Evolution of genes duplicated through fish-specific genome doubling

A Dissertation

Submitted in Partial Fulfillment

of the Requirements for

the Degree of Doctor of Agriculture

by

Yukuto Sato

The University of Tokyo

December 2007

Contents

Chapter 1.

General introduction

1.1. Evolution of novel genes via gene duplication

Elucidating the origin of genes and proteins with new function is essential to understand evolutionary transitions in biological functions and characteristics, such as metabolic abilities, physiological traits, immune systems, or morphological designs. It also provides a foundation for studies on protein engineering as well as conservation of natural genetic resources. Although novel genes can be generated by a point mutation, many of them appear to be produced primarily through gene duplication. This idea was widely popularized by the book "*Evolution by Gene Duplication*" (Ohno, 1970). It was not until the late 1990s, however, that the prevalence and importance of gene duplication has been evidently demonstrated on the basis of analysis of whole genome sequences from various organisms; the genome of each of bacteria, archaebacteria, and eukaryotes consists of a large number of duplicated genes (Zhang, 2003), indicating that the genomes have become complex mostly through repeated gene duplications. The importance of gene duplication as a fundamental process of evolution has been well known among biologists today (Ohno, 1999).

In response to the recent progress in genetics and genomics, studies on the evolution of duplicated genes begun to investigate the mechanisms by which duplicate genes persist in genomes and the detailed process by which new gene functions and protein properties can evolve (Force et al., 1999; Prince and Pickett, 2002; Zhang, 2003; He and Zhang, 2005; Rastogi and Liberles, 2005; Sato and Nishida, 2007). These subjects have now become one of the main concerns of the study of molecular evolution.

1.2. Whole genome doubling and the evolution of genes and genomes

Although duplication of genes can result from unequal crossing over, retroposition, or chromosomal duplication, the most dynamic event that accompanies extensive gene duplication is whole genome doubling (WGD). WGD, which occurs via polyploidization, generates a remarkable number of redundant genes. The occurrence of many redundant genes, in turn, provides mass opportunities for the evolution of novel genes. Therefore, WGDs may have played a critical role in evolution. Recent genomic studies have actually shown that a large, complex genome of eukaryotes including yeasts, land plants, and vertebrates have underwent several rounds of WGD during evolution (Blanc and Wolfe, 2004; Kellis et al., 2004; Adams and Wendel, 2005; Dehal and Boore, 2005; Panopoulou and Poustka, 2005; Froschauer et al., 2006). The findings imply that ancient WGDs are important to understand the evolution of complex eukaryotic genomes. A WGD also provides a unique opportunity to study the evolution of duplicated genes as mentioned below.

Here, I focus on genes generated via WGDs in vertebrate evolution. It is considered that vertebrates, after the split from the common ancestor shared with other deuterostomes, have undergone several rounds of WGD early in their evolution (Fig. 1-1A; Ohno,1970; Lundin, 1993; Panopoulou et al., 2003; Taylor et al., 2003). This notion is consistent with finding in earlier studies that the genomes of jawed vertebrates (gnathostomes) contain four or more clusters of *Hox* genes (Fig. 1-2), and also major histocompatibility complex (MHC) paralogous regions, respectively (Holland et al., 1994; Kasahara et al., 1997; Amores et al., 1998; Abi-Rached et al. 2002; Hoegg and Meyer 2005). Subsequent whole-genome analyses of mammals and fishes have provided further evidences for the WGD events occurred during vertebrate evolution (Jaillon et al., 2004; Dehal and Boore, 2005; Panopoulou and Poustka, 2005; Kasahara et al., 2007). The WGDs are considered to have occurred successively; first at the common ancestor of jawless and jawed vertebrates (one-round whole genome doubling,

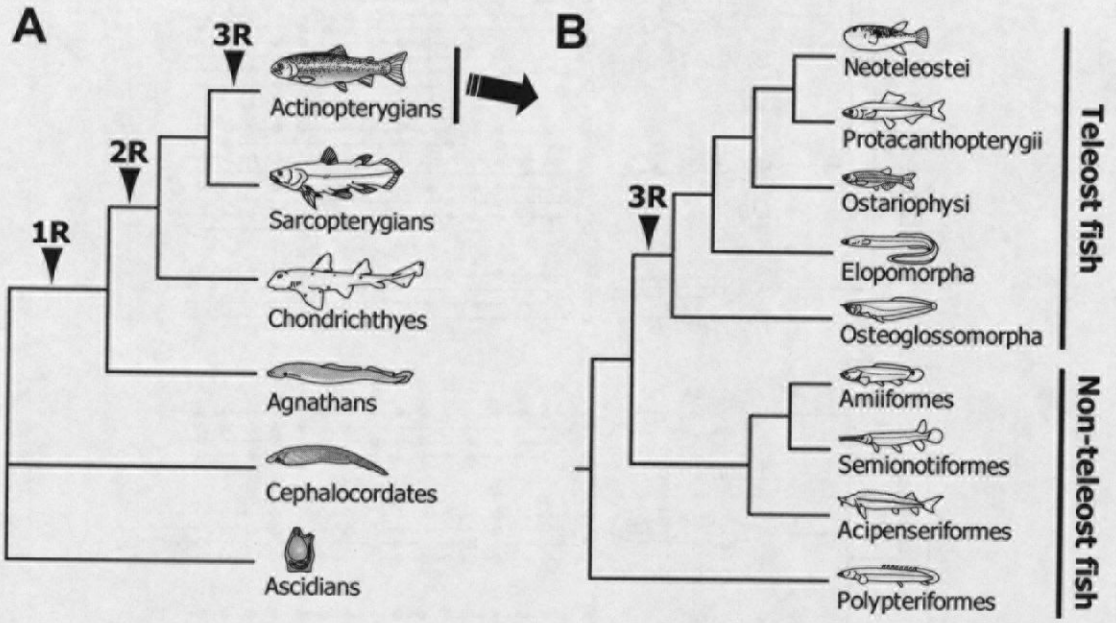

Fig. 1-1. Supposed whole-genome doubling events during vertebrate evolution. 1R, 2R, and 3R indicate one, two, and three-round whole genome doubling, respectively. **(A)** A known vertebrate phylogeny and the estimated timing of the 1R (Stadler et al., 2004) and 2R (Robinson-Rechavi et al., 2004). **(B)** A phylogeny of actinopterygians (ray-finned fishes; Inoue et al., 2003) and the estimated timing of the 3R (Chiu et al., 2004; Hoegg et al., 2004; Sato and Nishida, 2007).

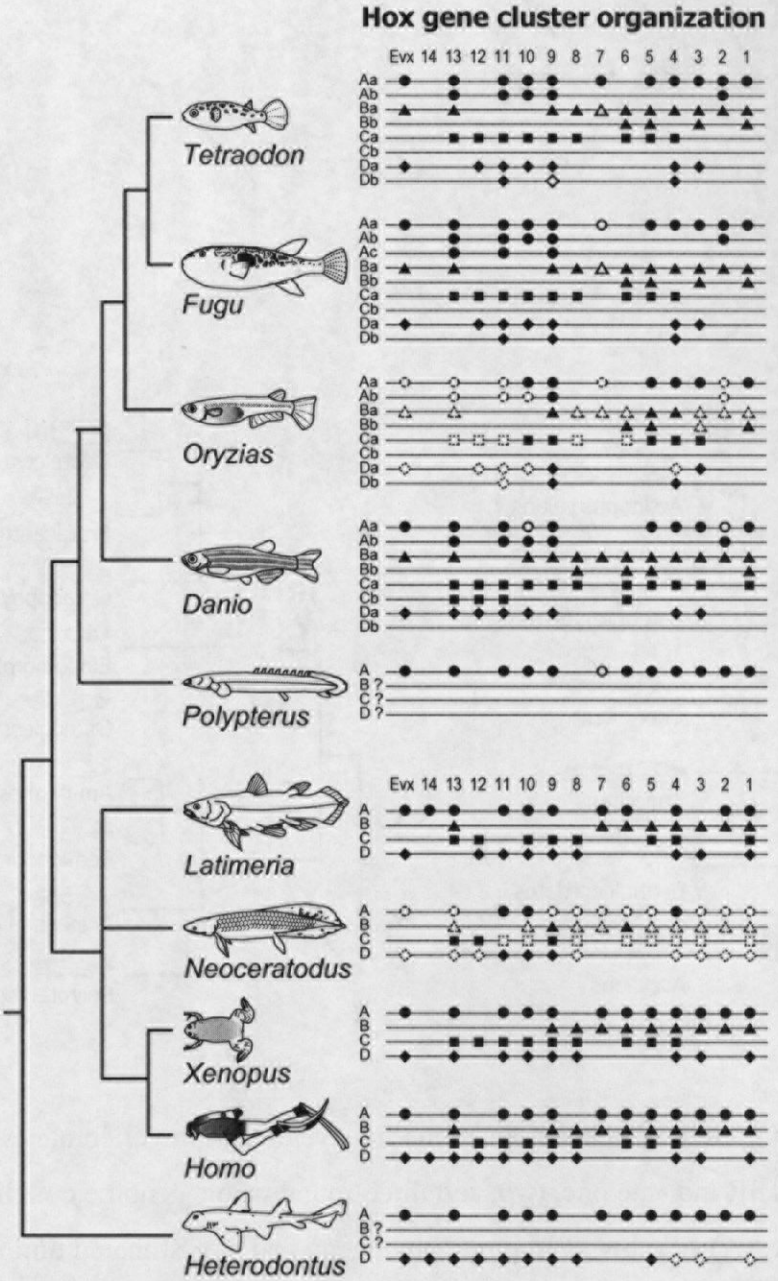

Fig. 1-2. A known phylogeny of jawed vertebrates and their repertoire of clusters of *Hox* genes. Circles, triangles, squares, and diamonds denote the *Hox* genes belong to cluster *A*, *B*, *C*, and *D*, respectively. Filled and open symbols indicate intact genes and pseudogenes, respectively. Dashed symbols idicate the hypothetical genes that have not been identified but estimated to be present (Hoegg and Meyer, 2005). Sources of data: pufferfish, *Tetraodon nigroviridis*: Jaillon et al. (2004); Fugu, *Fugu rubripes*: Aparicio et al. (1997); Amores et al. (2004); medaka fish, *Oryzias latipes*: Naruse et al. (2000); Kurosawa et al. (2006); zebrafish, *Danio rerio:* Van der Hoeven et al. (1996); Amores et al. (1998); bichir, *Polypterus senegalus*: Chiu et al. (2004); Australian lungfish, *Neoceratodus forsteri*: Longhurst and Joss (1999); clawed frog, *Xenopus tropicalis* and human, *Homo sapiens*: Hoegg and Meyer (2005); horn shark, *Heterodontus francisci*: Kim et al. (2000); Chiu et al. (2002).

1R-WGD; Stadler et al., 2004), second at the ancestor of jawed vertebrates (2R-WGD; Prohaska et al., 2004; Robinson-Rechavi et al., 2004), and third at the ancestor of teleost fish (3R-WGD; Amores et al., 1998; Chiu et al., 2004; Hoegg et al., 2004; Jaillon et al., 2004; Kasahara et al., 2007; Sato and Nishida, 2007), although the detailed phylogenetic positions of these WGD events are not fully determined (see Fig. 1-1).

The 1R-WGD and 2R-WGD events contributed to the formation of genomes of land vertebrates including mammals appear to have occurred before the split of lobe-fin fish (sarcopterygians) and ray-finned fish (actinopterygians) at 450 million years ago (MYA), and thus being very old (Fig. 1-3; Panopoulou et al., 2003; Christoffels et al. 2004; Vandepoele et al. 2004; Panopoulou and Poustka, 2005). Probably because of this antiquity, the 1R- and 2R-WGDs have left little unmistakable trace in duplicated genes or tree topology of duplicated genes, when currently available genome sequences are analyzed (Venter et al. 2001; Wolfe 2001; Dehal and Boore 2005). This implies that it is now difficult to study the evolution of duplicated genes on the basis of 1R- and 2R-WGDs.

1.3. Ray-finned fishes as a comparative model system to study duplicate gene evolution

Fish-specific genome doubling, or 3R-WGD, is the most recent one among the three WGD events occurred in vertebrate evolution (see Fig. 1-1A). The existence of this event has been confirmed by analyses of medaka and pufferfish genomes (see Fig. 1-3; Christoffels et al. 2004; Jaillon et al., 2004; Kasahara et al., 2007). From the analyses of several nuclear gene families, this 3R-WGD is estimated to have occurred in the common ancestor of teleost fishes, but after their divergence from an ancestor of the basal non-teleost fishes (Fig. 1-1B; Chiu et al., 2004; Hoegg et al., 2004; Sato and Nishida, 2007). The divergence time between teleosts and non-teleosts is estimated to be 359 to 404 MYA, and timing of the beginning of teleost radiation is estimated to be 309 to 324 MYA, on the basis of molecular clock analyses of

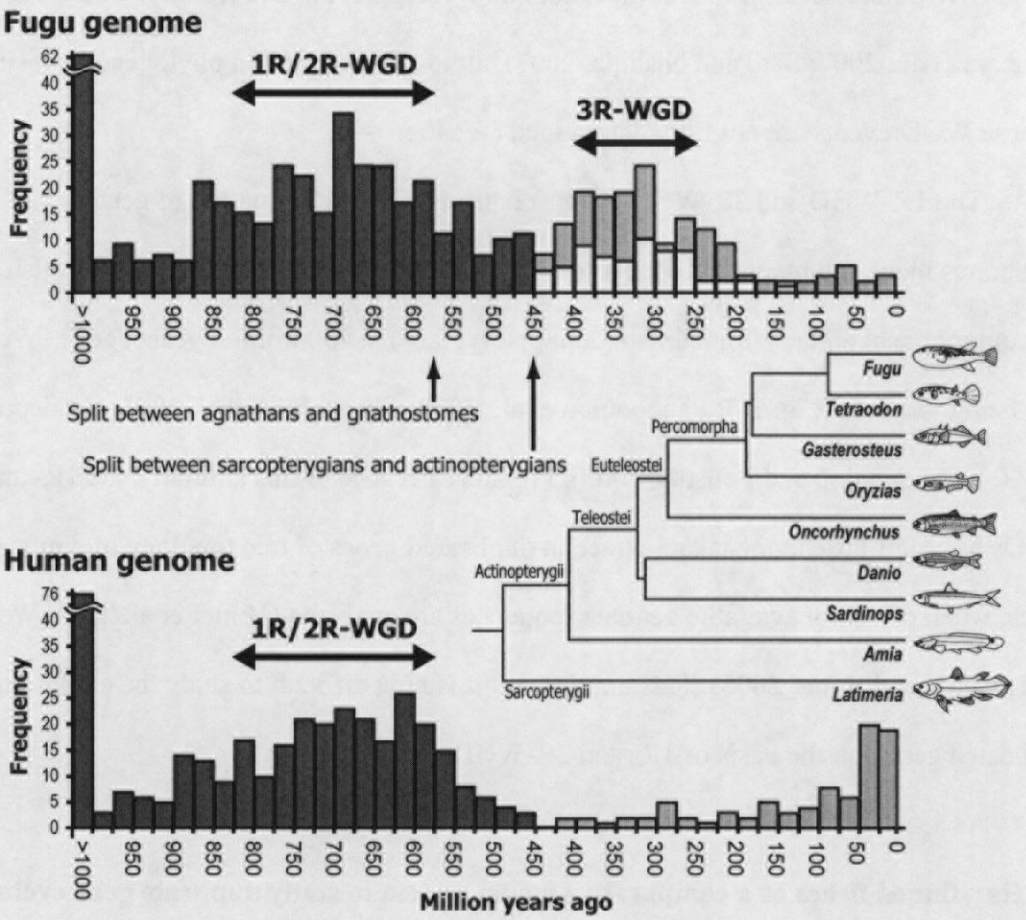

Fig. 1-3. Estimated age distribution of duplicated genes in the Fugu and human genome (source of data: Vandepoele et al., 2004), and molecular phylogeny and divergence time estimates for major lineages of Teleostomi (bony fish) (Yamanoue et al., 2006). The dark- and light-gray bars correspond to gene duplication events that have occurred before and after the split between actinopterygians and sarcopterygians, respectively. White bars refer to duplication events that were confirmed to have occurred via chromosomal block duplications.

whole mitochondrial genome sequences (Inoue et al., 2005; Yamanoue et al., 2006). This estimated time range for appearance of the common ancestor of current teleosts matches with the putative occurrence time of the 3R-WGD of 320 to 400 (average 353.0) MYA inferred from the analysis of pufferfish genomes (Fig. 1-3; Christoffels et al. 2004; Vandepoele et al. 2004; Christoffels et al. 2006). These suggest that the 3R-WGD has occurred at 300 to 400 million years ago in the common ancestor of teleost fish. Thus, the 3R-WGD would provide a nice opportunity to investigate the evolution of duplicated genes that arose through WGD, as compared to the 1R- and 2R-WGDs, which appear to be very old (see Fig. 1-3).

Owing to recent progress in molecular phylogenetic studies, a reliable phylogenetic framework for ray-finned fishes and divergence time estimates between their lineages, which are essential for comparative evolutionary analyses, are available (Inoue et al. 2003; Miya et al. 2003; Kikugawa et al., 2004; Inoue et al. 2005; Lavoué et al. 2005; Miya et al. 2005; Yamanoue et al. 2006). Furthermore, the full-genome sequence data of five species of teleosts are available in public databases. These five species of teleosts (zebrafish *Danio rerio*, medaka *Oryzias latipes*, stickleback *Gasterosteus aculeatus*, pufferfish *Tetraodon nigroviridis*, and Fugu *Fugu rubripes*) cover a wide evolutionary time scales, because zebrafish was estimated to have diverged from other lineages about 320 MYA, while pufferfish and Fugu diverged about 90 MYA (Yamanoue et al. 2006; see Fig. 1-3). In addition, the basal lineages of ray-finned fish, including Polypteriformes (bichir), Acipenseriformes (sturgeon), Semionotiformes (gar), and Amiiformes (amia), were estimated to have not experienced the 3R-WGD as mentioned above (see Fig. 1-1B). These non-teleost fishes may be suitable as outgroups for comparison to analyze the gene evolution after 3R-WGD. Because of these uniqueness, ray-finned fishes would be a good model system for studying the evolution of duplicated genes after WGD.

1.4. Persistence and the evolution of duplicated genes following fish-specific genome doubling

Although it was shown that teleosts underwent a fish-specific genome doubling or 3R-WGD (Jaillon et al., 2004; Kasahara et al., 2007), it is unclear how the duplicated genes generated by 3R-WGD have been lost or retained through diversification of teleost lineages, and then, how many genes have remained in duplicated condition in current teleost genomes. The current numbers of pairs of 3R-WGD-derived duplicated genes per genome were estimated to be 750 to 2,100, by performing pairwise comparisons between teleost (pufferfish or medaka) and human genomes (Jaillon et al., 2004; Kasahara et al., 2007). These comparisons, however, do not provide insight into the evolutionary process of duplicated genes derived by 3R-WGD; the number of duplicate genes contained in the genome of a common ancestor of teleost fish, and subsequent process of lineage-specific loss or gain of new functions of the genes remain unclear.

At present, the detailed process of how a new function or protein property evolves after gene duplication is still poorly understood. Duplicated genes can become novel genes by acquisition of new function or novel property of encoding protein through fixation of beneficial mutations (neofunctionalization; Fig. 1-4A; Ohno, 1970). Beneficial mutations, however, are generally rarer than loss-of-function mutations (Lynch and Walsh, 1998), such as truncations, frame shift mutations, and deletions of *cis*-regulatory region. In spite of this predominance of loss-of-function mutations, a large number of duplicated genes were found to be maintained on the genomes of various organisms, without becoming pseudogenes (nonfunctionalization; Fig. 1-4B, C) (reviewed in Force et al., 1999). This fact implies that many of duplicated genes were retained in genomes by some mechanisms independent of neofunctionalization with beneficial mutations.

On the basis of these findings, Force et al. (1999) and Lynch and Force (2000A)

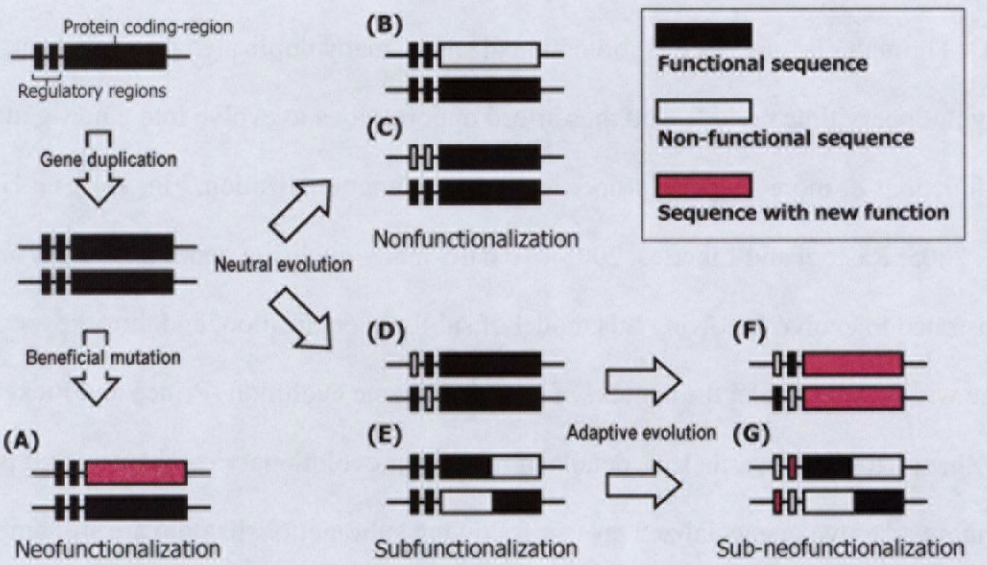

Fig. 1.4. A schematic view of current models for duplicate gene preservation and evolution. Small and large boxes indicate *cis*-regulatory and protein-coding sequences on the genome, respectively. Black and white boxes denote functional and non-functional sequences, respectively. Magenta boxes indicate the sequences that acquired new function or more adaptive property.

proposed that the loss-of-function mutations accumulating in duplicated genes promote the preservation of duplicated genes. If the function of the both of duplicated genes was partially and differentially defeated by loss-of-function mutations, then both of them become essential to accomplish the entire task of the ancestral gene (subfunctionalization, Fig. 1-4D, E), and thus be retained in the genome by natural selection (Force et al., 1999; Lynch and Force, 2000A). Through this process of subfunctionalization, many duplicated genes can persist for long evolutionary time periods, and thus afford opportunities to evolve into a new gene with novel functions or more adaptive properties (sub-neofunctionalization, Fig. 1-4F, G; He and Zhang, 2005; Rastogi and Liberles, 2005). To date, many duplicate genes have been demonstrated to evolve following this model of subfunctionalization, and thus the model has become widely accepted in the context of duplicated gene evolution (Prince and Pickett, 2002; Zhang, 2003). Nevertheless, details of molecular evolutionary mechanism and process into a more adaptive or specialized protein following subfunctionalization are still ambiguous (e.g., reviewed in Lynch and Katju, 2004; Hughes, 2005).

1.5. Purposes of this study

This study aimed to explore the molecular evolutionary processes and mechanisms whereby duplicated genes evolve into new genes with novel function or more adaptive property (neofunctionalization), specifically after subfunctionalization. As a framework for comparative genetics/genomics of this study, I focused on well-established ancient WGD in vertebrates, the 3R-WGD or fish-specific genome doubling (see Fig. 1-1), and reliable molecular phylogeny and divergence time estimates for ray-finned fish lineages, which were mainly inferred from whole mitochondrial genome sequence analyses (see Fig. 1-3). In the molecular evolutionary analysis, basal non-teleost fishes were used as appropriate outgroups, because these groups were estimated to had diverged before the occurrence of 3R-WGD as

mentioned above.

To accomplish the comparative genetic/genomic analyses of this study, the following methods were integratively utilized: (*i*) data-mining in the teleost genomes available (zebrafish, medaka, stickleback, and pufferfish); (*ii*) contemporary computational methods for molecular phylogeny and ancestral sequence reconstruction based on Bayesian and advanced maximum-likelihood inference; (*iii*) comparative molecular evolutionary analysis of duplicated genes on the basis of three-dimensional (3-D) structural information on a protein. The wet lab experiments, RT-PCR, cDNA cloning, and DNA sequencing, were carried out to obtain data for tissue-specific gene expression and gene sequences on basal non-teleosts because such data were not available from public databases.

In Chapter 2, I conducted a comparative analysis of sets of more than 100 gene families to clarify the numbers and characteristics of duplicated genes that have persisted in teleost genomes since the 3R-WGD. As a study model, I chose gene families involved in several signal transduction and metabolism pathways in human. Their orthologous genes in other vertebrate species including teleosts were mined from whole-genome sequence data available. Based on these data, gene duplication events occurred by the 3R-WGD were reliably identified by careful systematic analyses of molecular phylogeny for each gene family. Using information from these analyses, the process of gene loss or retention following the 3R-WGD were confidently inferred.

Next, I focused on the fish-specific duplicated genes *Pgi* (phosphoglucose isomerase, EC 5.3.1.9), which appear to have persisted through subfunctionalization after 3R-WGD. The underlying evolutionary process producing novel protein properties after subfunctionalization was analyzed by focusing on the structural properties of a protein, specifically the electric charge (Chapter 3). The aim of Chapter 4 was to ascertain the generality of findings in the Chapter 3. The *Ald* genes (fructose-1,6-bisphosphate aldolase, EC 4.1.2.13) of vertebrates, which were duplicated probably through 1R-WGD and 2R-WGD (Merritt and Quattro, 2002;

Steinke et al., 2006), were chosen and analyzed in the same manner as applied to *Pgi* genes. The results were examined in comparison to those obtained from the analysis of *Pgi* gene (Chapter 3) and existing information of the duplicated genes, pancreatic ribonuclease genes of leaf-eating monkey (Zhang et al., 2002; Zhang, 2006) and triose phosphate isomerase genes of ray-finned fish (Merritt and Quattro, 2001). Finally, I have discussed the mode of adaptive molecular evolution that plays a substantial role in the evolution of novel genes and proteins.