Chapter 3.
Post-duplication charge evolution of phosphoglucose isomerases in teleost fishes through weak selection on many amino acid sites

3.1. Introduction

Proteins that arise through gene duplication can become novel proteins through fixation of beneficial mutations (Ohno, 1970), but because beneficial mutations are generally rare, the partitioning of ancestral functions among duplicated genes by neutral evolution, or subfunctionalization, has been considered the primary process for the evolution of novel proteins (Force et al., 1999; Lynch and Force, 2000A; He and Zhang, 2005; Rastogi and Liberles, 2005). To date, many duplicate genes have been demonstrated to evolve following this model of subfunctionalization, and this model thus has become widely accepted in the context of duplicated gene evolution (Prince and Pickett, 2002; Zhang, 2003; Yu et al., 2003; de Souza et al., 2005; Tocchini-Valentini et al., 2005).

Nonetheless, how a more adaptive or specialized protein property evolves after subfunctionalization is poorly understood, mainly due to the limited resolution power of current analytical methods, which seek to detect positive selection on individual amino acid substitutions involved in adaptive molecular evolution. Such methods can recognize substitutions expected to be driven by strong selection, many of which are usually function-altering substitutions at important amino acid sites, such as enzyme active sites or viral epitopes (e.g., Bielawski and Yang, 2005). However, adaptive substitutions by relatively moderate or weak selection may not be recognized by these methods. Therefore, to detect adaptive protein evolution under a much wider range of selection pressure, a novel approach is required. In this chapter, I utilized a comparative evolutionary approach to this problem, focusing on differences in the high-dimensional properties of a protein, specifically the electric charge, encoded by a pair of duplicated genes.

As a model protein system for this study, I chose an important enzyme involved in glycolysis and gluconeogenesis, phosphoglucose isomerase (PGI; EC 5.3.1.9). The gene encoding this enzyme (*Pgi*) is present as a single copy in tetrapods, whereas two copies exist in most groups of ray-finned fishes (Avise and Kitto, 1973; Kao and Lee, 2002; Steinke et al., 2006). The fact that these duplicated *Pgi-1* and *Pgi-2* genes in fishes are expressed in different organs (Avise and Kitto, 1973; Dando, 1980) implies that these fish-specific duplicate *Pgi* genes are subfunctionalized with respect to their expression, and are thus good candidates for the model of subfunctionalized genes. For ray-finned fishes, a reliable phylogenetic framework, which is essential for comparative evolutionary analyses, is available due to recent progress in molecular phylogenetic studies (Venkatesh et al., 1999, 2001; Inoue et al., 2003; Miya et al., 2003; Kikugawa et al., 2004; Lavoué et al., 2005). In addition, the basal lineages of ray-finned fishes, including Semionotiformes (gar) and Amiiformes (amia), have only one *Pgi* locus (Avise and Kitto, 1973). This single-copy gene may be the direct descendant of the ancestral unduplicated *Pgi* in ray-finned fishes, and thus can be considered an appropriate outgroup gene for comparison between the duplicated *Pgi* genes.

The *Pgi-1* and *Pgi-2* genes, which were expected to be subfunctionalized in their expression, also differ in the net electric charge of their encoded proteins (Avise and Kitto, 1973; Dando, 1980; Kao and Lee, 2002). The electric charge of soluble proteins such as PGIs is a structural property brought by a large number of multiple amino acid residues and is involved in the adaptive evolution of several soluble proteins (e.g., Frolow et al., 1996; Merritt and Quattro, 2001; Zhang et al., 2002), such as an acquisition of protein thermostability (Alsop et al., 2003; Yano and Poulos, 2003; Robinson-Rechavi et al., 2006). Therefore, the evolution of electric charge in the duplicated PGI proteins is an interesting subject to investigate regarding the evolution of novel protein properties after subfunctionalization.

In this chapter, I examined first whether the spatial expression patterns of duplicated *Pgi* genes in ray-finned fishes are compatible with predictions based on the subfunctionalization model of duplicate gene evolution. Next, by focusing on the electric charges of the PGI proteins, I analyzed the underlying evolutionary process producing novel protein properties after gene duplication through ancestral sequence inference using the maximum likelihood (ML) method based on a reliable phylogenetic framework of ray-finned fishes, and also using three-dimensional (3-D) structural information on the protein.

3.2. Materials and Methods

3.2.1. Taxonomic sampling

I chose 10 representative species from divergent lineages of ray-finned fishes, as follows—basal non-teleost ray-finned fishes: *Polypterus ornatipinnis* (bichir), *Acipenser ruthenus* (sturgeon), *Amia calva* (amia), and *Lepisosteus osseus* (gar); teleosts: *Osteoglossum bicirrhosum* (arowana) and *Anguilla anguilla* (eel) from basal groups, *Plecoglossus altivelis* (smelt) and *Danio rerio* (zebrafish) from intermediate groups, and *Mugil cephalus* (mullet) and *Fugu rubripes* (fugu) from derived groups. Live specimens, which were obtained either from local shops or other investigators in Japan, were treated according to the ethical recommendations of the Ichthyological Society of Japan and the University of Tokyo.

3.2.2. Cloning and sequencing

Total RNA was extracted from fresh skeletal muscle and liver tissue using TRIzol reagent (Invitrogen) and reverse-transcribed into first-strand cDNA with oligo-dT adaptor primer using an RNA PCR kit (TaKaRa). Partial *Pgi* cDNA was amplified using PCR with vertebrate universal degenerate primers (Kao and Lee, 2002). The well amplified DNA fragments were purified using a MinElute gel extraction kit (Qiagen), ligated into the pGEM-T Easy Vector system (Promega), transmitted into competent *E. coli* (Competent High DH5a, Toyobo), and sequenced with an ABI PRISM 3100 (Applied Biosystems) using T7 or SP6 primers. The partial *Pgi* sequences were used to design gene-specific primers (GSPs; shown in Table 3-1) for RACE PCR. 3' RACE PCR was conducted with the sense GSP and M13 primer M4 (TaKaRa) and the first-strand cDNA as the template. For 5' RACE, double-stranded cDNA PCR libraries were synthesized from 1 μg of total RNA using the cDNA synthesis kit (M-MLV version; TaKaRa) combined with the cDNA PCR library kit (TaKaRa). Then, 5' RACE PCR was conducted with the antisense GSP and CA primer (TaKaRa). Subcloning and sequencing were performed as above.

3.2.3. Phylogenetic analysis

The *Pgi* genes from 20 vertebrates were phylogenetically analyzed with the Bayesian and ML methods using the programs MrBayes 3.0B4 (Ronquist and Huelsenbeck, 2003) and PAUP 4.0b10 (Swofford, 2002), respectively. The species used [GenBank accession numbers or Ensembl Transcript IDs of the *Pgi* gene(s)] were as follows: bichir (AB282684*), sturgeon (AB282688*), amia (AB282681*), gar (AB282687*), arowana (*Pgi-1*: AB282682* and *Pgi-2*: AB282683*), eel (*Pgi-1*: AB282685* and *Pgi-2*: AB282686*), smelt (*Pgi-1*: AB282690* and *Pgi-2*: AB282691*), zebrafish (*Pgi-1*: AJ306395 and *Pgi-2*: AJ306396), mullet (*Pgi-1*: AJ306392 and *Pgi-2*: AJ306393), fugu (*Pgi-1*: NEWSINFRUT00000145974 and AB282689*, and *Pgi-2*: NEWSINFRUT00000159975), *Homo sapiens* (human; K03515), *Sus scrofa* (pig; X07382), *Oryctolagus cuniculus* (rabbit; AF199601), *Cricetulus griseus* (hamster; Z37977), *Mus musculus* (mouse; M1422), *Rattus norvegicus* (rat; ENSRNOT00000032613), *Gallus gallus* (chicken; ENSGALT00000007948), *Boiga kraepelini* (snake; AJ306394), *Bufo*

Table 3-1.

Gene-specific primers used for cDNA cloning or partial amplification of *Pgi* in this study.

Primers	Purposes	Sequence (5'→3')[1]
For bichir (*Polypterus ornatipinnis*)		
PoorPGI-3'-1	3'-RACE of *Pgi*	GGC CAA CAA ATT CAA TGG TC
PoorPGI-3'-2	3'-RACE of *Pgi*	CCT TCA TTT TAG GTG CAC TGA
PoorPGI-5'-1	5'-RACE of *Pgi*	CAG TGG TGT ATT GTG GAA GTG
For sturgeon (*Acipenser ruthenus*)		
AcruPGI-3'-1	3'-RACE of *Pgi*	GAG CAC AAG ATC TTC GTA CAG G
AcruPGI-5'-1	5'-RACE of *Pgi*	GCT CCA CTC AGA AGA TGC TC
AcruPGI-5'-2	5'-RACE of *Pgi*	ACA GGG ACA TTY TTW TCC AAG
For amia (*Amia calva*)		
AmcaPGI-3'-1	3'-RACE of *Pgi*	AAC TGC TGC CTC ATA AGG TC
AmcaPGI-3'-2	3'-RACE of *Pgi*	TCT TCA CGA AAC TGA ATC CC
AmcaPGI-5'-1	5'-RACE of *Pgi*	CGG AAA TGA TTG TCC ATC CAG T
For gar (*Lepisosteus osseus*)		
LeosPGI-3'-1	3'-RACE of *Pgi*	TGA TTG CTA TGT ATG AAC AC
LeosPGI-3'-2	3'-RACE of *Pgi*	ATC GGC ATT GGT GGA TCT G
LeosPGI-5'-1	5'-RACE of *Pgi*	ATG GAT GAG CTG GTA GAA GG
For arowana (*Osteoglossum bicirrhosum*)		
OsbiPGI1-3'-1	3'-RACE of *Pgi-1*	CAC AAA GTG TTT GAG GGA A
OsbiPGI1-5'-1	5'-RACE of *Pgi-1*	AAG GCA TGA GTC TCT GCT T
OsbiPGI2-3'-1	3'-RACE of *Pgi-2*	AAT TTG CAC CAC AAG ATC C
OsbiPGI2-3'-2	3'-RACE of *Pgi-2*	ACA TAG GCT TTG AGA ACT T
OsbiPGI2-5'-1	5'-RACE of *Pgi-2*	TGG AAG AAG TTG ATG TAC CAG
For eel (*Anguilla anguilla*)		
AgagPGI1-3'-1	3'-RACE of *Pgi-1*	GTT CAA GAA GTT GAC CCC TTT C
AgagPGI1-5'-1	5'-RACE of *Pgi-1*	TGG AAG AAG TTG ATG TAC CAG A
AgagPGI2-3'-1	3'-RACE of *Pgi-2*	AGA AGC TGA CAC CAT TCA TCC
AgagPGI2-5'-1	5'-RACE of *Pgi-2*	AGT GGT TGT CCA TCC AGT GAG

For smelt (*Plecoglossus altivelis*)		
PlalPGI1-3'-1	3'-RACE of *Pgi-1*	TCG CTG CAT ACT TCC AAC AG
PlalPGI1-5'-1	5'-RACE of *Pgi-1*	TGT CTC TGA TAG GGT GTT GGG
PlalPGI2-3'-1	3'-RACE of *Pgi-2*	ACC AAG GTA CTC GCA TGG TC
PlalPGI2-5'-1	5'-RACE of *Pgi-2*	TCC TTC TTA GCC TCC TCT GTT G
For fugu (*Fugu rubripes*)		
FuruPGI1-5'	*Pgi-1* (partial)	CAC GGA TGT AAA GAG CGT CTC CT
FuruPGI1-3'	*Pgi-1* (partial)	GCA GCA GTG GTA CAA AGC CAA

[1]Positions with mixed bases are designated by their IUB codes: R = A/G; Y = C/T; K = G/T; M = A/C; S = G/C; W = A/T.

melanostictus (toad; AJ306397), and *Paramyxine yangi* (hagfish; AJ306391). Newly cloned sequences in this study (marked with asterisks) were named under the denomination of PGI isozymes (Avise and Kitto, 1973). Bayesian and ML trees were constructed under the GTR + I + Γ model (Yang, 1994), which was selected as the best-fitting model of nucleotide substitution by hierarchical likelihood ratio tests (hLRTs) (Nylander, 2004; Posada and Crandall, 1998) with 1100 base pairs (bp) of the *Pgi* coding region (excluding the third codon position). The Bayesian posterior probabilities of the phylogeny and its branches were determined from 9901 trees. Support for heuristic ML analysis was assessed using 100 bootstrap replications.

3.2.4 Synteny analysis

The genomic regions around the *Pgi* locus (or loci) in the human, chicken, and zebrafish genomes were investigated and compared. Genomic data from the pufferfishes *Fugu rubripes* and *Tetraodon nigroviridis* were not useful in this analysis because the locations of their *Pgi* loci were not determined. Data on the neighborhood of the *Pgi* locus in the human and chicken genomes were obtained from the NCBI Mapviewer Web site (http://www.ncbi. nlm.nih.gov/mapview/). Twenty-seven protein-coding genes were identified around the human PGI locus, within a 1.8-Mb-long region on chromosome 19. The nucleotide sequences of these human genes were subjected to BLASTN searches against the zebrafish genome sequences using the Ensembl BLASTN search service (http://www.ensembl.org/Multi /blastview). The matches detected with an *E*-value threshold of $<10^{-3}$ were checked visually. Then, I selected identifiable genes described as putative orthologs of the queries. Their genomic location data were used to rebuild the synteny maps around the zebrafish *Pgi* loci.

3.2.5. Gene expression analysis

RT-PCR was performed for expression analysis of the *Pgi* genes. The primers used are described in Table 3-2. They were designed as follows: to distinguish the duplicate *Pgi* loci in teleosts, the 3' region of one primer from each primer pair was made to locate the differential nucleotide site between the two loci of the species concerned, and to avoid false amplification from genomic DNA contaminants, each primer pair was designed to span a *Pgi* exon/intron boundary considered conservative among vertebrates. Total RNA was extracted from liver, skeletal muscle, heart, gill filament, brain, and kidney tissues of fresh fish samples. RNA extraction, reverse-transcription into first-strand cDNA and PCR were performed in the same manner as mentioned in the *3.2.2. Cloning and sequencing* section. The thermal-cycle profile was as follows: 1 cycle at 94°C for 2 min; 30 cycles at 94°C for 30 sec, 60°C for 30 sec, and 72°C for 30 sec; followed by 1 cycle at 72°C for 7 min. As a positive control for gene expression, *ß-actin* cDNA was amplified using the primers 5'-GACATGGAGAAGATCTGGCA-3' and 5'-TGATCCACATCTGCTGGAAGGT-3' (predicted product size = 834 bp), which were designed by Dr. Kaoru Kuriiwa of the National Museum of Nature and Science, Tokyo. These primer sequences were based on a highly conserved region of the *ß-actin* gene in mangrove killifish, *Rivulus marmoratus* (GenBank accession number AF168615). The amplified DNA fragments were separated on a 2.0% L03 agarose gel (TaKaRa), stained with ethidium bromide, and visualized under UV light. GeneRuler^{TM} 100 bp DNA Ladder Plus (MBI Fermentas) was used as a size marker for electrophoresis.

3.2.6. Charge evolution analysis

The ML inference of the ancestral sequences of *Pgi* genes was performed by BASEML (Yang, 1997) based on the phylogeny of ray-finned fishes using whole

Table 3-2.

Gene-specific primers used for RT-PCR analysis of *Pgi* genes in this study.

Primers	Target gene (product size)	Sequence (5'→3')
For bichir (*Polypterus ornatipinnis*)		
PoorPGI-F	*Pgi* (491 bp)	TTC CAG CAG GGT GAC ATG GA
PoorPGI-R		ACA CCC CAC TGG TCA TAG CTG
For sturgeon (*Acipenser ruthenus*)		
AcruPGI-F	*Pgi* (344 bp)	ACA CCA AGG AAC ACG CAT GA
AcruPGI-R		AGC TGT TGA TGT CCC AGA TGA C
For amia (*Amia calva*)		
AmcaPGI-F	*Pgi* (422 bp)	GGG GCT CAC TGG ATG GAC AA
AmcaPGI-R		CCC TTC ATC AGG GCC TCA GT
For gar (*Lepisosteus osseus*)		
LeosPGI-F	*Pgi* (299 bp)	GGA GCT CAC TGG ATG GAC AA
LeosPGI-R		CCT TGA TGG ATG AGC TGG TAG A
For arowana (*Osteoglossum bicirrhosum*)		
OsbiPGI1-F	*Pgi-1* (266 bp)	ACG TGG TCA ATA TCG GTA TC
OsbiPGI1-R		AAC CAC AGA TGG ATC AGT AG
OsbiPGI2-F	*Pgi-2* (266 bp)	CGA TGT CGT CAA TAT TGG CAT T
OsbiPGI2-R		CAG CAG ATT TGT CCT TGG CG
For eel (*Anguilla anguilla*)		
AgagPGI1-F	*Pgi-1* (285 bp)	TCA TCA AGG AAC ACG CAT GA
AgagPGI1-R		AGC GCT CCA AGA ATG AAA GG
AgagPGI2-F	*Pgi-2* (351 bp)	AAC CTG CAC CAC AAG ATC CT
AgagPGI2-R		AAC CTC AGT GGT GTC CTG AA
For zebrafish (*Danio rerio*)		
DarePGI1-F	*Pgi-1* (243 bp)	GAG ATA ACC TGC ATC ATA AGA TC
DarePGI1-R		TCT TGT GCT CAT ACA TCG CA

DarePGI2-F	*Pgi-2* (179 bp)	CTG ATG AAG GGG AAA ACA ACA GAA
DarePGI2-R		TCA TAC ATG GCG ATC AGC ACA
For smelt (*Plecoglossus altivelis*)		
PlalPGI1-F	*Pgi-1* (302 bp)	CAA GGC ACT CGT ATG ATT CC
PlalPGI1-R		TTG TGC TCG TAC ATT GCA AC
PlalPGI2-F	*Pgi-2* (357 bp)	ACT TAT CCA CCA AGG TAC TC
PlalPGI2-R		TGG TCA AAA CTG TTG ATC TC
For mullet (*Mugil cephalus*)		
MucePGI1-F	*Pgi-1* (386 bp)	CCA AAG TAC TCG TCT GAT TCC
MucePGI1-R		TCT TCT TAG CGA GTT GCT TC
MucePGI2-F	*Pgi-2* (363 bp)	ATT GCT TTG CAC ATT GGC TTC
MucePGI2-R		AGA AGG CAC CAT TCG TGT TCC
For fugu (*Fugu rubripes*)		
FuruPGI1-F	*Pgi-1* (311 bp)	GCT CAT CCA CCA AGG AAC TC
FuruPGI1-R		TGT GCT CGT ACA TGG CAA TC
FuruPGI2-F	*Pgi-2* (387 bp)	TGC GTG TAA ACT ACC ACA CT
FuruPGI2-R		TCT TGT GCT CAT ACA TCG CT

mitochondrial genome data (Inoue et al., 2003). Tetrapods were excluded from this analysis because of their absence in this tree. Nucleotide sequence alignments of the coding region of *Pgi* genes (1650 bp, without ambiguous regions) from 10 ray-finned fishes plus hagfish were used. The GTR + Γ (Yang, 1994) model was selected as the best fitting model by the hLRTs (Table 3-3). The average overall accuracy of the reconstructed sequences (#1–#15) was 0.948 ± 0.003 SE. The pI values were estimated from the deduced amino acid sequences using the ProtParam tool (Gasteiger et al., 2005). The solvent-accessible surface area (SASA) of each amino acid residue was estimated with GETAREA 1.1 (Fraczkiewicz and Braun, 1998) for the dimeric PGI protein structure using a solvent radius of 1.4 Å (approximately the size of a water molecule). The X-ray refined crystal structure of Rabbit PGI (PDB ID: 1XTB; Lee and Jeffery, 2005) was used as a reference data. The structural portion of the PGI composed of amino acid residues with more than 20 Å^2 SASA was considered “molecular surface.” This boundary mostly agrees with other criteria based on the ratio of side-chain surface area to random coil value per residue (Fraczkiewicz and Braun, 1998). A three-dimensional graphical model of the PGI molecule was constructed using RasMol (Sayle and Milner-White, 1995).

3.2.7. 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 (550-amino-acid sequence of PGIs from 11 fishes and the known phylogenetic framework of ray-finned fishes; Inoue et al., 2003), 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 = –5851.41); the second-best matrix was Dayhoff (Dayhoff et al., 1978) (lnL = –5865.41). Using the JTT matrix (m_{ij}), transition rates between pairs of amino acids (P_{ij}) were calculated by the equation

Table 3-3.

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

Models	Likelihood	Likelihood ratio tests[1]			
	scores (lnL)	F81 + Γ [6]	HKY85 + Γ [7]	TN93 + Γ [8]	GTR + Γ [11]
JC69 + Γ [3]	-16039	46.0* (3)	790.9* (4)	835.8* (5)	852.4* (8)
F81 + Γ [6]	-16016	-	744.9* (1)	789.8* (2)	806.4* (5)
HKY85 + Γ [7]	-15644	-	-	44.9* (1)	61.5* (4)
TN93 + Γ [8]	-15621	-	-	-	16.6* (3)
GTR + Γ [11]	-15613	-	-	-	-

*$P < 0.01$

[1] The likelihood ratio test statistic (2ΔL) 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.

$$P_{ij} = \frac{m_{ij} f_i \mu_i}{\sum_{j=\mathrm{Ala}}^{\mathrm{Val}} m_{ij}}, \quad (i \neq j)$$

where f_i is the normalized frequency and μ_i is the relative mutability of each amino acid. The parameter f_i was estimated separately for the surface and interior portions of the inferred common ancestral protein of PGI-1 and PGI-2 to consider differential amino acid composition in different parts of the protein (Table 3-4). Based on the resultant P_{ij}, I estimated the theoretical ratio of the charge-changing substitutions to charge-neutral substitutions ($\Sigma P_{\text{charge-changing}}$:$\Sigma P_{\text{charge-neutral}}$) of the surface ($r_1$:$r_2$) and interior ($r_3$:$r_4$) portions of the PGI protein molecule under the assumption of random mutation.

According to the null hypothesis that all pairs of amino acid substitutions occur regardless of their spatial locations, the amino acid substitution events would be spatially distributed into the surface and interior portions of the PGI protein along the ratio of the numbers of amino acid substitution sites at the surface (132 sites) to the interior (80 sites) of the PGI protein since their gene duplication. Accounting for the spatial-differential amino acid composition as described above, the 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 = $132r_1$:$132r_2$:$80r_3$:$80r_4$.

Table 3-4.

Inferred normalized amino acid frequencies (f_i) of the surface and interior portions of the common ancestral protein of PGI-1 and PGI-2.

Residue	Normalized frequency	
	Surface	Interior
Ala (A)	0.05556	0.09494
Arg (R)	0.05128	0.01582
Asn (N)	0.08120	0.04747
Asp (D)	0.06410	0.03165
Cys (C)	0.00000	0.00316
Gln (Q)	0.05983	0.02215
Glu (E)	0.14530	0.01266
Gly (G)	0.05556	0.08861
His (H)	0.03846	0.03165
Ile (I)	0.01709	0.09177
Leu (L)	0.02564	0.14241
Lys (K)	0.15812	0.01899
Met (M)	0.00855	0.05063
Phe (F)	0.02137	0.08228
Pro (P)	0.04274	0.03165
Ser (S)	0.04274	0.04747
Thr (T)	0.07265	0.06013
Trp (W)	0.02991	0.01582
Tyr (Y)	0.01709	0.03165
Val (V)	0.01282	0.07911

3.3. Results

3.3.1. Duplication and subfunctionalization of the *Pgi* genes in teleost fishes

The present molecular phylogeny of vertebrate *Pgi* genes (Fig. 3-1) suggests that the *Pgi-1* and *Pgi-2* genes in teleost fishes resulted from a gene duplication event that occurred before the radiation of teleosts but after the separation of basal non-teleost ray-finned fishes (Fig. 3-1 arrow). The *Pgi* duplication appears to have derived from the ancient teleost genome duplication (Amores et al., 1998; Taylor et al., 2003; Jaillon et al., 2004) because the phylogenetic position of the *Pgi* duplication confirmed here is the same as that of the estimated teleost-specific genome duplication event (Chiu et al., 2004; Hoegg et al., 2004), and gene content around the *Pgi* locus in the human genome is partly conserved in the corresponding regions of both zebrafish *Pgi* loci (Fig. 3-2), which rules out the possibility of a tandem or single-gene duplication of the *Pgi* in teleosts. This condition allows us to study the duplicated *Pgi* genes without considering interlocus concerted evolution, which complicates the analysis of functional divergence in duplicated genes.

A reverse transcriptase-polymerase chain reaction (RT-PCR)-based expression analysis showed that the *Pgi* gene in non-teleost ray-finned fishes was expressed in all tissues examined (Fig. 3-3; for details, Fig. 3-4), confirming that this gene is the direct descendant of the ancestral unduplicated *Pgi* with no tissue specificity, as in tetrapods (Avise and Kitto, 1973; Dando, 1980; Kao and Lee, 2002). In contrast, the teleost *Pgi-1* gene was expressed mainly in internal organs, including the liver, heart, gill, brain, and kidney, and weakly in the muscle, whereas *Pgi-2* was expressed mainly in the heart and muscle. The differential expression patterns of *Pgi-1* and *Pgi-2* support the concept of subfunctionalization (Force et al., 1999; Lynch and Force, 2000A), which is the complementary loss of subsets in the expression organs of the ancestral gene. Thus, the *Pgi* gene family in ray-finned fishes is an

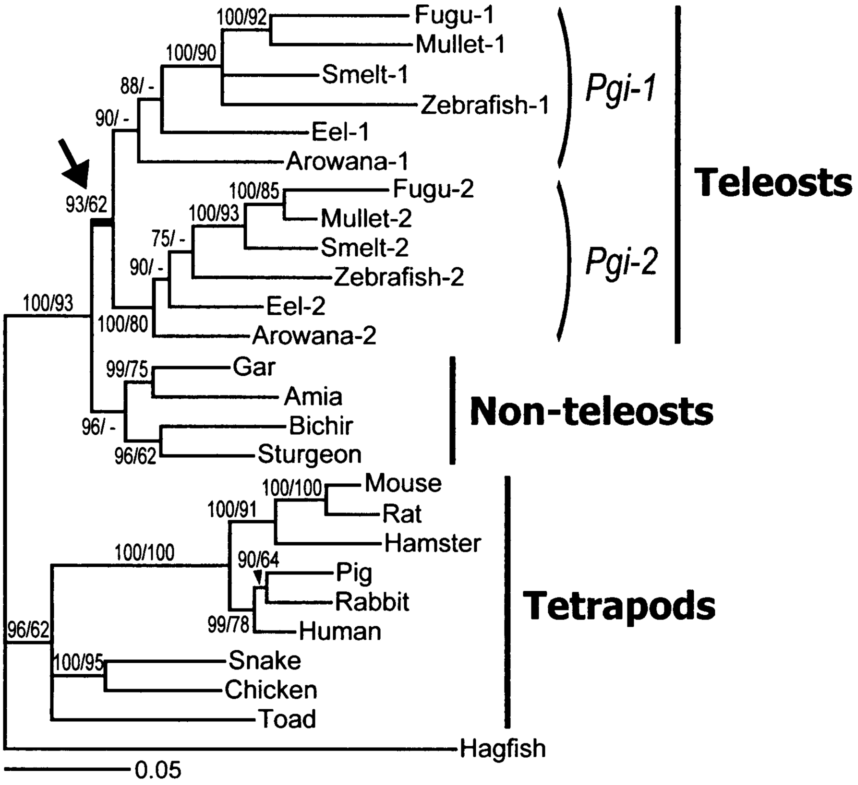

Fig. 3-1. Bayesian-maximum likelihood tree of *Pgi* genes derived from 20 vertebrates. Numbers indicate percent posterior probabilities for the Bayesian tree (left) and bootstrap support values for the maximum likelihood tree (right). Arrow denotes gene duplication. In cDNA clones, only one *Pgi* was identified from non-teleosts, whereas two *Pgi* were identified from teleosts. The two *Pgi* differed by about 20% in amino acid sequence, and were grouped into separate clades (*Pgi-1* and *Pgi-2*). In both clades, the gene relationships were consistent with the evolutionary relationships of the teleosts (Inoue et al. 2003; Miya et al. 2003; Lavoue et al. 2005).

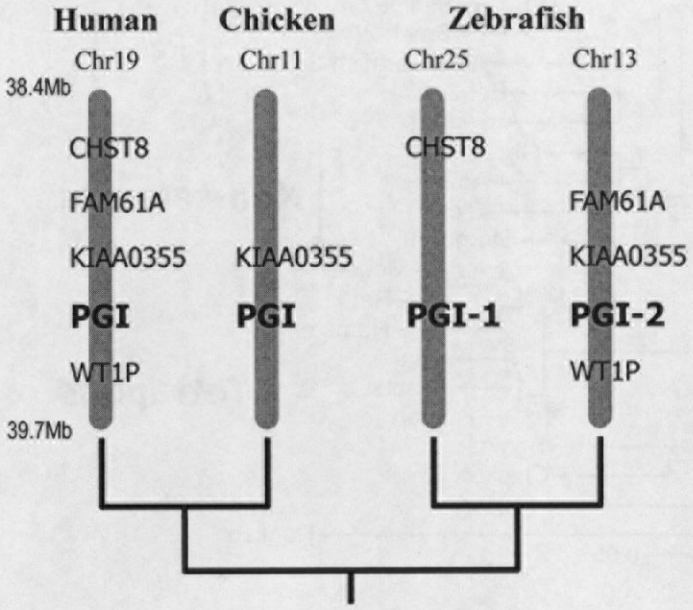

Fig. 3-2. Gene contents around the *Pgi* locus (or loci) in the human, chicken and zebrafish genomes according to the assembly versions of the human genome from October 2005 (NCBI 36), chicken genome from March 2004 (WASHUC 1) and zebrafish genome from March 2006 (Zv 6). The grey horizontal bars denote chromosomes with gene names showing their relative physical locations.

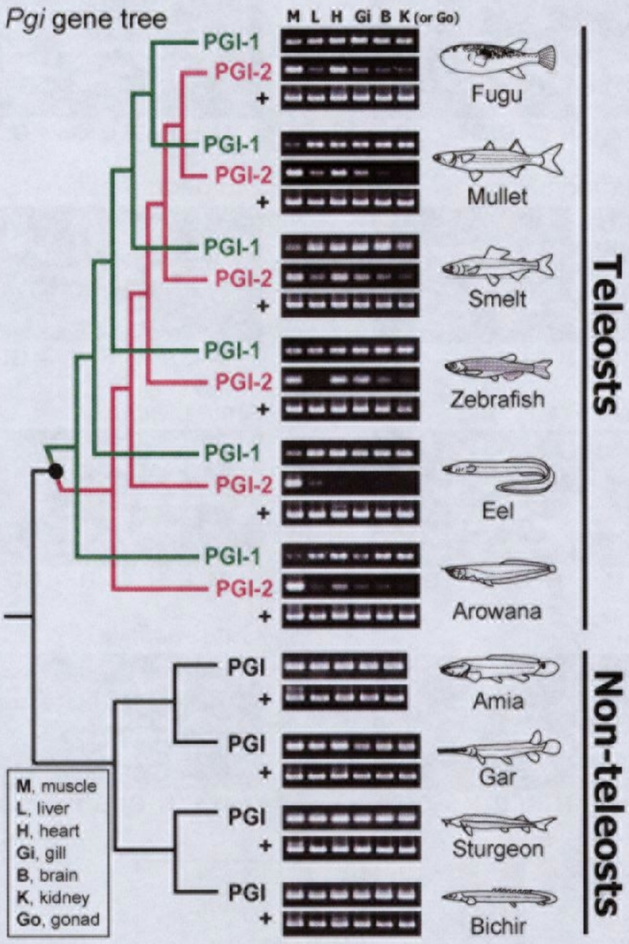

Fig.3-3. Partial-length gel images of the RT-PCR expression analysis of *Pgi* genes and positive control (β-actin) genes in ray-finned fishes. The tree in the left panel show the relationships among the *Pgi* genes inferred in this study. The black circle on the tree denotes the timing of *Pgi* gene duplication event. Letters indicate tissues: M, muscle; L, liver; H, heart; Gi, gill; B, brain; K, kidney; Go, gonad (in bichir and sturgeon). Full-length gels, including negative controls and size markers, are presented in Fig. 3-4.

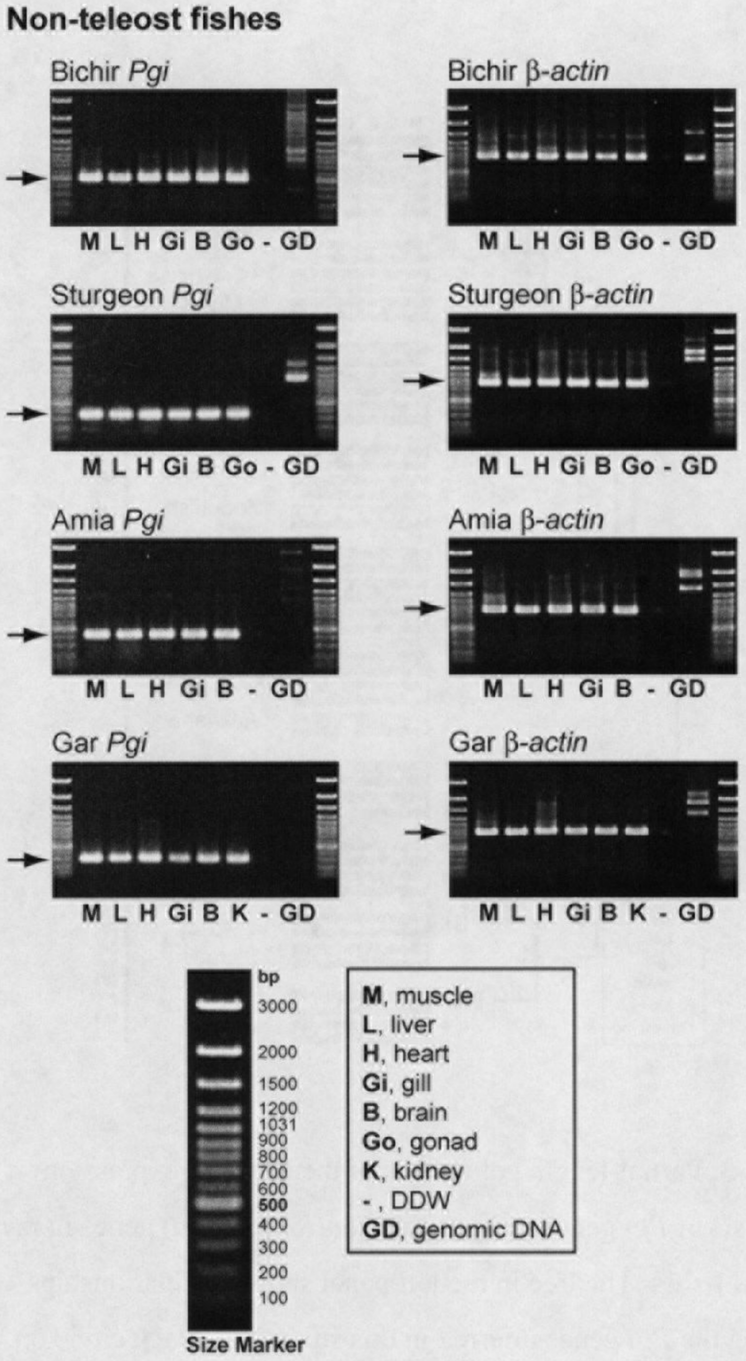

Fig. 3-4 (to be continued). Full-length gel images of the RT-PCR expression analysis of *Pgi* and β-*actin* genes. The letters below each gel image indicate the tissues tested: M, muscle; L, liver; H, heart; Gi, gill; B, brain; Go, gonad; K, kidney; -, sterilized deionized water; GD, genomic DNA.

Teleost fishes

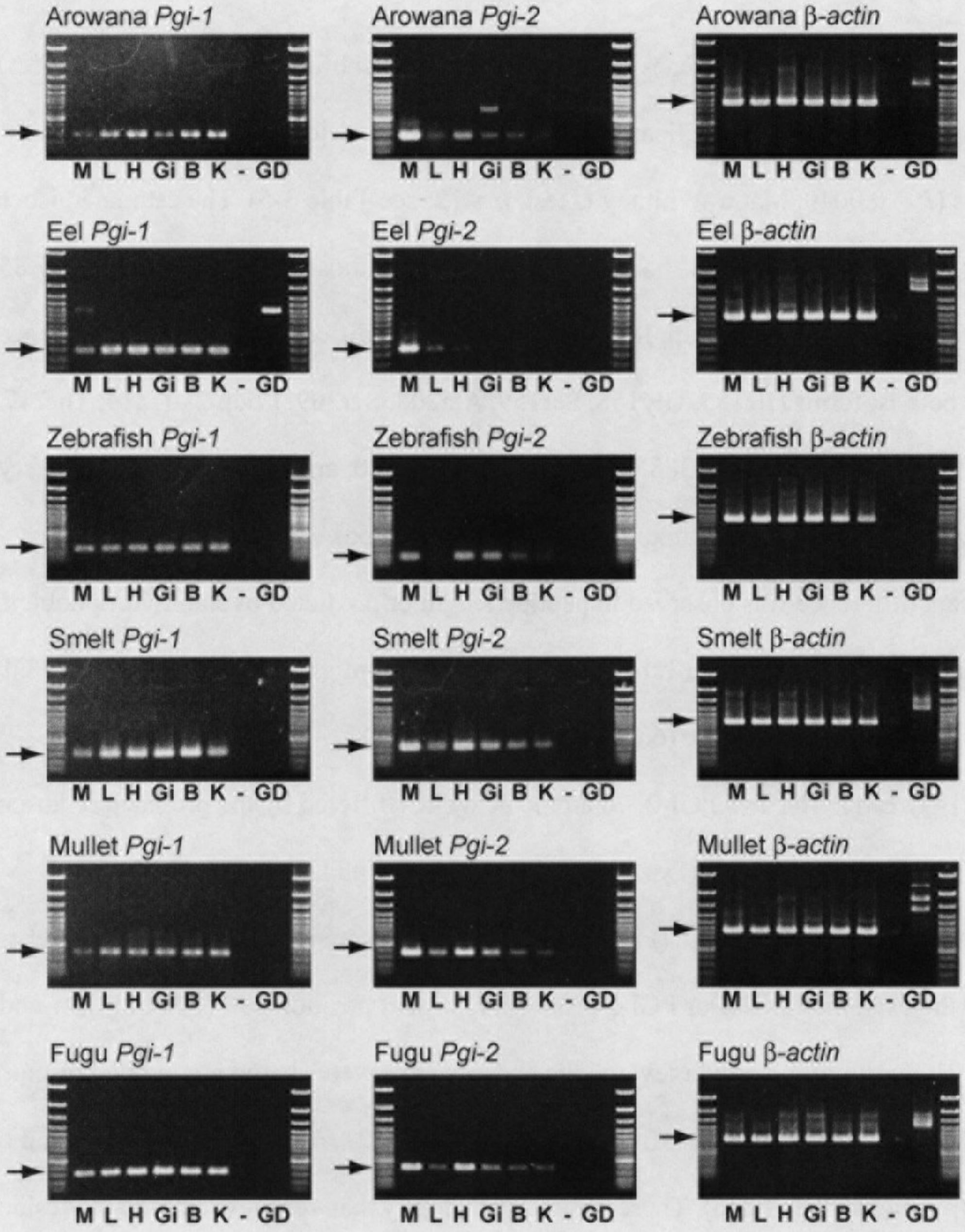

Fig. 3-4 (continued).

appropriate model for studying molecular evolution after subfunctionalization.

3.3.2. Evolution of the electric charges of duplicated PGI proteins in teleost fishes

The *Pgi-1* and *Pgi-2* genes in teleosts, which are subfunctionalized with respect to their expression, differed significantly in the predicted electric charge of their encoded proteins ($P = 0.0040$, Mann–Whitney U test, $n = 12$; see Table 3-5). The estimated isoelectric points (pI) of PGI-1 were 6.21–6.45 (average, 6.31), and those of PGI-2 were 6.75–7.85 (average, 7.17), with no overlap. In contrast to this clear difference, the PGI enzyme active sites in both isoforms (Ile156, Gly158, Ser159, Ala208, Ser209, Loop 210–214, Thr217, Arg272, Gln353, Glu357, His388, Gln511, Helix 512–520, and Lys518; Lee and Jeffery, 2005) were totally conserved among all fishes and tetrapods examined. Moreover, no significant difference was observed in peptide length or predicted overall hydrophobicity between the two isoforms (see Table 3-5). The estimated pI values for the ancestral PGI in non-teleosts were intermediate (6.62–6.84; average, 6.78).

Between PGI-1 and PGI-2, 76 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. 3-5B). These sites were not fixed for a unique charge state in the examined PGI-1 or PGI-2 proteins, except at position 294 (Gln in PGI-1 and Lys in PGI-2). Furthermore, only a few unique charged sites were shared among two or more genealogically related isoforms: five in PGI-1 (positions 27, 61, 78, 199, and 454) and two in PGI-2 (positions 17 and 226). These observations imply that very few amino acid residues were acquired specifically to the ancestral proteins of PGI-1 and/or PGI-2, and involved in differences in electric charge between current PGI-1 and PGI-2.

Table 3-5.

Biochemical parameters of vertebrate PGI 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]
Teleost fish PGI-1						
Fugu-1	552	-0.261	6.33	58	64	-6
Mullet-1	553	-0.282	6.30	57	63	-6
Smelt-1	553	-0.284	6.21	54	62	-8
Zebrafish-1	553	-0.265	6.45	53	58	-5
Eel-1	553	-0.359	6.36	57	63	-6
Arowana-1	553	-0.277	6.22	56	64	-8
Teleost fish PGI-2						
Fugu-2	553	-0.294	6.96	59	61	-2
Mullet-2	553	-0.265	7.85	61	60	1
Smelt-2	552	-0.337	7.36	64	64	0
Zebrafish-2	553	-0.304	6.82	59	61	-2
Eel-2	553	-0.285	6.75	58	61	-3
Arowana-2	553	-0.280	7.07	63	64	-1
Non-teleost fish PGI						
Sturgeon	555	-0.356	6.82	59	61	-2
Gar	555	-0.308	6.83	59	61	-2
Amia	555	-0.355	6.62	60	63	-3
Bichir	556	-0.278	6.84	58	60	-2
Tetrapod PGI						
Toad	553	-0.226	7.68	59	58	1
Snake	553	-0.237	8.72	62	57	5
Chicken	553	-0.255	8.34	64	61	3
Mouse	558	-0.294	7.75	61	60	1
Pig	558	-0.340	7.79	62	61	1
Rat	558	-0.285	7.38	61	61	0
Hamster	558	-0.322	7.08	59	60	-1
Rabbit	558	-0.292	7.11	58	59	-1
Human	558	-0.344	8.42	62	59	3

Jawless fish PGI						
Hagfish	554	-0.238	7.82	57	56	1

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

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

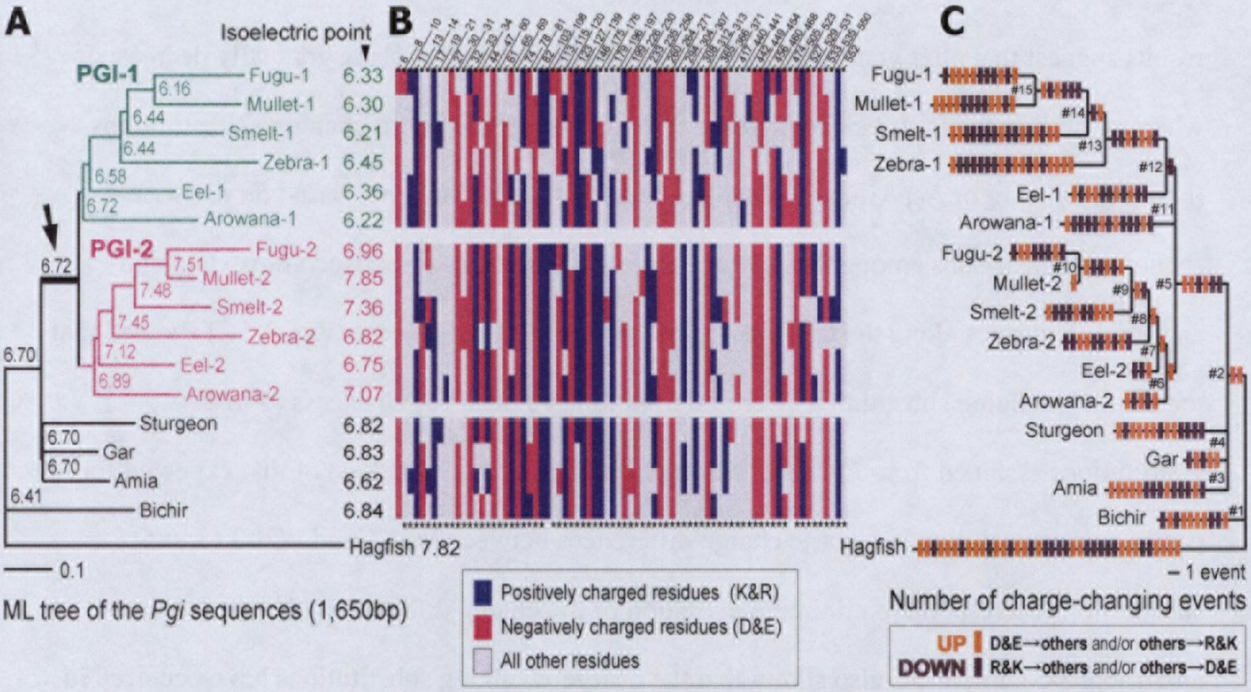

Fig. 3-5. Current states and inferred evolutionary process of electric charge of PGI isoforms. **(A)** Maximum likelihood tree of *Pgi* genes in ray-finned fishes inferred by BASEML (Yang, 1997) with known phylogeny (Inoue et al., 2003). Numbers indicate estimated pI. Arrow denotes gene duplication. **(B)** Amino acid sites that differ by the presence or absence of hydrophilic charged residues between current PGI-1 and PGI-2. Positively charged residues are colored blue; negatively charged residues, red; other residues, light grey. The numbers above refer to the amino acid positions of PGI (Lee and Jeffery, 2005). The stars below indicate sites located on the molecular surface. **(C)** Inferred charge-changing substitution events mapped over the PGI phylogeny. Orange and brown bars denote upward and downward direction of charge change, respectively.

The underlying process of the electric charge evolution of the PGI proteins can be inferred by ML sequence reconstructions (Yang, 1997) based on the recent ray-finned fish phylogeny (Inoue et al., 2003). I applied this approach and show the results in Fig. 3-5A. The results suggest that after gene duplication, pI values in the PGI-1 clade gradually decreased, whereas those in the PGI-2 clade increased. Next, 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. This result of assignment (Fig. 3-5C) showed that the charge-changing substitutions were inferred to have occurred in excess (5 to 28 substitutions assigned from *Pgi* duplication [#5] to tips of the branches) of that expected for parsimonious evolution in electric charge differences between PGI-1 and PGI-2 (3 to 5 substitutions correspond to minimum evolution of the charge differences; Figure 3-5C; see also Table 3-5). Fig. 3-5C also shows that the charge-changing substitutions have occurred in both directions (either upward or downward) on most branches at various amino acid sites (76 sites shown in Figure 3-5B). An analysis using parsimony yielded similar results (Fig. 3-6).

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

Based on 3-D structural information on the PGI protein molecule, I further examined whether the inferred charge-changing substitutions were actually involved in the evolution of electric charge. The results of this analysis on the inferred substitution sites and number of substitutions are shown in Fig. 3-7A and 3-7B, respectively. Fig. 3-7A shows that the inferred charge-changing substitution sites after the *Pgi* gene duplication (colored in magenta) were concentrated at the surface of the PGI molecule, in contrast to the inferred charge-neutral substitution sites (colored in dark gray) that contribute little or nothing to net protein charge. The inferred number of charge-changing and charge-neutral substitutions that can potentially

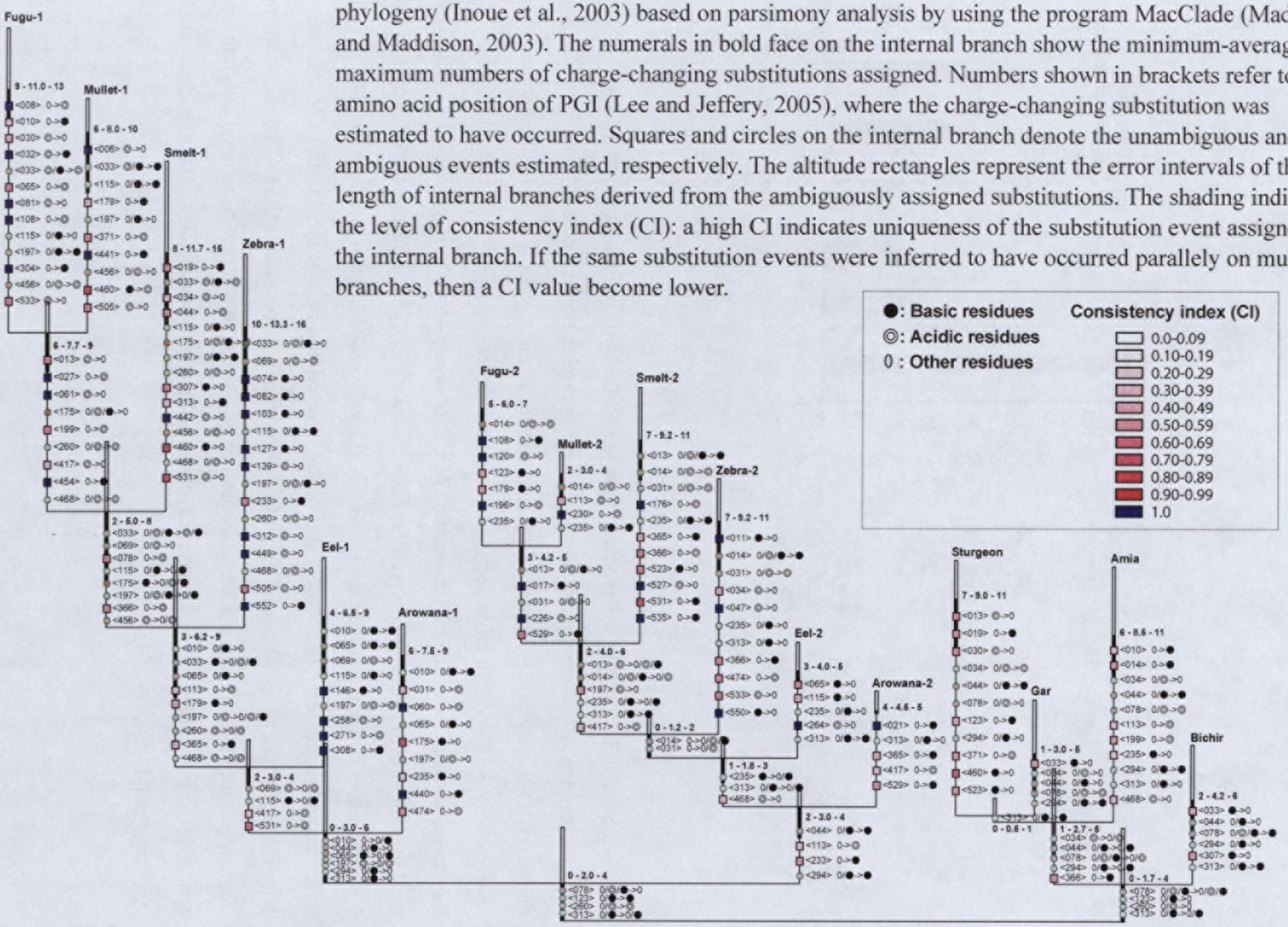

Fig. 3-6. Charge-changing substitutions occurred in the evolution of fish PGIs assigned on the known phylogeny (Inoue et al., 2003) based on parsimony analysis by using the program MacClade (Maddison and Maddison, 2003). The numerals in bold face on the internal branch show the minimum-average-maximum numbers of charge-changing substitutions assigned. Numbers shown in brackets refer to the amino acid position of PGI (Lee and Jeffery, 2005), where the charge-changing substitution was estimated to have occurred. Squares and circles on the internal branch denote the unambiguous and ambiguous events estimated, respectively. The altitude rectangles represent the error intervals of the length of internal branches derived from the ambiguously assigned substitutions. The shading indicates the level of consistency index (CI): a high CI indicates uniqueness of the substitution event assigned on the internal branch. If the same substitution events were inferred to have occurred parallely on multiple branches, then a CI value become lower.

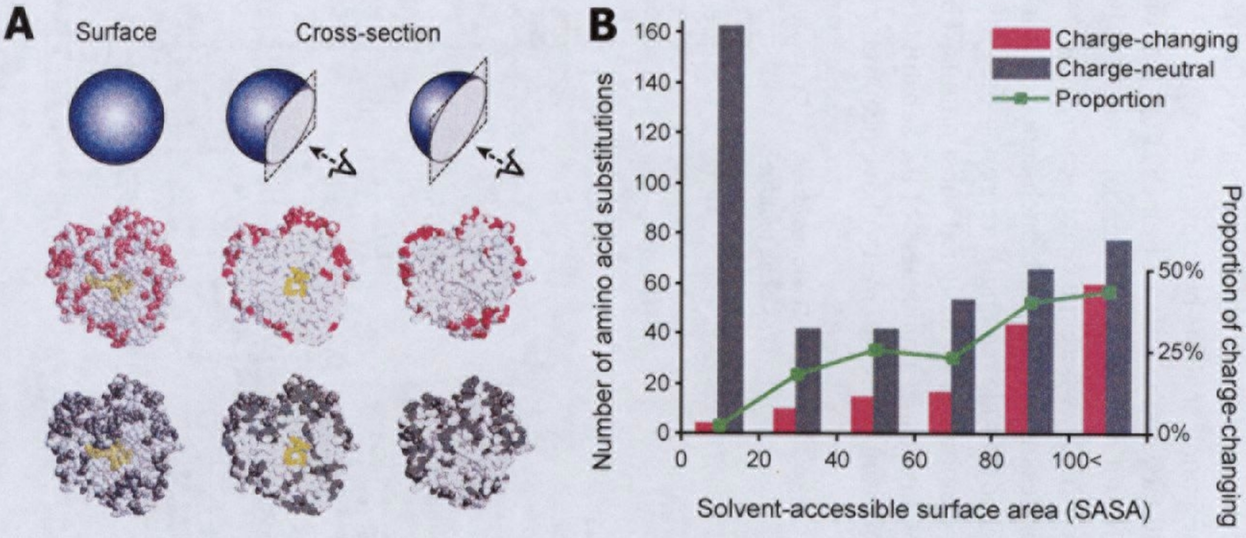

Fig. 3-7. Spatial locations of inferred amino acid substitutions in PGI structure. **(A)** ML-inferred charge-changing (CC) substitution sites after the *Pgi* duplication are colored magenta; charge-neutral (CN) substitution sites, dark grey; enzyme active sites, yellow. Full molecular models are shown on the left, and two cross-sections are shown center and right. The inferred CC sites localize to the surface of the PGI molecule (73 CC sites / 234 total surface sites, 3 CC sites / 316 total interior sites; $P = 0.0000$, two-tailed Fisher's exact test), in contrast to the inferred CN sites (106 CN sites / 234 total surface sites, 183 CN sites /316 total interior sites; $P = 0.1040$, two-tailed Fisher's exact test). **(B)** Histograms of the inferred numbers of the CC and CN substitutions after the *Pgi* duplication. The solid green line denotes the proportion of the CC substitutions per total substitutions within the site classes based on solvent accessibility (horizontal axis): this proportion significantly increases with solvent-accessible surface area ($P = 0.0000$, Cochran-Armitage trend test, $n = 584$).

occur at identical site classes also followed the same trend (Fig. 3-7B).

However, because water-soluble proteins such as PGI are generally surrounded by a hydrophilic shell containing a high density of polar residues, it is natural to expect that random mutations cause charge-changing substitutions to occur more frequently on the protein surface without any selection. Considering this expected mutation bias, further analysis was performed (Table 3-6; for details, see *section 3.2. Materials and Methods*). This comparison of theoretically expected and ML-inferred numbers of charge-changing and charge-neutral substitutions imply that charge-neutral substitutions have occurred more frequently than expected at the molecular surface [ML-inferred value, 63.1% = 277/(162 + 277); expected value, 55.5% = 230.17/(184.46 + 230.17)], consistent with the general observation that molecular evolutionary rates are faster at the surface than in the interior portions of water-soluble proteins (Bustamante et al., 2000; Choi et al., 2006). However, what is most important in this table is that the proportion of charge-changing substitutions concentrated at the surface of the PGI molecule is much greater than that expected by chance [ML-inferred value, 97.2% = 141/(4 + 141); expected value, 78.8% = 133.45/(35.92 + 133.45)]. These charge-changing substitutions do not appear to be derived from differential neutral evolution of base composition or codon usage between *Pgi-1* and *Pgi-2* genes, as demonstrated by the fact that GC content and codon usage frequencies are not significantly different between *Pgi-1* and *Pgi-2* (GC content: $P = 0.0782$, Mann–Whitney U test, $n = 12$; rank order of codon usage: $r_s = 0.9509$, $n = 64$).

Table 3-6.

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	4	141	162	277	584
Theoretical prediction	35.92	133.45	184.46	230.17	584.00
P value*	0.00000		0.02408		

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

3.4. Discussion

The results of phylogenetic analysis, RT-PCR-based expression analysis, and sequence comparison of *Pgi* genes in teleost fishes suggest that after subfunctionalization of the duplicated *Pgi* genes in the ancestor of teleost fishes, the electric charges of the PGI-1 and PGI-2 proteins diverged. This evolution can be interpreted according to the sub-neofunctionalization model of 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 alters the selective environment at each locus, resulting in structural fine-tuning or adaptation of the encoded proteins by positive selection. That is, the divergent evolution of the electric charges in the duplicated PGI isoforms is the consequence of specialization for the specific function (glycolysis or gluconeogenesis) or distinct cellular environment of tissues where each isoform is predominantly expressed (see Fig. 3-3), as suggested for other water-soluble proteins (Frolow et al., 1996; Merritt and Quattro, 2001; Zhang, et al., 2002; Zhang, 2006).

The present comparative evolutionary analysis implies that since the gene duplication event, the electric charges of the two PGI isoforms changed steadily through many charge-changing substitutions in both directions of charge change; only a few charged amino acid sites were specific to PGI-1 or PGI-2 (Fig. 3-5). Such charge-changing substitutions concentrated at the surface of PGI molecule (Fig. 3-7) were inferred to have occurred much more frequently than expected in the parsimonious evolution of electric charge difference between the two isoforms (see Fig. 3-5B and Table 3-5). From these observations, two possible scenarios are proposed for the evolution of protein charge in duplicated PGI isoforms: protein charges in PGI-1 decreased gradually, while those in PGI-2 increased; alternatively, charge divergence between PGI-1 and PGI-2 was completed soon after the duplication and before the radiation of teleosts, followed by maintenance of the protein charges under purifying selection, while stochastic charge-changing substitutions by drift occurred among lineages. In either scenario, We can conclude that the surface charge evolution of PGI proteins was not driven by strong selection on individual amino acid sites leading to permanent fixation of a particular residue, but rather was driven by weak selection on a large number of amino acid sites and consequently by steady directional or purifying selection on the overall structural properties of the protein, which is derived from many modifiable sites. This mode of molecular evolution agrees with the understanding that most proteins are substantially tolerant of a broad spectrum of substitutions and thus may harbor many amino acid sites available for evolutionary modification (Lynch, 2005). This study provides the first plausible evidence of protein evolution through such selection.

The mode of molecular evolution proposed in this study would be difficult to find using existing methods that detect strong selection for particular substitutions. I applied such an analysis to identify positively selected sites after *Pgi* gene duplication using the program DIVERGE version 1.04 (Gu and Vander Velden, 2002); however, the results were not clear (data not shown). Further analysis using the program CODEML (Yang, 1997) did not detect the acceleration of the rate of nonsynonymous substitution, not showing selection for amino acid changes (estimated ω was 0.01 to 0.23). Even if, in general, a significant excess of amino acid change is detected, such methodology itself cannot rule out possible confounding effects, or alternative interpretations, particularly the relax of purifying selection (Hughes, 2007). In a previous study, an analysis using the program HonNew (Zhang, 2000) also failed to detect selection for charge-changing substitutions in teleost PGIs, leading to the conclusion that the charge change in the duplicate PGIs of teleosts may be selectively neutral (Kao and Lee, 2002). The mode of molecular evolution presented here, in which diverse evolutionary resolutions exist at the level of a primary sequence that corresponds to a certain selective pressure on a protein property, may be relevant to various cases of adaptive modification in proteins, such as hydrophobic properties, molecular size, and electric charge. This may be an important pathway underlying physiological adaptation, along with protein evolution by simple amino-acid changes, gene deletion or silencing, and possibly *cis*-regulatory changes (Hoekstra and Coyne, 2007; Hughes, 2007).

In summary, this study provide the evidence that relatively weak selection on a large number of amino acid sites drives the evolution of novel charge-state of duplicated phosphoglucose isomerases, which are subfunctionalized in teleost fishes. Such mode of adaptive molecular evolution, which was hardly recognizable by existing analytical methods aiming to detect strong selection on individual amino acid changes, may play a substantial role in the evolution of novel proteins.