comparative genomics using Fugu reveals insights into regulatory subfunctionalization. Genome Biol. 8:R53.

Yamanoue Y, Miya M, Inoue JG, Matsuura K, Nishida M. 2006. The mitochondrial genome of spotted green pufferfish *Tetraodon nigroviridis* (Teleostei: Tetraodontiformes) and divergence time estimation among model organisms in fishes. Genes Genet. Syst. 81:29–39.

Yang Z. 1994. Estimating the pattern of nucleotide substitution. J. Mol. Evol. 39:105–111.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13:555–556.

Yano JK, Poulos TL. 2003. New understandings of thermostable and peizostable enzymes. Curr. Opin. Biotechnol. 14:360–365.

Yu WP, Brenner S, Venkatesh B. 2003. Duplication, degeneration and subfunctionalization of the nested *synapsin-Timp* genes in *Fugu*. Trends Genet. 19:180–183.

Yu WP, Yew K, Rajasegaran V, Venkatesh B. 2007. Sequencing and comparative analysis of fugu protocadherin clusters reveal diversity of protocadherin genes among teleosts. BMC Evol. Biol. 7:49.

Zhang J. 2000. Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. J. Mol. Evol. 50:56–68.

Zhang J, Zhang YP, Rosenberg HF. 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. Nat. Genet. 30:411–415.

Zhang J. 2003. Evolution by gene duplication: an update. Trends Ecol. Evol. 18:292–298.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. 22:2472–2479.

Zhang, J. 2006. Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. Nat. Genet. 38:819–823.