Molecular Phylogeny and Functional Genomics of β-Galactoside α2,6-Sialyltransferases That Explain Ubiquitous Expression of st6gal1 Gene in Amniotes*1

Daniel Petit1, Anne-Marie Mir5, Jean-Michel Petit5, Christine Thisse5, Philippe Delannoy1, Rafael Oriol1, Bernard Thisse5, and Anne Harduin-Lepers51

From the 1Unité de Génétique Moléculaire Animale, Université de Limoges Faculté des Sciences et Techniques, INRA UMR 1061, 123 Avenue Albert Thomas, 87060 Limoges, France, the 5Unité de Glycobiologie Structurale et Fonctionnelle, Université Lille Nord de France, Lille 1, CNRS UMR 8576, IFR 147, 59655 Villeneuve d’Ascq, France, the 5Department of Cell Biology, School of Medicine, University of Virginia, Charlottesville, Virginia 22908, and the 5Unité de Microenvironnement et Physiologie de la Différenciation, Université de Paris Sud XI, INSERM U1004, 16 Avenue Paul Vaillant-Couturier, 94807 Villejuif, France

Sialyltransferases are key enzymes in the biosynthesis of sialylglycoconjugates that catalyze the transfer of sialic residue from its activated form to an oligosaccharidic acceptor. β-Galactoside α2,6-sialyltransferases ST6Gal I and ST6Gal II are the two unique members of the ST6Gal family described in higher vertebrates. The availability of genome sequences enabled the identification of more distantly related invertebrates’ st6gal gene sequences and allowed us to propose a scenario of their evolution. Using a phylogenomic approach, we present further evidence of an accelerated evolution of the st6gal1 genes both in their genomic regulatory sequences and in their coding sequence in reptiles, birds, and mammals known as amniotes, whereas st6gal2 genes conserve an ancestral profile of expression throughout vertebrate evolution.

Sialyltransferases described in higher vertebrates are glycosyltransferases that mediate the transfer of sialic acid residues from activated sugar donors (CMP-β-Neu5Ac,2 CMP-β-Neu5Gc, and CMP-β-KDN) to terminal non-reducing positions of oligosaccharide chains of glycoproteins and glycolipids (reviewed in Refs. 1–3). Classically, the vertebrate sialyltransferase superfamily is divided into four families, namely the ST6Gal, ST3Gal, ST6GALNAC, and ST8Sia, depending on the glycosidic linkage formed and the monosaccharide acceptor used.3 Members of the mammalian and avian ST6Gal family catalyze the transfer of sialic acid residues to the terminal galactose residues of the type 2 disaccharide (Gal(NAc)β1,4GlcNAc), resulting in the formation of an α2–6 glycosidic linkage (for reviews, see Refs. 3–10). Unlike the other sialyltransferase families, this family comprises only two paralogs in the human genome named ST6GAL1 and ST6GAL2, respectively (1, 2). The human ST6GAL1 gene is ubiquitously expressed in a broad variety of tissues, whereas the ST6GAL2 gene is expressed in a tissue-specific (adult brain) and stage-specific (embryonic) manner. Mammalian st6gal1 gene expression is regulated by multiple promoters governing the expression of several transcripts encoding identical polypeptide enzyme, and high levels of mRNA are detected in hematopoietic cells and in liver (11–13).

Sialylated α2,6-lactosaminyl structures (Neu5Acα2–6Galβ1–4GlcNAc; sia6LacNAc) found on N-glycosylproteins and also, to a lesser extent, on O-glycosylproteins, glycolipids, and free oligosaccharides (14) are involved in a highly specific recognition phenomenon (15). In the mammalian immune system, B cells highly express ST6Gal I (11, 16, 17), and sialylated α2,6-lactosaminyl structures generated on CD45 and immunoglobulin M (IgM) are the preferred ligands of CD22 (Siglec 2), a sialic acid-binding Ig-like lectin found exclusively on B-lymphocytes and involved in B cell immunologic activation and signaling as evidenced in KO mice (16, 18, 19). Overexpression of ST6Gal I has been reported in several human malignancies, and clinical and experimental studies suggest a positive correlation between high ST6Gal I levels and invasive behavior of cancer cells (14, 20). Integrin-mediated adhesion is based on protein interactions, and binding can be significantly modulated by sia6LacNAc structures on β1-integrin in vivo and in vitro in cancer cells, leading to enhanced cell motility and invasiveness (21–23). ST6Gal I plays a role in inflammation (24, 25), and in mammals, transient up-regulation occurs during acute phase reaction when the organism experiences trauma or infection (26, 27). Finally, in contrast to avian and other mammalian influenza viruses, human influenza virus A and B prefer the α2,6-linked sialic acid found in abundance in human upper airways over the α2–3-linked sialic acid (28–30). On the other hand, the ST6Gal II function remains unknown.

st6gal homologs have been cloned from several higher vertebrate species (1). Furthermore, a ST6Gal cDNA named DSIaT was cloned from Drosophila melanogaster (31), suggesting that the ST6Gal family was present in insects, although not much Neu5Ac could be detected (32–34). DSIaT is detected almost
Evolutionary Study of st6gal Gene Expression in Vertebrates

exclusively in central nervous system (CNS) neurons in the embryonic stage 17, in the optic lobe of third instar larva, and in adult head (35). Targeted disruption of the DSiaT gene results in a neurological phenotype, suggesting that DSiaT modulates the nervous system function of voltage-gated sodium channel (36). Because the mammalian st6gal2 gene is detected mainly in CNS as well, it has been suggested that ST6Gal II might have conserved an ancestral function, whereas ST6Gal I would have developed new functions in vertebrates. Further understanding of the evolutionary history of st6gal genes through molecular phylogenetic analysis will shed light on the functions of these genes maintained during evolution.

In the era of genomics, we have developed the ability to investigate the genomic sequences of the sialyltransferase genes that modify glycans in different animal lineages, thus providing a powerful means of reconstructing the evolutionary history of sialylation, determining key genetic events in the establishment of glycan sialylation machinery (2, 37). In the present work, we address the fate of vertebrate st6gal genes. We take advantage of the wealth of data provided by complete genome projects to refine the molecular relationship of ST6Gal and to address st6gal gene evolutionary trends in terms of gene gain and loss and also translocation and mutation rate, those mechanisms that were instrumental in establishing modern functions of vertebrate ST6Gal I and ST6Gal II. We have traced the environment of these genes (i.e. the set of orthologous genes around st6gal gene loci). In parallel, we have compared the expression pattern of st6gal genes in the vertebrate lineage, through molecular cloning of bony fish (teleost) (Danio rerio) and amphibian (Silurana tropicalis) st6gal. Our phylogenetic and expression analysis provide valuable insights into st6gal gene evolution in vertebrates and a model of duplication events whereby the st6gal1 genes have undergone neofunctionalization in higher vertebrates.

EXPERIMENTAL PROCEDURES

In Silico Sialyltransferase Retrieval—Only eukaryote sequences were considered for this study. Homologous st6gal sequences were searched through exploration of all genomic and expressed sequence tags (ESTs) available from general databases, such as NCBI (see the BLAST (Basic Local Alignment Search Tool) Web site) for the green lizard Anolis carolinensis, DDBJ, or Ensembl, or in specialized databases JGI for Branchiostoma floridae, the Genome Sequencing Center at the Washington University School of Medicine (St. Louis, MO) for the lamprey Petromyzon marinus, KEGG GENES (38–40), the Genome Sequencing Center at the Baylor College of Medicine for Homo sapiens and the sea urchin Strongylocentrotus purpuratus, and the Institute of Molecular and Cell Biology for the elephant shark Callorhinchus milii using BLASTN, TBLASTN, and PSI-BLAST (41) with default parameters (an e-value cut-off at 0.01 was used in all BLAST searches). Human and mouse sequences were used as first queries in the first round of search. The assignment of these sequences to ST6Gal was determined by the specific motifs that are hallmarks of this family (1, 42). All genomic sequences allowing generation of a complete catalytic domain were considered. Splice site prediction analysis was achieved at the Berkeley Drosophila Genome Project. The structure of the genes, in terms of exon/intron boundaries, was deduced from several non-exclusive strategies: (i) comparing the boundaries proposed by Genscan (MIT server), (ii) comparing EST from genomic assemblages (scaffolds or contigs), (iii) comparing the boundaries to those present in known genes.

Phylogenetic Analysis—The alignment of amino acid sequences was conducted using ClustalX software (43). The selection of informative sites was helped by G-BLOCKS (44) with the options of less stringent selection. Phylogeny trees were produced by maximum likelihood (ML) using PHYML, version 2.4.4 (45), with the Jones-Taylor-Thornton (JTT) model of amino acid substitution, neighbor joining (NJ), and minimum evolution (ME) using MEGA4.0 (46), and bootstrap percentages were calculated from 2000 replicates. The numbers of site changes in each branch were calculated with the Protpars program included in the PHYLIP Package (47), using 228 sites, under the constraint of the user tree produced by ME (see Refs. 48 and 49 for details).

The calibration used for dating the divergence between ST6Gal I and ST6Gal II in vertebrates was as follows: amphioxus/vertebrates, 650 MYA (50); lamprey/gnathostoma, 575 MYA; gnathostoma/osteichthians, 460 MYA; osteichthians/other vertebrates, 450 MYA; tetrapods/actinopterygians, 360 MYA; amniotes/other vertebrates, 310 MYA; genome duplication in teleosts (R3), 320 MYA (51). We calculated the regression equations between linearized branch Pearson’s correlations, and associated probabilities were calculated with PAST version 2.01 (52).

Synteny Analysis and Paralogon Detection—Synteny between vertebrate st6gal and related genes in invertebrates was assessed by chromosomal walking and reciprocal BLAST searches of genes adjacent to st6gal loci in human (HSA), mouse (MMU), chicken (GGA), medaka (OLA), zebrafish (DRE), Takifugu rubripes (TRU), and amphioxus (BFL) genome databases (Ensembl). The detection of paralogous blocks (53) was done using the latest Ensembl data set (version 5.28). The Web site for these paralogons (see the Trinity College Dublin Web site) offers the possibility to carry out block detection in humans with self-defined parameters.

Expression Analysis—Unigene at the NCBI data base was used to quantify the number of ESTs identified for each tissue in the following species: H. sapiens, Mus musculus, Gallus gallus, S. tropicalis, and D. rerio. In order to homogenize the different overall values among organisms, we divided the number of ST6Gal ESTs by the total number of ESTs per tissue. Second, we removed the tissues for which only one organism was recorded. Third, the table containing 22 columns (tissues or developmental stages) and $5 \times 2$ (species $\times$ st6gal1 and st6gal2 genes) lines was submitted to a principal component analysis (PCA) using PAST 2.01 (52). According to the method described by Ermonval et al. (54), PCA allows projecting the data set onto a two-dimensional plan, each column factor represented by a vector according to pair-wise correlations; the higher the correlation between two factors, the more acute the angle between the vectors. In this plan, the st6gal1 or st6gal2 genes corresponding to a given species are projected in the direction of their greatest values. The EST ratios per tissue were multiplied by 10$^6$ and log-transformed to normalize the distri
bution and then submitted to a two-way clustering using Euclidean distance as measure of similarity, using PAST 2.01. The coloration intensity of each case in the table was in proportion to the values.

**Animals and Maintenance**—Zebrafish (*D. rerio*) and clawed frog (*S. tropicalis*) were maintained in our aquatic biology facility, as described previously (55, 56). All experimental procedures adhered to the CNRS guidelines for animals use.

**Isolation of RNA, cDNA Synthesis, and PCR Analysis**—Total RNA was extracted from various *S. tropicalis* and *D. rerio* tissues using the nucleospin RNA II kit (Macherey-Nagel, Hoerdt, France). A proteinase K digestion step (55 °C, 10 min) and phenol/chloroform extraction were inserted into the protocol after Dounce homogenization of the tissues and before column purification of total RNA. Cellular RNA was quantified using a NanoDrop® ND-1000 UV-visible spectrophotometer (NanoDrop Technologies, Wilmington, DE). RNA integrity was further assessed using the RNA 6000 Nano LabChip® kit on an Agilent Bioanalyzer (Agilent Technologies, Stratagene, La Jolla, CA). For subsequent PCR amplifications, first strand cDNA was synthesized from total RNA using an oligo(dT) primer and the AffinityScript Q-PCR cDNA synthesis kit according to the manufacturer’s protocol (Agilent Technologies). Based on the nucleic acid sequences determined in silico, oligonucleotide primers were designed (Eurogentec, Herstal, Belgium) in the open reading frame (see supplemental Fig. 5).

PCR amplifications were carried out with the Taq core kit DNA polymerase (Qiagen, Courtaboeuf, France) or Jena DNA polymerase (Jena Bioscience, Euromedex, Souffelweyersheim, France) using buffer solutions provided by the manufacturer. Annealing temperatures ranged from 48 to 55 °C, and amplified fragments were subjected to 2% agarose gel electrophoresis, visualized by ethidium bromide, gel-extracted, and subcloned in the pCR®2.1-TOPO vector (TOPO TA Cloning, Invitrogen, Jolla, CA). For subsequent PCR amplifications, first strand cDNA was synthesized from total RNA using an oligo(dT) primer and the AffinityScript Q-PCR cDNA synthesis kit according to the manufacturer’s protocol (Agilent Technologies). Based on the nucleic acid sequences determined in silico, oligonucleotide primers were designed (Eurogentec, Herstal, Belgium) in the open reading frame (see supplemental Fig. 5).

**Evolutionary Study of st6gal Gene Expression in Vertebrates**

To test the conservation of st6gal expression, we compared the expression for the *st6gal1* and *st6gal2* genes in different species. The expression level of the *st6gal1* and *st6gal2* genes in the species studied were compared by RT-qPCR using primers that were specific to each gene. The expression level of the genes was normalized to the *β*-actin gene, which was used as an internal control. The values were then analyzed by the 2^{-ΔΔCt} method to determine the relative expression levels. The expression levels were then compared between the species studied.

**RESULTS**

**Identification of ST6Gal-related Genes in Bilaterians**

In order to identify putative genes encoding proteins with significant similarity to the known mammalian st6gal genes in animals with bilateral symmetry (bilaterians), we carried out a BLAST search on various invertebrate and vertebrate nucleotide databases using the known ST6Gal sequences. The search was based on the fact that the highly conserved sialylmotif peptide consensus sequences (L, S, III, and VS) are characteristic of all animal sialyltransferases and consequently serve as hallmarks for their identification.

A broad phylogenetic distribution of st6gal genes was observed in multicellular animals (metazoans). It should be noted that a short EST (NCBI, EST division: EC377350) from the sponge *Oscarella carmella* is attributable to ST6Gal. Despite an extensive examination of EST and whole genome shotgun sequences in data banks (JGI), no homologous st6gal gene was identified in the cnidaria *Nematostella vectensis*, the lophophorechozoa (polychete annelid *Capitella teleta* and mollusk *Lottia gigantea*), the hymenoptera insects *Apis mellifera* and *Nasonia vitripennis*, or in the nematoda *Caenorhabditis elegans* genome. It appears that among bilaterian animals developing first the mouth (protostomes), only one copy of *st6gal* gene sequence was retrieved from arthropods, like arachnida (*Ixodes scapularis* and *Varroa destructor*), crustacea (*Daphnia pulex* and *Calligus rogercresseyi*), and insects diptera (*D. melanogaster*, *Anopheles gambiae*, *Aedes aegypti*, and *Culex quinquefasciatus*), homoptera (*Acyrthosiphon pisum*), lepidoptera (*Bombyx mori*), phthiraptera (*Pediculus humanus corporis*), and coleoptera (*Tricholium castaneum*). Among bilaterian animals developing first the anus (deuterostomes), we found one copy of the *st6gal* gene in the hemichordata *Saccoglossus kowalevskii* and two copies in the Amphioxus (*B. floridae*), but none was found in the sea urchin (*S. purpuratus*) or in the tunicates *Ciona intestinalis* and *Ciona savignyi*. In vertebrates, most examined genomes contain two members the *st6gal1* and *st6gal2* paralogous genes, except in the lamprey *Petromyzon marinus*, where three *st6gal* copies were found. In teleosts, we also describe three members named *st6gal1*, *st6gal2*, and *st6gal2-1* in the zebrafish (*D. rerio*) genome. In order to gain further insights into lower vertebrate *st6gal* genes, we carried out a BLAST search on the *st6gal* gene from zebrafish and *D. rerio*.
Evolutionary Study of st6gal Gene Expression in Vertebrates

Molecular Phylogeny Analysis

**Catalytic Domain**—As a first step in the analysis, we assessed the orthology of the catalytic domain of vertebrate and invertebrate ST6Gal-related protein sequences by multiple sequence alignments with ClustalW (supplemental Fig. 2). The G-BLOCKS server evidenced 200 informative sites to construct the phylogenetic trees. The three tested methods to infer ST6Gal phylogeny, NJ, ME, and ML, using JTT as transition model of amino acid substitution: 36 ST6Gal sequences, 11 vertebrate ST6Gal I, 11 vertebrate ST6Gal II. G-BLOCKS selected 200 positions. Bootstrap values were calculated from 500 replicates, and values of >50% are reported at the left of each divergence point. The tree is rooted with the invertebrate arthropod sequences as the outgroup.

In order to estimate the time of divergence of the vertebrate st6gal gene subfamilies, we reconstructed linearized trees for duplicate genes under the assumption of a molecular clock using MEGA4.0 (60). The results obtained with NJ, ME, and ML are given in Fig. 1 and give an estimated divergence time in the range of 473 MYA by ME (Fig. 1B), 499 MYA by NJ, and 508 MYA according to ML.

We also observed on the phylogenetic tree that the branch lengths in the vertebrate ST6Gal I clade were longer than in the vertebrate ST6Gal II clade (Fig. 1A). We thus tested the significance of these differences for each internal branch (e.g. from the ancestor of osteichthyan to the ancestor of teleosts). For each branch, we counted the numbers of site changes using the parsimony program Protpars in PHYLIP in the ST6Gal I and ST6Gal II sequences of the catalytic domain (Table 1). The χ² tests show that there is a highly significant accumulation of mutations in the ST6Gal I branches leading to mammals and to teleosts and to a lesser extent to amphibians, relative to ST6Gal II branches. In contrast, we observe an accumulation of substitutions in the branch leading to osteichthyans ST6Gal II compared with the ST6Gal I counterpart.

In order to better understand the significance of these changes, we also compared the substitution numbers in the conserved motifs between ST6Gal sequences for each branch (Table 2). The greatest amounts were observed in the sialylmotifs L and S and in the family motif b (1), with a regular excess of changes found in ST6Gal I sequences; it concerns the sialylmotifs L and family motif b in the transition amnion-mammals, the sialylmotif L in the transition tetrapods-amnion, and the family motif b and sialylmotif S in the transition osteichthyans-teleosts.

**NTerminal of ST6Gal**—The length of the protein sequences encoded by the first exon of vertebrate st6gal genes, encom-

---

**TABLE 1**

|                      | ST6Gal I | ST6Gal II | χ²       | Significance |
|----------------------|----------|-----------|----------|--------------|
| Amniotes/mammals     | 34       | 8         | 16.10    | ***          |
| Tetrapods/amnion     | 28       | 13        | 5.49     | *            |
| Tetrapods/Xenopus    | 58       | 30        | 8.91     | **           |
| Osteichthyans/tetrapods | 25     | 23        | 0.08     | NS           |
| Osteichthyans/teleosts | 48     | 17        | 14.78    | ***          |
| R2/vertebrates       | 18       | 32        | 3.92     | *            |

---

*Significance

The table above shows the numbers of site changes in the catalytic part (228 sites) during transitions of vertebrate evolution.
passing the cytoplasmic and transmembrane domains and the stem region, varies from 201 amino acid residues in the human ST6Gal I sequence to 329 amino acids in the fugu ST6Gal II sequence. Multiple sequence alignments of this region using ClustalW revealed weak sequence conservation upstream of the tryptophan residue Trp$^{96}$ and Trp$^{208}$ in human ST6Gal I and ST6Gal II protein sequences, respectively, and among tetrapod ST6Gal I protein sequences (QVW-KDP) (61). Local alignments performed by ClustalX allowed refinement of the correspondences between the amino acid sequences (supplemental Fig. 3). These alignments revealed several insertion events, such as a poly(E) in T. nigroviridis, D. rerio, O. latipes, and T. rubripes ST6Gal II sequences and a poly(QLEREK) in the amphibian S. tropicalis ST6Gal I sequence of unknown biological relevance. Altogether, these observations suggest that the ancestral st6gal1 and st6gal2 genes have undergone small insertion/deletion (indel) events during vertebrate evolution that led to changes in the reading frame.

At the gene level, we pointed out previously overall gene organization conservation in five coding exons of the st6gal1 vertebrate genes with the notable exception of fish st6gal1 genes, which exhibit additional coding exons in their 5’ region (1, 2) or, alternatively, two additional intron sequences. The position of the teleost second intron is inside a relatively well conserved protein sequence, downstream to the amino acid corresponding to the human Trp$^{96}$. These results do not support the exon shuffling hypothesis.

We then tested if the indel events in st6gal genes could be linked to evolutionary change amounts in the catalytic regions. We took into account the events encompassing at least three codons retrieved in the sequences coding the stem region but absent in the sequences coding the catalytic domain. We also considered the two introns located in the region encoding the stem region of the teleost st6gal1 gene, which could be interpreted as insertions. Except in this last case, most indels could be considered as deletions compared with arthropod sequences. The largest deletion, denoted ID6 in Fig. 2A, concerns tetrapod ST6Gal I and comprises around 70 codons. A 17-codon-long deletion (ID5) characterizes vertebrate ST6Gal I sequences. A 15-codon-long insertion is only shared by T. rubripes and T. nigroviridis ST6Gal II sequences (supplemental Fig. 3). Three indels remain ambiguous and may correspond to insertions in the ancestor to vertebrate ST6Gal I and ST6Gal II or to deletions in arthropod ST6Gal I/II (Fig. 2A). Interestingly, most indel events are clearly hallmarks of ST6Gal subfamilies in different subsets of vertebrates (Fig. 2B). We tested if these indel events were linked to the length of corresponding branches in the phylogeny tree constructed from the comparison of the catalytic part of the protein. Because the branch

**TABLE 2**

|                | Sialylmotif L | Family motif b | Sialylmotif S |
|----------------|---------------|---------------|---------------|
|                | ST6Gal I | ST6Gal II | ST6Gal I | ST6Gal II | ST6Gal I | ST6Gal II | ST6Gal I | ST6Gal II |
| Amniotes/mammals | 5        | 1           | 4        | 1           | 2        | 1          |
| Tetrapodes/ammniotes | 4        | 0           | 2        | 2           | 3        | 1          |
| Osteichthyans/tetrapodes | 7        | 6           | 2        | 2           | 1        | 2          |
| Osteichthyans/teleosts | 4        | 5           | 9        | 1           | 4        | 0          |
| R2/osteichthyans | 4        | 4           | 1        | 1           | 4        | 3          |

**FIGURE 2.** Indel insertion/deletion events in the ST6Gal I and ST6Gal II protein sequences. A, alignments of protein sequences without gap are schematized in gray; indels are indicated by white boxes. Intron insertions are indicated by black bars in the teleost st6gal1 sequence. B, simplified phylogenetic tree and insertion/deletion distribution. The tree was constructed with 200 positions within the catalytic domain, using ME. Filled square, insertion; open squares, indel; filled rectangles, intron insertion; open triangles, deletion. C, correlation between Indel numbers and branch lengths. The tree was constructed as in B. The regression equation was given by PAST 2.01.
Evolutionary Study of \textit{st6gal} Gene Expression in Vertebrates

In order to investigate the dynamic of \textit{st6gal} gene evolution across vertebrate genomes and to explain the appearance of the two vertebrate \textit{st6gal} gene subfamilies, we first analyzed the evolutionary history of \textit{st6gal} in the context of the two rounds of whole genome duplications (WGD), also known as the 2R hypothesis (62). We assessed the paralogy and synteny relationships of the identified \textit{st6gal} genomic loci in various vertebrate genomes. The presence of two or more orthologous gene pairs on two distinct chromosomes in a single species can define paralogons issued from WGD events R1 and R2. In the human genome, using the Paralogon program (53), we found a statistically significant \((sm > 3)\) block limited to three genes (data not shown). We then studied a larger segment around both \textit{st6gal} genes, using Ensembl and found a set of 11 putative paralogous genes on HSA 3q27 and HSA 2q11.3 (Fig. 3) emphasizing the involvement of a genome doubling event. Taken together, these approaches support the hypothesis of WGD as a cause of \textit{st6gal} gene duplication in vertebrates.

Next, we examined the two \textit{st6gal} loci and their neighbors in the genome of various vertebrate species using the Synteny Database (63) (available on the World Wide Web). A conserved synteny refers to the existence of two or more orthologous genes that are co-localized on the same chromosome in two or more animal species, although their gene order on each chromosome can be different (64). The synteny including the \textit{st6gal2} gene is simple because the synteny data base site gave a set of 10 genes common to human HSA2q12 and zebrafish DRE9 (Fig. 4A). In the other examined teleost genomes (medaka and fugu), only one chromosome bears the \textit{st6gal12} synteny. In the \textit{S. tropicalis} genome, a series of four scaffolds corresponds to this synteny, suggesting their colinearity (supplemental Fig. 4). In addition, a paralogon of four genes, including the \textit{st6gal2} gene, was found in the zebrafish genome on DRE9 and DRE6 (Fig. 4A), suggesting a genome doubling event in teleosts (WGD R3).

For the synteny around the \textit{st6gal1} gene, the situation appears to be more complex because two different sets of genes can be defined in teleosts and in amnions (Fig. 4B), both well conserved within these two vertebrate groups. On one hand, HSA3 and GGA9 share 261 orthologous genes, among which 21 are present on \textit{S. tropicalis} scaffold 55 (supplemental Fig. 4). On the other hand, the fish chromosomes DRE21, GAC7, and TN17 share six genes (sclc6a/n, trpc2, ca4, pura, \textit{st6gal1}, rhodb), but only one gene, \textit{st6gal1}, is common to both groups of vertebrate genes (Fig. 4B). Further analysis performed in the synteny data base revealed seven genes shared by GGA9 and DRE21, including \textit{st6gal1}, 15 shared by GGA9 and DRE15, and 18 shared by GGA9 and DRE2 (supplemental Fig. 4). Five genes (rbp2, itm2c, clsn2, crbp2, and atp1b) have paralogs on DRE2 and DRE15, indicating that these segments result from a WGD R3 event that occurred at the base of teleost radiation, \(~350\) MYA (65–67). In summary, we can infer that in teleosts, a block of at least seven genes has been translocated to the equivalent chromosome of DRE21, from the protochromosome DRE15–2 of their common ancestor. Interestingly, there are two paralogous genes on DRE21 and DRE15 (neu2 and gpcr-rhod) that are absent from DRE2, suggesting that the seven-gene block has been translocated from the DRE2 ancestral chromosome, after the WGD R3 event (Fig. 4C). In addition, several genes around \textit{st6gal1}/2 in the \textit{B. floridiae} genome (scaf V2 104q) are retrieved around both the \textit{st6gal1} and \textit{st6gal2} genes in the human and chicken genomes (Fig. 4D), further suggesting conservation of synteny for \textit{st6gal} genes from cephalochordates to mammals and a disruption of \textit{st6gal1} synteny in teleosts.

**FIGURE 3.** Genomic organization of human \textit{st6gal1}/\textit{st6gal2} cluster paralgon and putative orthologous counterpart is indicative of a WGD event. Eleven putative paralogous genes spanning regions on human chromosomes HSA2q11.3 and HSA3q27 were found. Genes represented were chosen from analysis performed at Ensembl by chromosomal walking and reciprocal TBLASTN searches of genes adjacent to \textit{st6gal} loci.

**TABLE 3**

| ME   | NJ   | Site change numbers |
|------|------|---------------------|
| Pearson’s \(r\) | 0.885 | 0.833 | 0.785 |
| \(p\) | 0.15% | 0.53% | 1.21% |
**Evolutionary Study of st6gal Gene Expression in Vertebrates**

ST6Gal Gene Expression in Vertebrates and EST Analysis

Several studies have noted the differential expression pattern of α2,6-sialylation and ST6Gal genes in various mammalian species (37, 68–71). To estimate the breadth of st6gal gene expression, we looked at various tissue EST libraries from several representative animal species. We statistically analyzed the expression profiles of ESTs from the information retrieved on the Unigene site of NCBI. Tissue-dependent expression patterns were inferred from the EST profile accessible from the Unigene database. The multivariate approach of PCA gave a quite satisfactory result. The plan defined by the two first axes takes into account about 80% of information of the data set (Fig. 5A). The first axis of PCA expresses nearly 52% of variance, whereas the second axis represents more than 28% of variance. It appears that the projections of most gene expression profiles are gathered on the right side of the plane, whereas mammal and bird st6gal1 appears apart on the left side. This observation suggests that the expression profile of anomiote st6gal1 genes is almost ubiquitous, whereas teleost and amphibian st6gal1 genes have a more similar profile of expression compared with vertebrate st6gal2 genes. Furthermore, direction of the vectors corresponding to each tissue indicates preponderant expression of the pointed gene. As an example, the avian st6gal1 gene is more expressed in thymus, testis, or muscle compared with its mammalian counterpart, which is predominantly expressed in lung, kidney, or brain. The heat map (Fig. 5B) constructed using PAST 2.0 with log-transformed values illustrates the subequent expression of the st6gal1 gene in mammals and bird and indicates that testis, brain, kidney, and embryo tissues frequently express the st6gal2 gene.

RT-PCR in Adult Fish and Amphibian Tissues

To substantiate these observations and gain further insights into the expression of lower vertebrate st6gal genes, we designed oligonucleotides primers in the amphian S. tropica!

**FIGURE 4. Syntenic relationships of the st6gal gene loci in vertebrate and amphioxus genomes.** Physically mapped genomes of human (HSA), mouse (MMU), chicken (GGA), O. latipes (OLA), T. nigroviridis (TNI), G. aculeatus (GAC), and D. rerio (DRE) in Ensembl were used to identify conserved gene neighbors of the st6gal genes. A, conserved syntenic blocks found around the st6gal2 gene; B, disrupted synteny around the st6gal1 gene found in ammies and in teleost genomes. C, schematic diagram depicting the genetic events that occurred early in the jawed vertebrate lineage and led to modern st6gal genes in fish. The last common ancestor (LCA) of teleost fishes and ammies around 500 MYA bore an ancestral st6gal1/2 gene, which was located on the protocarchromosome DRE 15-2-6-9 HSA 2-3; WGD R2 (∼475 MYA) gave rise to st6gal1 and st6gal2 genes found on HSA3 and HSA2, respectively, and on the protocarchromosome DRE 15-2 and DRE 6-9 of the last common ancestor. The DRE 2/DRE 15 and DRE 6/DRE 9 pairs of chromosomes derived from these single common protocarchosomes DRE 15-2 and DRE 6-9 after WGD R3. Rearrangement of a block of genes including st6gal1 occurred in teleosts, resulting in the translocation of st6gal1 on DRE21, whereas the other duplicate, named st6gal1-r, disappeared from DRE 15 through pseudogenization or deletion. st6gal2 and st6gal2-r were maintained on DRE 6 and DRE 9. D, conserved synteny observed between the amphioxus genomic region hosting the st6gal-like gene (scaf V2 104q) and the two sets of human paralogons described in the legend to Fig. 3.
Evolutionary Study of st6gal Gene Expression in Vertebrates

DISCUSSION

Because many biological processes are governed by carbohydrate-protein interactions involving sialic acids, the evolutionary approach to gain further insights into the biological relevance of sialyltransferases is of particular interest (72, 73). The β-galactoside α2,6-sialyltransferases ST6Gal I and ST6Gal II mainly described in mammals mediate the addition of α2,6-linked sialic acid to Galβ1–4GlcNAc and GalNAcβ1–4GlcNAc disaccharides, respectively (8). Our phylogenetic and gene expression studies provide insights into the regulation and function of these conserved genes as well as important clues to the evolutionary events and functional changes that have occurred in different animal species. To date, such results on phylogenetic relationships and expression patterns of a glycosyltransferase family are quite unique (1, 37, 55).

st6gal Gene Sequences Appeared Early in Metazoans—The mRNA fragment identified from O. carmella, a sponge with chemical conduction, epithelial-like cells, and sensory-like cells from the porifera phylum (74), suggests that an ancestral st6gal1/2 gene was already present in the earliest metazoans. This gene could be orthologous to the one present in the sponges Geodia cydonium, in which Muller et al. (75) and we noticed an increased expression during late stages of development. The highest level of expression was found in the brain and in non-neuronal territories, such as the proctodeum, gall bladder, and intestinal bulb. st6gal2 is expressed in the marginal zone of the CNS, stronger in anterior diencephalon and in lateral anterior hindbrain, and in the ganglion cell layer of retina, except in the proliferative zone.

Regulatory Evolution of st6gal1 Genes; 5′-RLM-RACE

The transcriptional start site(s) (TSS) and complete 5′-untranslated region (5′-UTR) were determined by 5′-RLM-RACE in lower vertebrate st6gal1 genes of the zebrafish D. rerio and the frog S. tropicalis using total RNA extracted from zebrafish eggs and intestine tissues or frog liver and intestine tissues, respectively (data not shown). Unique 5′-RACE amplification products of about 160 bp in zebrafish tissues and of about 1060 bp in frog tissues were obtained and subcloned in TOPO TA pCRII vector, and several clones were fully sequenced. The results demonstrated the existence of a unique TSS for zebrafish and frog st6gal1 genes in these tissues. Comparison of these cDNA sequences with genomic databases indicated that these unique zebrafish and frog transcripts show either one or two additional 5′-UT exons, respectively (Fig. 8), located far upstream the first coding exon. In contrast to the higher vertebrates, where a complex 5′-UTR with multiple upstream non-coding exons and multiple start sites has been described for the st6gal1 genes, there is a unique st6gal1 transcript in lower vertebrates showing a simple 5′-UTR with one or two non-coding exons.

A.

FIGURE 6. Expression pattern of the zebrafish and amphibian st6gal genes in various adult tissues using RT-PCR. A, fish st6gal genes. Relative expression levels of zebrafish st6gal and β-actin mRNA were evaluated by RT-PCR as described under “Experimental Procedures,” among various adult tissues. Oligonucleotide primer sequences specific to fish st6gal1, st6gal2, and st6gal2-r are given in supplemental Fig. 5 and yield 236- and 465-bp PCR fragments, respectively. Lane 1, kidney; lane 2, intestine; lane 3, brain; lane 4, liver; lane 5, muscle; lane 6, heart; lane 7, eggs; lane 8, oocytes. The zebrafish β-actin (378 bp) was amplified as a control of CDNA synthesis and purity. B, amphibian st6gal genes. Oligonucleotides specific to S. tropicalis st6gal1 and st6gal2 are indicated in supplemental Fig. 5 and yield 236- and 465-bp PCR fragments, respectively. Lane 1, muscle; lane 2, eyes; lane 3, heart; lane 4, skin; lane 5, lung; lane 6, stomach; lane 7, ovaries; lane 8, intestine; lane 9, liver; lane 10, brain; lane 11, kidney. The amphibian β-catenin (327 bp) was amplified for 35 cycles as a control of CDNA synthesis and purity.

B.

restricted to intestine, kidney, and ovaries. It is also expressed in liver at a low level in fish and to a larger extent in frogs. Conversely, the st6gal1 gene is largely expressed in adult fish brain, whereas it is almost not detected in frog brain. Altogether, both st6gal genes have a similar expression profile, and they are not detected in muscle and heart. The amphibian and zebrafish st6gal2 gene expression is maintained in adult brain, ovaries, and intestine with overlapping territories of expression for st6gal1.

In Situ Hybridization (ISH) during Zebrafish Development

To examine the expression pattern among st6gal paralogs during zebrafish embryonic development, we performed whole-mount RNA ISH with zebrafish embryos (Fig. 7). st6gal2 and st6gal2-r gene expression was detected from gastrulation until larva stage (5 days postfertilization), whereas st6gal1 gene expression was not detected before 24 h postfertilization (hpf) or after hatching (48 hpf). Our ISH analysis indicated that at embryonic developmental stage 48 hpf, st6gal1 and st6gal2 genes are expressed in overlapping brain territories of zebrafish. We found a continuous expression of the two st6gal2-related genes during development, from egg to larva stages. Both genes are detected in hatching gland cells. As for the st6gal2-r gene, the overall level of expression is rather low,

and we noticed an increased expression during late stages of development. The highest level of expression was found in the brain and in non-neuronal territories, such as the proctodeum, gall bladder, and intestinal bulb. st6gal2 is expressed in the marginal zone of the CNS, stronger in anterior diencephalon and in lateral anterior hindbrain, and in the ganglion cell layer of retina, except in the proliferative zone.

Regulatory Evolution of st6gal1 Genes; 5′-RLM-RACE

The transcriptional start site(s) (TSS) and complete 5′-untranslated region (5′-UTR) were determined by 5′-RLM-RACE in lower vertebrate st6gal1 genes of the zebrafish D. rerio and the frog S. tropicalis using total RNA extracted from zebrafish eggs and intestine tissues or frog liver and intestine tissues, respectively (data not shown). Unique 5′-RACE amplification products of about 160 bp in zebrafish tissues and of about 1060 bp in frog tissues were obtained and subcloned in TOPO TA pCRII vector, and several clones were fully sequenced. The results demonstrated the existence of a unique TSS for zebrafish and frog st6gal1 genes in these tissues. Comparison of these cDNA sequences with genomic databases indicated that these unique zebrafish and frog transcripts show either one or two additional 5′-UT exons, respectively (Fig. 8), located far upstream the first coding exon. In contrast to the higher vertebrates, where a complex 5′-UTR with multiple upstream non-coding exons and multiple start sites has been described for the st6gal1 genes, there is a unique st6gal1 transcript in lower vertebrates showing a simple 5′-UTR with one or two non-coding exons.

FIGURE 6. Expression pattern of the zebrafish and amphibian st6gal genes in various adult tissues using RT-PCR. A, fish st6gal genes. Relative expression levels of zebrafish st6gal and β-actin mRNA were evaluated by RT-PCR as described under “Experimental Procedures,” among various adult tissues. Oligonucleotide primer sequences specific to fish st6gal1, st6gal2, and st6gal2-r are given in supplemental Fig. 5 and yield 236- and 465-bp PCR fragments, respectively. Lane 1, kidney; lane 2, intestine; lane 3, brain; lane 4, liver; lane 5, muscle; lane 6, heart; lane 7, eggs; lane 8, oocytes. The zebrafish β-actin (378 bp) was amplified as a control of CDNA synthesis and purity. B, amphibian st6gal genes. Oligonucleotides specific to S. tropicalis st6gal1 and st6gal2 are indicated in supplemental Fig. 5 and yield 236- and 465-bp PCR fragments, respectively. Lane 1, muscle; lane 2, eyes; lane 3, heart; lane 4, skin; lane 5, lung; lane 6, stomach; lane 7, ovaries; lane 8, intestine; lane 9, liver; lane 10, brain; lane 11, kidney. The amphibian β-catenin (327 bp) was amplified for 35 cycles as a control of CDNA synthesis and purity.

B.

restricted to intestine, kidney, and ovaries. It is also expressed in liver at a low level in fish and to a larger extent in frogs. Conversely, the st6gal1 gene is largely expressed in adult fish brain, whereas it is almost not detected in frog brain. Altogether, both st6gal genes have a similar expression profile, and they are notably not detected in muscle and heart. The amphibian and zebrafish st6gal2 gene expression is maintained in adult brain, ovaries, and intestine with overlapping territories of expression for st6gal1.

In Situ Hybridization (ISH) during Zebrafish Development

To examine the expression pattern among st6gal paralogs during zebrafish embryonic development, we performed whole-mount RNA ISH with zebrafish embryos (Fig. 7). st6gal2 and st6gal2-r gene expression was detected from gastrulation until larva stage (5 days postfertilization), whereas st6gal1 gene expression was not detected before 24 h postfertilization (hpf) or after hatching (48 hpf). Our ISH analysis indicated that at embryonic developmental stage 48 hpf, st6gal1 and st6gal2 genes are expressed in overlapping brain territories of zebrafish. We found a continuous expression of the two st6gal2-related genes during development, from egg to larva stages. Both genes are detected in hatching gland cells. As for the st6gal2-r gene, the overall level of expression is rather low,
detected a sialyltransferase activity at the cell surface involved in cell-cell recognition. Although the relationships between all of the sialyltransferase families are not yet established, the st6gal gene family could constitute the most ancient sialyltransferase family described in animals (2). Because this st6gal1/2 gene is retrieved from most studied arthropod and deuterostome genomes, we can deduce that it has disappeared independently in several lineages, as in the cnidarians, the lophotrochozoa (mollusks and annelids), the hymenoptera insects A. mellifera and N. vitripennis, nematodes such as C. elegans (76), the sea urchin S. purpuratus, and the tunicates C. intestinalis and C. savignyi. The reason for st6gal1/2 gene loss in these taxa must be related to the primary function of this gene product. Given the small number of invertebrate genomes explored so far, the information available in protostomes and deuterostomes is quite fragmentary and has been mainly documented in Drosophila and vertebrates. Sialylation in insects has long been controversial (32, 35, 77), and recently, DSiaT, a unique st6gal gene, has been characterized in Drosophila (31). It is exclusively expressed in a subset of neurons in late embryonic stage 17, in the optic lobe of third instar larva and in the region of olfactory projection neurons in adult head (35). The encoded enzyme was found to be involved in the function of a voltage-gated sodium channel and neuromuscular junction and appears to be

![Figure 7. Expression pattern of the zebrafish st6gal genes during zebrafish embryonic development using whole mount ISH at early developmental stages (A) and later developmental stages (B). The st6gal1 gene shows no labeling at early developmental stages (gastrula (G), early somitogenesis (ES), and middle somitogenesis (MS)). st6gal1 is found in the ventricular part of telencephalon, in the ventral part of diencephalon, in the lateral part of the hindbrain (discontinuous patches of cells), and in the lateral part of the spinal cord at 24 and 36 hpf. st6gal1 gene expression is detected in the ganglion cell layer of the retina (except in the proliferative zone), cranial ganglia, and brain marginal zone at 48 hpf and is not detected later. The st6gal2 gene is expressed at the gastrula stage in the prechordal plate mesendoderm and at early somitogenesis in polster, anterior border of neural plate, adaxial cells, formed somites, and also weakly in segmental plates. At middle somitogenesis, it is expressed in polster, anterior head epidermis, caudal adaxial cells, and the last formed somites. At 24 hpf, the st6gal2 gene is expressed in tail somites and adaxial cells and weakly in the CNS; later, at 36 hpf, a stronger signal is found in the marginal zone, and no labeling is detected in tectum. At 48 hpf and 5 days postfertilization, the st6gal2 transcript is detected in anterior diencephalon and in the ganglion cell layer of retina but not in the proliferative zone. The st6gal2-r gene is weakly expressed at the gastrula stage and shows also a weak basal level of expression at early somitogenesis and middle somitogenesis and more labeling in hatching glands. At later developmental stages, from pharyngula to larva stages (24 hpf to 5 days postfertilization (dpf)), the st6gal2-r gene is detected in hatching glands, mucus cells, proctodeum, intestinal bulb, gall bladder, inner nuclear cell layer of retina (except in the proliferative zone), nucleus in diencephalon, nucleus in anterior tectum, and nucleus in hindbrain at the larval stage. hg, hatching gland; inl of retina, inner nuclear layer of retina; i bulb, intestinal bulb; n in diencephalon, nucleus in diencephalon; n in ant tectum, nucleus in anterior tectum; n in hb, nucleus in hindbrain; up mesoderm, unsegmented paraxial mesoderm; tg, tegmentum; ventral dc, ventral diencephalon. LV, lateral view; DV, dorsal view.

| Time   | st6gal1 | st6gal2 | st6gal2-r |
|--------|---------|---------|-----------|
| 24 hpf | LV      | LV      | LV        |
| 36 hpf | LV      | LV      | LV        |
| 48 hpf | LV      | LV      | LV        |
| 5 dpf  | LV      | LV      | LV        |
essential for the regulation of nervous system function (36). Moreover, this Drosophila protein exhibits notable preferred enzymatic activity toward LacdiNAc substrates over LacNAc termini in in vitro assays (31), despite the fact that no evidence for the presence of LacdiNAc or LacNAc could be established in vivo (34). Mammalian st6gal1 and st6gal2 genes described previously have counterparts in all vertebrates examined, except for the lampreys, the living representatives of jawless vertebrates (agnatha), in which the three st6gal gene sequences form a sister group to all other vertebrate sequences. Two of
these three \textit{st6gal} genes were amplified by PCR (supplemental Fig. 5) from a 6–10-day embryonic cDNA library kindly provided by Prof. J. Langeland (78), indicating their expression during embryogenesis, whereas the third one appears to be absent (data not shown). Our phylogenetic analysis indicates that the single \textit{st6gal} gene found in arthropods, the two copies found in amphioxus, and the three copies found in lampreys are orthologous to all vertebrate \textit{st6gal} genes.

In order to explain the origin of \textit{st6gal1} and \textit{st6gal2} gene duplication in vertebrates, we compared the environment of each identified \textit{st6gal} gene locus and determined their paralogy or orthology relationships. We pointed out a disruption in the conserved synteny of \textit{st6gal1} loci in teleost fishes further suggesting a chromosomal rearrangement. These translocation events are known to occur at higher rates in fish genomes compared with tetrapod genomes (79). On the other hand, \textit{st6gal2} synteny was maintained during vertebrate evolution. Moreover, intraspecific comparisons of chromosome segments inside vertebrates revealed that blocks of paralogous genes, named paralogons, can be identified (EPGD (80), CHSMiner (81)). Large sets of paralogons have been interpreted as a result of two rounds of genome duplications that occurred early in vertebrate evolution. The first round R1 probably occurred around 550 MYA, before the separation of lampreys from jawed vertebrates (gnathostomata). The second round R2 dates to about 474 MYA, after the emergence of lamprey and before cartilaginous fishes (chondrichthyan) divergence (82, 83). Identification of paralogons in the vertebrate genome and our calculations indicate that the \textit{st6gal1} and \textit{st6gal2} split dates back to this period and lead us to assume that one of the \textit{st6gal} genes duplicated from R1 was lost. Subsequently, a third WGD R3 occurred ~350 MYA in the ray-finned fish lineage, after emergence of lobe-finned fishes (65, 79, 84–86), leading to the paralogon pair including \textit{st6gal2} and \textit{st6gal2-r} genes found in the zebrafish genome. The \textit{st6gal2-r} gene was maintained in zebrafish but lost over time in other fish lineages, probably due to functional redundancy because both genes show similar patterns of expression during development.

A Scenario of Tissue Expression Evolution in Vertebrates—Our EST analysis using PCA highlighted another differential profile of expression of lower vertebrate \textit{st6gal1} genes compared with higher vertebrates. \textit{st6gal1} genes from fishes and amphibians form a cluster with all of the vertebrate \textit{st6gal2} genes, whereas mammalian and avian \textit{st6gal1} genes are found apart. This suggests an evolutionary change of the expression profile of \textit{st6gal1} gene in amniones. Using ISH in embryonic zebrafish tissues, we found overlapping territories of expression of \textit{st6gal1} and \textit{st6gal2} genes maintained in the adult brain in several vertebrate species (1). Surprisingly, \textit{st6gal2} and \textit{st6gal2-r} genes exhibit differential patterns of expression. Both genes are expressed at early developmental stages, and the gastrula stage marks their onset of expression. The \textit{st6gal2-r} gene is primarily detected in hatching gland cells, which produce metalloprotease choriolytic enzymes HCE and LCE digesting egg envelope (chorion) at the time of embryo hatching (87). This suggests a role in the process of hatching gland differentiation (same time as differentiation of notochord and paraxial mesoderm), in mucous cells and proctodeum.

We next analyzed adult tissue distribution of \textit{st6gal} genes using RT-PCR in lower vertebrates. We observed that adult \textit{D. rerio} and \textit{S. tropicalis} express the \textit{st6gal2} gene mainly in the brain, as previously reported for the mammalian \textit{st6gal2} gene (5, 9). It is also highly expressed in ovaries and to a lesser extent in intestine. Such slight variations in the \textit{st6gal2} expression profile have been reported for the bovine gene, which is significantly amplified from lung and intestine adult tissues (88). Both organisms also express the \textit{st6gal1} gene in ovaries and intestine, but their expression profile is more heterogeneous in lower vertebrates, which is in sharp contrast to the ubiquitous mammalian \textit{st6gal1} gene expression profile. Interestingly, in \textit{D. rerio}, \textit{st6gal1} is found in kidney and is notably absent in liver, whereas in \textit{S. tropicalis}, it is amplified in liver tissue and is almost not detected in kidney. Analysis of the EST profile of the chicken (Gga.1148) provided by the GenBank\textsuperscript{TM} data base illustrates expression of the \textit{st6gal1} gene in several adult tissues, such as brain, liver, thymus, muscle, ovary, or bursa of Fabricius. Because it is expressed in the zebrafish kidney, the frog liver and the bird bursa of Fabricius, which are the chief organs of B-cell development corresponding to the mammalian bone marrow (89), we hypothesize that the \textit{st6gal1} gene product would have gained a progressive function in lymphoid organs during evolution. Indeed, genetically \textit{st6gal1}-altered mice provided fragmentary insights into \textit{st6gal1} biological function, showing that the enzyme is implicated in immune system function (16, 90).

We could also predict that this gene is expressed in the thymus of teleosts and amphibians.

In summary, we observe a relative conservation of the \textit{st6gal2} expression profile in vertebrates, suggesting that it could be involved in molecular mechanisms that support neurogenesis (91) and thus would have conserved this role in the CNS already recorded in \textit{Drosophila}. However, its expression is also maintained in ovaries and intestine in lower vertebrates and mammals, further suggesting that the \textit{st6gal2} gene might have evolved new functions acquired within the 75 million years that elapsed between the R2 and the osteichthyans radiation because during that period, the \textit{st6gal2} gene evolved more rapidly than the \textit{st6gal1} gene, as illustrated by its longer branch lengths (Fig. 1). Up to now, its physiological function in vertebrates remains unknown, although it has been shown to be implicated in apoptosis (92).

Analysis of 5′-End of \textit{st6gal} Genes—In order to better understand the pattern of \textit{st6gal} gene expression diversification in vertebrates, we assessed which factors might have influenced their expression at the genomic level. BLAST searches of zebrafish and amphibian EST resources of the NCBI data base tend to demonstrate conservation over vertebrate evolution of the number of TSS for the \textit{st6gal2} gene already described in mammals (88, 93), correlating with their conserved pattern of expression (data not shown). We thus focused on the \textit{st6gal1} gene, which is ubiquitously detected in mammalian adult tissues, strongly expressed by the human liver, and transiently up-regulated during inflammation and in several cancers, due to multiple promoter–driven 5′-UT exons (1, 94). Our 5′-RACE analysis of fish and frog \textit{st6gal1} exons in several adult tissues clearly demonstrated the use of a unique TSS and the presence of one 5′-UT exon in \textit{D. rerio} and two 5′-UT exons in \textit{S. tropi-
Evolutionary Study of st6gal Gene Expression in Vertebrates

calis tissues. Our data further suggest that major changes have occurred at the level of regulatory cis-acting sequences and point to a still hypothetical rapid evolution of their regulatory genomic sequences that might be due to greater relaxation of evolutionary constraints often considered to be the driving force in the evolution of genetic networks (95). This rapid complication of the genetic/epigenetic regulation of expression of the st6gal1 gene has led to a diversification of the tissue distribution and also of function in higher vertebrates. Indeed, phenotypic variation in α2,6-sialylation of N-glycosylproteins has been observed in various animals and in particular in mammals despite genetic conservation of their translated gene sequences (24). The patterns of tissue α2,6-sialylation of N-glycosylproteins differ widely among mammals, even among closely related taxa, such as mice and humans, which diverged only 96 MYA (68, 96–98), or great apes and humans, which diverged 13–14 MYA (28, 96). We suggest a still on-going evolution and neo-functionalization of st6gal1 genes in mammals, which could explain differences in influenza virus infection of airway epithelial cells (24).

In the context of the rapid evolution of functions of st6gal genes in vertebrates, our data further suggest that the st6gal2 genes might have maintained an ancestral function due to their localized expression in vertebrate CNS and similar biochemical activity compared with DSiaT (31). Mammalian recombinant ST6Gal I and ST6Gal II enzymes produced in heterologous systems like Spodoptera frugiperda (SF-9) mediate the addition of α2,6-linked sialic acid to Galβ1–4GlcNAc (LacNac) and to GalNAcβ1–4GlcNAc (LacDiNac) disaccharides, respectively (7, 8). It has been previously shown that the ST6Gal I/II enzyme from D. melanogaster prefers LacDiNac-bearing substrates over LacNac (31), an enzymatic characteristic that was maintained in mammalian ST6Gal II enzymes (8, 10). Although information on these enzymes is lacking for lower vertebrates, we postulate that these biochemical properties extend to all of the vertebrate ST6Gal enzymes. Interestingly, the distribution of LacDiNac in mammals is very limited, and LacDiNac might be substituted by 4-O-sulfated-, α1,3-fucosylated, or α2,6-sialylated derivatives (99). As indicated by these authors, the glycans bearing LacDiNac are notably recorded in pituitary glycoprotein hormones and tenascin-R produced by oligodendrocytes and small interneurons in the hippocampus and cerebellum. Other glycoproteins concerned are glycodelin, with potent immunosuppressive and contraceptive activities in humans, and zona pellucida glycoproteins from murine eggs. We suggest that the ancestral ST6Gal I/II accept GalNac substrates better than Gal substrates and that the new properties of amniote ST6Gal I toward Gal substrates may help to evade pathogens, as suggested previously (98).

Variations in the ST6Gal Sequences during Evolution—Comparison of cumulative numbers of amino acid substitutions in the catalytic part of the ST6Gal enzymes between corresponding branches in the ST6Gal I and ST6Gal II trees raised intriguing points that deserve discussion. Within the ST6Gal II tree, we noticed short branches between each species, supporting our hypothesis of a conserved role throughout animal evolution, due to selective pressure. However, the higher number of substitutions at the base of the ST6Gal II clade compared with the base of the ST6Gal I clade is indicative of changes difficult to interpret because no significant changes within the conserved motifs are recorded (see the R2/osteichthyan line in Table 2). In contrast, there is a long branch leading to ST6Gal I teleosts from osteichthyan ancestors, associated with an accumulation of substitution within the family motif b and sialylmotif S (1). This feature suggests an original function of this enzyme within bony fishes, although this hypothesis requires further study. In the tetrapod clade, we observed a greater number of substitutions in the branch leading to amniote ST6Gal I compared with the amniote ST6Gal II branch, associated with changes in the sialylmotif L. More interestingly, this accumulation is more clear in the branch leading to mammals, and this affects the sialylmotif L and family motif b. We can deduce that there is a progressive change in the function of ST6Gal I from tetrapods to mammals and that these changes are probably not of the same nature as those observed in teleosts.

As mentioned previously, there is a shift in the preferred specificity of acceptor substrate during the evolution of ST6Gal I in vertebrates (8). It is interesting to note that site-directed mutagenesis of the sialylmotif L and S conserved amino acids in the rat ST6Gal I and in the human ST3Gal I indicated that they are implicated in the donor (CMP-Neu5Ac) and acceptor binding, respectively (100–102). Thus, the present analysis of substitution changes suggests that the LacdiNac to LacNac shift would be correlative to the accumulation of substitutions during amniote to mammal evolution.

The stem region of vertebrate ST6Gal enzymes and their coding sequences also show important variations, and we observed more indel events in the branches leading to teleosts and to amniote ST6Gal I than in the ST6Gal II counterparts. The fish st6gal1 genes exhibit two additional intron sequences that probably result from successive insertions of two spliceosomal introns in the first exon of the fish st6gal1 gene, after teleost radiation. Moreover, we noticed an insertion of repeated genetic sequences in the amphibian st6gal1 gene, leading to the formation of an acidic supercoiled region in the ST6Gal I stem region that may have variable impacts on the subcellular distribution in the trans-Golgi network and enzymatic activity of ST6Gal I. We molecularly cloned the amphibian st6gal1 cDNA sequence and confirmed the presence of acidic repetitive sequences (REKDEL) in the S. tropicalis ST6Gal I protein sequence, which is also found in the bifunctional α2,3/α2,8-sialyltransferase of Helicobacter acinonychis Sheeba (YP_665016). The signals and mechanisms mediating Golgi localization have been studied extensively for various mammalian glycosyltransferases (reviewed in Refs. 103–106). The cystolic tail/transmembrane domain/stem region of human ST6Gal I is probably implicated in subcellular traffic through functional homodimerization and/or interactions with other proteins, such as COP-I coated vesicle anterograde traffic of GT or COG (103, 107–111) or Golgi retention (112–114) and in the modulation of its enzymatic activity through substrate recognition (61). It has also been shown that the cystolic tail/transmembrane domain/stem region impacts mammalian ST6Gal I secretion via BACE-1 aminopeptidase activity in Alzheimer disease (115–118). Altogether, our data on the molecular cloning of lower vertebrate st6gal1 genes and their molecular evo-
Evolutionary Study of st6gal Gene Expression in Vertebrates

...tion raise the question of the evolution of vertebrate ST6Gal I with regard to their subcellular localization, interaction with other glycosyltransferases, and activity in vivo.

The traditional paradigm is that duplication releases a selective constraint on one paralogous gene, offering the possibility of the appearance of new function(s). Here, we show on-going neofunctionalization of st6gal1 genes in amniotes and maybe in teleosts. The consequence of neofunctionalization of st6gal1 genes is a net increase in expression complexity following duplication. Those genes, such as st6gal1, implicated in immunity, host defense, reproduction, and olfaction are fractionally evolving, whereas those, such as st6gal2, implicated in intracellular signaling, neurogenesis, and neurophysiology are slowly evolving (119). Alternatively, another model of evolution of duplicated genes named subfunctionalization suggests duplication-degeneration-complementation (the DDC model), leading to pleiotropic expression (120). This appears to be more or less the case with st6gal2 and st6gal2-r genes in zebrafish, issued from a specific teleost WGD R3, which has a complementary pattern of expression in adult tissues and embryos. Overlapping expression domains could produce fine graining of gene function (85).

A relationship has been established between the breadth of expression, expressed as the number of tissues in which ESTs are recorded, and evolutionary rates. Briefly, the wider the tissue expression, the weaker the evolutionary rate, a fact attributable to a greater selective pressure when a gene is expressed over a variety of tissues (121, 122). In the case of st6gal genes, we observe quite the opposite because st6gal2 shows low variations of sequence despite the reduced number of tissues in which it is expressed. Thus, the conservation of the function of this gene would be driven by purifying selection. In contrast, the high evolutionary rates observed in st6gal1 gene sequences instead result from changes of function and of specificity and an increase in expression breadth, through an increase in the number of TSSs. ST6Gal I progressively acquired functions in the immune system from a probable ancestral role in embryonic and adult CNS.

Acknowledgments—We acknowledge Prof. Mohammed Lemdani (University of Lille Nord de France, Lille 2, France) and Dr. François Foulquier (University of Lille Nord de France, Lille 1, France) for helpful discussions; Marianne Gérard, Amandine Verlande, and Leïla Bekri for excellent technical assistance; and Dr. Rosella Molli-...
Evolutionary Study of st6gal Gene Expression in Vertebrates

Citron, M., Shitara, H., Taya, C., Yonekawa, H., Paulson, J. C., Miyoshi, E., Taniguchi, N., and Hashimoto, Y. (2005) J. Biol. Chem. 280, 8589–8595

116. Kitazume, S., Oka, R., Ogawa, K., Futakawa, S., Hagiwara, Y., Takikawa, H., Kato, M., Kasahara, A., Miyoshi, E., Taniguchi, N., and Hashimoto, Y. (2009) Glycobiology 19, 479–487

117. Kitazume, S., Saito, T. C., and Hashimoto, Y. (2002) Seikagaku 74, 1180–1183

118. Kitazume, S., Tachida, Y., Oka, R., Kotani, N., Ogawa, K., Suzuki, M., Dohmae, N., Takio, K., Saito, T. C., and Hashimoto, Y. (2003) J. Biol. Chem. 278, 14865–14871

119. Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., Yamada, T., Nagayasu, Y., Doi, K., Kasai, Y., Jindo, T., Kobayashi, D., Shimada, A., Toyoda, A., Kuroki, Y., Fujiyama, A., Sasaki, T., Shimizu, A., Asakawa, S., Shimizu, N., Hashimoto, S., Yang, J., Lee, Y., Matsushima, K., Sugano, S., Sakaizumi, M., Narita, T., Ohishi, K., Haga, S., Ohta, F., Nomoto, H., Nogata, K., Morishita, T., Endo, T., Shin-I, T., Takeda, H., Morishita, S., and Kohara, Y. (2007) Nature 447, 714–719

120. Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., and Postlewait, J. (1999) Genetics 151, 1531–1545

121. Duret, L., and Mouchiroud, D. (2000) Mol. Biol. Evol. 17, 68–74

122. Subramanian, S., and Kumar, S. (2004) Genetics 168, 373–381

123. Tsuji, S., Datta, A. K., and Paulson, J. C. (1996) Glycobiology 6, v–vii