Phylogenetic Analysis of the Vertebrate Galectin Family

Denis Houzelstein, Isabelle R. Gonçalves,† Andrew J. Fadden,‡ Sukhvinder S. Sidhu,*[1] Douglas N. W. Cooper,§ Kurt Drickamer,‡ Hakon Leffler,|| and Françoise Poirier**

*Laboratoire de Génétique et Développement des Mammifères and †Laboratoire de Structure et dynamique des Génomes, Institut Jacques Monod, Paris, France; ‡Department of Biochemistry, Glycobiology Institute, University of Oxford, Oxford, United Kingdom; §Departments of Psychiatry and Anatomy, University of California at San Francisco; and ||Section MIG (Microbiology Immunology Glycobiology), Inst Laboratory Medicine, Lund University, Lund, Sweden

Galectins form a family of structurally related carbohydrate binding proteins (lectins) that have been identified in a large variety of metazoan phyla. They are involved in many biological processes such as morphogenesis, control of cell death, immunological response, and cancer. To elucidate the evolutionary history of galectins and galectin-like proteins in chordates, we have exploited three independent lines of evidence: (i) location of galectin encoding genes (*LGALS*) in the human genome; (ii) exon-intron organization of galectin encoding genes; and (iii) sequence comparison of carbohydrate recognition domains (CRDs) of chordate galectins. Our results suggest that a duplication of a mono-CRD galectin gene gave rise to an original bi-CRD galectin gene, before or early in chordate evolution. The N-terminal and C-terminal CRDs of this original galectin subsequently diverged into two different subtypes, defined by exon-intron structure (F4-CRD and F3-CRD). We show that all vertebrate mono-CRD galectins known to date belong to either the F3- or F4-subtype. A sequence of duplication and divergence events of the different galectins in chordates is proposed.

Introduction

Complex carbohydrates are prominent constituents of all living organisms and have been implicated in a variety of functions including cell-cell recognition, signaling, and host-microbe interactions. These functions often involve specific recognition of saccharide motifs by families of proteins called lectins (Gabius 2002; Kilpatrick 2002 for recent reviews). One such protein family, the galectins, is defined by both shared sequence elements and affinity for β-galactosides. Members of this family have been isolated from phyla ranging from vertebrates to sponges; about fourteen mammalian members have been reported so far (Cooper 2002).

X-ray crystallography of several galectins has shown that their carbohydrate recognition domains (CRD) consist of about 130 amino acids arranged in a tightly folded conserved β-sandwich structure formed by a six strand sheet (S1–S6) and a five strand sheet (F1–F5), with the conserved carbohydrate-binding amino acids in strands S4–S6 (Rini and Lobsanov 1999; Loris 2002). In the mammalian galectin genes published so far, the CRDs are encoded by three consecutive exons (Gitt and Barondes 1991; Barondes et al. 1994*b*; Gitt et al. 1998*b*). The middle exon encodes almost all the conserved residues that make up the galectin signature and are known to interact with bound carbohydrates (Lobsanov et al. 1993; Barondes et al. 1994*b*).

Galectins are usually grouped based on their architecture (Hirabayashi and Kasai 1993). The prototype of this family, galectin-1, is a noncovalent dimer of a subunit essentially consisting of a CRD. Many other mono-CRD galectins have subsequently been discovered, which may or may not form dimers (for example mammalian galectin-5, -7, -10 [Kopitz et al. 2003]). The second class of galectins only encompasses Galectin-3, which contains a single CRD, fused to an N-terminal unrelated sequence of about 120 amino acids. Galectin-3 is referred to as a chimeric galectin. A third class of galectins contains two CRDs within the same peptide chain joined by a linker of up to 70 amino acids in between. These bi-CRD galectins are named tandem-repeat type galectins.

The cellular properties of galectins are unusual in that they are synthesized as cytosolic proteins but can be stimulated to secretion by nonclassical pathways or alternatively targeted to the nucleus (Hughes 2001; Leffler 2001). Extracellular galectins may modulate cell adhesion and induce intracellular signals by cross-linking cell-surface and extracellular glycoproteins, possibly forming supramolecular ordered arrays (Sacchettini, Baum, and Brewer 2001). Galectins may also bind intracellular noncarbohydrate ligands and have intracellular regulatory roles in processes such as RNA splicing, apoptosis, and cell cycle (Liu, Patterson, and Wang 2002). Thus, a huge body of evidence points to multifunctional effects for galectins (Leffler 2001; Gabius et al. 2002).

Among the best-documented potential functions for a galectin are the proinflammatory role of galectin-3 and the immunosuppressive role of galectin-1 (Leffler 2001 for review). Galectins have also been implicated in cancer. For instance, galectin-3 can enhance tumor growth by its anti-apoptotic and angiogenic functions and promote metastasis by affecting cell adhesion (Andre et al. 1999; Kim et al. 1999; Nangia-Makker et al. 2000; Honjo et al. 2001).

Despite all these activities in vitro, galectin-1 and galectin-3 null mutant mice are viable and fertile and display very subtle phenotypes (Poirier and Robertson 1993; Colnot et al. 1998); thus the basic biological functions of galectins remain largely unclear (see Danguy, Camby, and Kiss 2002; Liu, Patterson, and Wang 2002; Rabinovich, Rubinstein, and Toscano 2002 for recent reviews). This raises a number of questions about the functional and evolutionary relationships between galectins. Can members of this family that display overlapping expression profiles compensate for the loss of another galectin? Can different galectins have agonistic or antagonistic actions within

[1] Present address: Department of Anatomy, University of California at San Francisco.

Key words: galectin, Lgals, vertebrate, chordate, evolution.

E-mail: houzelstein@ijm.jussieu.fr.

Mol. Biol. Evol. 21(7):1177–1187. 2004
doi:10.1093/molbev/msh082
Advance Access publication February 12, 2004

a given cell or tissue? To what extent do galectins in one species have orthologues in other species and if so, is it possible to extrapolate functional information? To address such questions, we have examined the history of the galectin encoding genes (named *LGALS* in human, *Lgals* in other chordates) by exploiting three independent lines of evidence: (i) location of *LGALS* genes in the human genome; (ii) *Lgals* gene exon-intron organization; and (iii) sequence comparison of available galectin CRDs from vertebrates. This analysis was also conducted on sequences which share a high degree of sequence similarity to the members of the galectin family. For the purpose of this manuscript, these related sequences are referred to as "galectins" whether or not they have demonstrated affinity for beta-galactosides, which has been included in the original definition of a bona fide galectin (Barondes et al. 1994*a*).

Materials and Methods

Sequences and Mapping Data Retrieval

Sequences were retrieved from the following Web sites: www.ncbi.nlm.nih.gov/_(Blast, mapview, LocusLink and unigene); www.ensembl.org/ (Homo_sapiens, /domainview, /Mus_musculus/domainview, /Fugu_rubripes/domainview); http://genome.jgi-psf.org/fugu6/fugu6.info.html; www.jgi.doe.gov/programs/ciona/ciona_mainpage.html; http://firefly.bio.indiana.edu/; www.wormbase.org/.

Physical maps of human chromosomes were recovered from: www.ensembl.org/Homo_sapiens/syntenyview.

To undertake the phylogenetic analysis of the galectins, we systematically screened databases for galectin sequences (both protein and nucleic acid; see tables and protein alignment in the online Supplementary Material) with Blast (Altschul et al. 1997). A few sequences were reconstructed from ESTs and translated using Lasergene from DNASTAR. Only sequences known to be transcribed, based on the presence of ESTs in the databases, were included in this analysis.

Sequence Alignments

All amino acid alignments were performed with ClustalW (Thompson, Higgins, and Gibson 1994) and these alignments were manually checked with SEAVIEW (Galtier, Gouy, and Gautier 1996) or Lasergene from DNASTAR.

Phylogeny

Maximum-likelihood distances of the protein alignment were computed with Tree-Puzzle, version 5.0 (Schmidt et al. 2002), using the VT matrix (Muller and Vingron 2000), considering rate heterogeneity between sites using a gamma distribution (parameter alpha estimated from data set = 1.98 ± 0.19). Tree construction was then carried out using the BIONJ algorithm (Gascuel 1997). Bootstrap values were computed using 1,000 replicates, created with the program SEQBOOT from the PHYLIP package (Felsenstein 1993). Maximum parsimony analyses were also undertaken with the program PROTPARS of PHYLIP.

Results

Galectin Gene Location on the Human Genome

In a search for *LGALS* genes in the human genome, we have noted that several genes encoding bi-CRDs have similar sets of neighboring genes, including members of the CAPN (calpain), ACTN (actinin), and RYR (ryanodine receptor) families (fig. 1). Interestingly, this synteny is conserved in both the human and the mouse genomes as well as in the pig genome, at least in the case of *Lgals-4* (Martins-Wess et al. 2002).

These similarities in genomic organization suggest that bi-CRD galectins arose by duplications of large chromosomal segments (or "en bloc" duplications [Abi-Rached et al. 2002]). Such en bloc duplication episodes early in vertebrate history, which might be the consequence of a complete polyploidization (see Holland et al. 1994; Makalowski 2001; Spring 2002, Larhammar, Lundin, and Hallbook 2002), are thought to account for the existence of the four different *HOX* gene clusters (see Prince 2002 for a recent review) as well as the four different *MHC* gene clusters (Abi-Rached et al. 2002). Similarly, the four bi-CRD galectin encoding genes might also derive from a common ancestral bi-CRD encoding *Lgals* gene located on a chromosomal segment that underwent these large-scale duplications giving rise to the conserved synteny in regions of chromosomes 1, 11, 17, and 19. This hypothesis is also supported by the presence of *LGALS-4*, *-8*, *-9*, *-12* orthologues in many vertebrates (see below and table in online Supplementary Material). *Lgals-6* is unique among bi-CRD encoding galectins because it is likely to come from a very recent tandem duplication of *Lgals-4* and has only been identified in mice so far (Gitt et al. 1998*a*, 1998*b*).

Exon-Intron Organization of Galectin Genes

The 130 amino acid long galectin CRD is formed by two anti-parallel β-sheets, composed of five and six β-strands (labeled F1 to F5 and S1 to S6). The extensive knowledge of several galectin structures allowed the positioning of the respective β-strands on the CRD sequence alignments (see online Supplementary Material). We used these structural features as landmarks to compare the exon-intron organization of twelve *Homo sapiens*, one *Mus musculus*, two *Ciona intestinalis* (sea squirt), two *Drosophila melanogaster*, and ten *Caenorhabditis elegans* galectin encoding genes (fig. 2).

Two Types of Galectin CRDs in Vertebrates

We first examined the gene organization of human *LGALS-1* to *–12*, *HSPC159*, *PP13* and mouse *grifin* (a potential human grifin gene has been located but no human grifin transcripts have been found yet). When we compared the relative positions of the β-strands with respect to the exon-intron boundaries in the nucleic acid sequence, we found that CRDs are always encoded by three exons (fig. 2; Barondes et al. 1994*b*; Gitt et al. 1998*b*). The first exon encodes the S1 and F2 β-strands. The 5′ boundary of the second exon, which we refer to as "W" exon because it contains a highly conserved tryptophan residue, interrupts

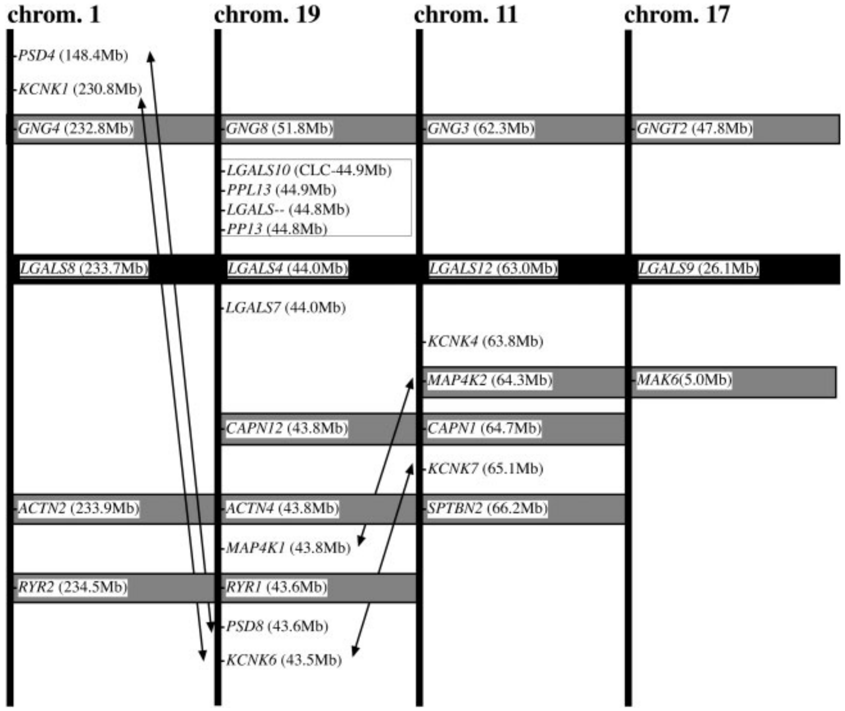

FIG. 1.—Physical map of the chromosomal regions surrounding the genes encoding the bi-CRD galectins in *Homo sapiens*. *LGLS-4*, *-8*, *-9*, and *-12* genes are located on paralogous chromosomal segments. Chromosome numbers are indicated at the top (chromosomes 1, 19, 11, 17). Gene positions along the chromosome are given in brackets as reported in www.ensembl.org/Homo_sapiens/. Galectin genes are shown in a black box. Genes shown in grey boxes have conserved positions relative to each other. Double-headed arrows point to genes of the same family, the positions of which have been shuffled relative to the others.

the codon immediately 5′ to the S3 β-strand in all *LGALS* genes except for *LGALS-10*, in which four codons separate it from the beginning of S3 (see sequences in online Supplementary Material). This exon encodes the three adjacent β-strands (S4, S5, S6) that form a pocket containing the seven residues directly involved in carbohydrate binding. The third exon encodes the F5, S2, and F1 β-strands.

We observed that there are two distinct subtypes of W exons: some ending within the sequence encoding the F4 β-strand, the others ending about 30 bases upstream, within the sequence encoding the F3 β-strand. We propose to call these two CRD subtypes, F4-CRDs and F3-CRDs, respectively. All human galectin CRDs belong to one subtype or the other and are thus encoded by one of two different exon-intron structures. Thus, the mono-CRD galectins fall in two classes: the F4-CRD galectins, which include *LGALS-7*, *LGALS-10*, and *PP13*; and the F3-CRD galectins, which include *LGALS-1*, *GALS-2*, *LGALS-3*, *grifin*, and *HSPC159*. Strikingly, in all four bi-CRD galectins, the N-terminal CRD is of the F4 subtype while the C-terminal CRD is of the F3 subtype. Interestingly, the exon/intron organization of ten *Takifugu rubripes* and four *danio rerio Lgals* genes is also very similar to their mouse and human orthologues (see online Supplementary Material).

This shared exon-intron organization strongly suggests that all vertebrate CRDs originate from a common ancestral CRD by a mechanism of duplication and divergence. Moreover, these results strongly suggest that the bi-CRD galectins come from an original bi-CRD galectin, which arose by tandem duplication of the gene encoding the ancestral mono-CRD galectin.

Two Bi-CRD Galectins in the Urochordate *Ciona intestinalis*

From ESTs and genomic sequences, we deduced the sequence of two *Lgals* genes encoding bi-CRD galectins from the urochordate *Ciona intestinalis*. The first one (*cionaLgalsa-a*) exhibits the same characteristic F4-CRD–linker–F3-CRD organization as all vertebrate bi-CRD encoding *Lgals* genes. This result strongly supports the hypothesis that the original bi-CRD encoding gene was already present in the ancestor common to the vertebrate and urochordate lineages. The genomic sequence of the second *C. intestinalis* (*cionaLgals-b*) bi-CRD encoding gene is still incomplete but it appears to have an F4-CRD–linker–F4-CRD structure (apart from an intron insertion in the N-terminal CRD), an organization never encountered in vertebrate *Lgals* genes.

Protostome Galectins

Two *D. melanogaster* bi-CRD galectins have been identified in the databases and affinity for β-galactosides has been reported for one of them (CG5335; Pace et al. 2002). Although these genes have only three exons, the border of one of them interrupts a codon 5′ to the S3 β-strand of the N-terminal CRD, as is the case in all chordate galectins. This is a rare enough feature to suggest

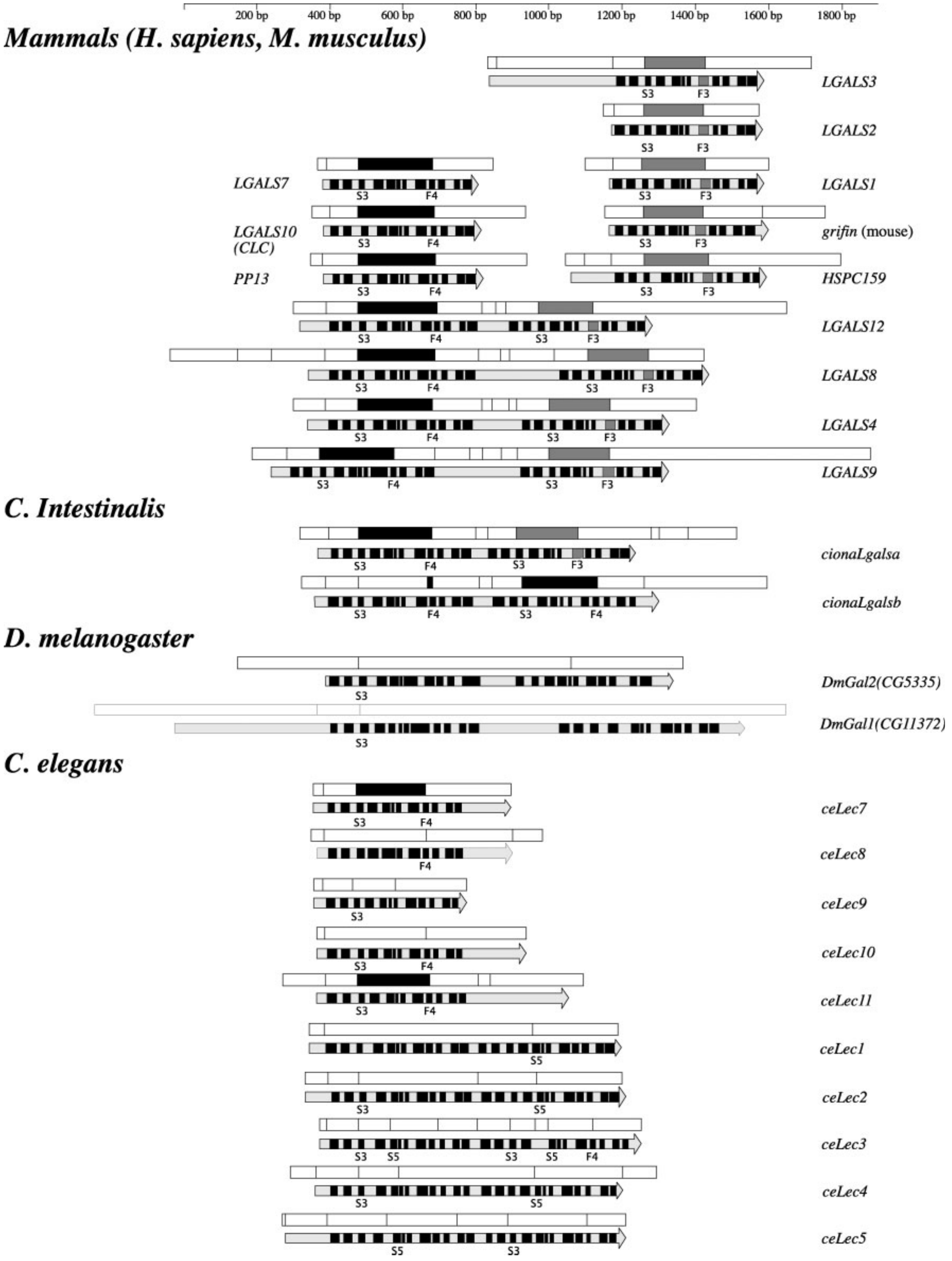

FIG. 2.—Comparison of the *Lgals* gene organization from different species. Each galectin open reading frame is indicated as a long grey arrow with boxes corresponding to the different β-strands that form the S and F β-sheets. The names of the relevant β-strands are indicated below each arrow. Exons are represented as boxes placed above each arrow. The W exons, which contain the highly conserved tryptophan residue of galectins, are either filled in black when they encode S3 to F4 sequences (F4 subtype), or they are filled in grey when they encode S3 to F3 sequences (F3 subtype). Note that all mammalian bi-CRD galectins are of the F4-F3 subtype.

a common ancestor to both chordate and *D. melanogaster* galectins.

In contrast to *D. melanogaster*, many genes encoding putative and demonstrated galectins have been identified in *C. elegans* and the genomic sequences of 10 of them are available; five have one CRD (*Lec7* to *Lec11*) and five have two CRDs (*Lec1* to *Lec5*). Although their exon-intron structures are rather different from that of chordates, *Lec7* and *Lec11* have a typical F4-CRD with a W exon terminating in the F4 β-strand. Moreover, *Lec8* and *10* have an F4-CRD-like structure, differing only by one missing intron from a typical F4-CRD. These results suggest that galectins from *C. elegans* and chordates are derived from an ancient gene present in the ancestor of both protostomes and deuterostomes. Moreover, it is possible that the original CRD motif present in the most distant species was an F4-CRD because this organization appears to be shared in both phyla whereas no F3-CRD has been identified in protostomes so far.

The bi-CRD galectins of *C. elegans* have a genomic organization different from that of vertebrates but similar to each other. In this case, the W exon often terminates just before S5 (5 out of 10 CRDs). This unique organization of genes encoding *C. elegans* bi-CRD galectins suggests that they appeared by duplication, within the protostomes lineage, of an ancestral bi-CRD galectin encoding gene. Therefore, genes from *C. elegans* and chordates that encode bi-CRD galectins are likely to come from two independent duplication episodes and there is no one-to-one relationship of orthology between bi-CRD galectins from the two phyla.

Phylogenetic Tree

We aligned the sequences from 142 CRDs: 4 from *C. intestinalis*, 30 from teleost fishes, 27 from amphibians, 8 from birds and lizards, and 73 from mammals. The alignment included only the folded CRD proper and was further refined by taking into account the knowledge of galectin structures. Three partial galectin sequences were then excluded (galectin-1 from *Podarcis hispanica* and *Ambystoma mexicanum*, and grifin from *Danio rerio*) because the sequences being too short, they reduced the number of informative sites. Sequences from *C. intestinalis* were also removed as the limits of resolution of the programs are immediately reached when introducing species that are too evolutionarily divergent (data not shown). Finally, a phylogenetic tree was built from the alignment of the 135 remaining CRDs and several points can be made (fig. 3).

As predicted from the genomic data, F4-CRDs and F3-CRDs are not intermingled with the F4-CRDs forming a clear group of their own. This clear distinction was obtained using two different methods and two different distance parameters in the phylogeny (data not shown). Therefore, the partition of galectin CRDs that we obtain by this phylogenetic analysis is remarkably identical to that obtained using the independent parameter of exon-intron structure.

The vast majority of CRDs, including those from birds, fish, and amphibians, clearly belong to one of the known groups of CRDs already defined in mammals. There is only some uncertainty for a few divergent sequences from *Xenopus laevis* (xgalectin-IIIs and VI for example; Shoji et al. 2003) and *Danio rerio* (galectin-9 like, for example), which are presumed to be polyploid species (see Aparicio 2000; Makalowski 2001; Prince 2002 for reviews).

The evolutionary tree provides further evidence in favor of the hypothesis that en bloc duplication accounts for the origin of bi-CRD galectins (see fig. 1). Because such duplications are thought to have occurred at about the origin of vertebrates, after their divergence from urochordates, it is expected that there would be one bi-CRD galectin in *C. Intestinalis*, and four in vertebrates. Indeed, figure 3 shows that bi-CRD galectins fit this prediction because four of them are found in many vertebrate species whereas only one bi-CRD with a similar F4-CRD–linker–F3-CRD organization could be identified in the urochordate *C. intestinalis*. In this regard, the fact that no galectin-12 sequences could be identified in teleost fishes comes as a surprise. However, the tree topology regarding galectin-8, -12, and HSPC159 strongly suggests the presence of galectin-12 in the ancestor common to teleost fishes and tetrapods. Its apparent absence is likely to be a result of incompleteness of the sequence databases or to its loss in the teleost fish lineage after divergence from the tetrapod lineages.

Several mono-CRD galectins are also found in a large range of species from teleost fishes to mammals (i.e., galectin-1/2, galectin-3, grifin, and HSPC159).

Galectin-1 and galectin-2 are two closely related galectins that have been isolated in a large variety of species and form their own subgroup on the evolutionary tree. Interestingly, the tree topology shows that the galectin-1/2 sequences identified in teleost fishes are equally related to galectin-1 and galectin-2 from other vertebrates. This result suggests that galectin-1 and galectin-2 diverged in tetrapod lineage after separation from the teleost fish lineage.

Grifin is a F3-CRD galectin that has been first identified in rat (Ogden et al. 1998). ESTs and genomic sequences have also been found in mice. Interestingly, ESTs corresponding to *D. rerio* and chick *grifin* transcripts have been recently identified, indicating that a *grifin* gene was already present in the ancestor common to teleost fishes and mammals. Unfortunately, the *D. rerio* sequence is only partial and could not be included in the phylogenetic tree.

The sequences of galectin-1/2, grifin, and galectin-3 are too divergent for deducing their precise evolutionary relationships with other chordate galectins in general and F3-CRD containing galectins in particular. In contrast, HSPC159 is very similar to the C-terminal CRD of galectin-12 with respect to gene organization (F3-CRD) as well as sequence. There is even more similarity between HSPC159 and galectin-12 than between galectin-8 and galectin-12 (fig. 3). These results suggest that HSPC159 originated by partial duplication of the C-terminal domain of galectin-12 after its divergence from galectin-8 and that this event occurred in the ancestor common to teleost fishes and mammals.

Some mono-CRD galectins are found in a limited number of species and, in some cases, their origin can be inferred from their position on the tree. Galectin-14, which has only been reported in sheep so far (Dunphy et al.

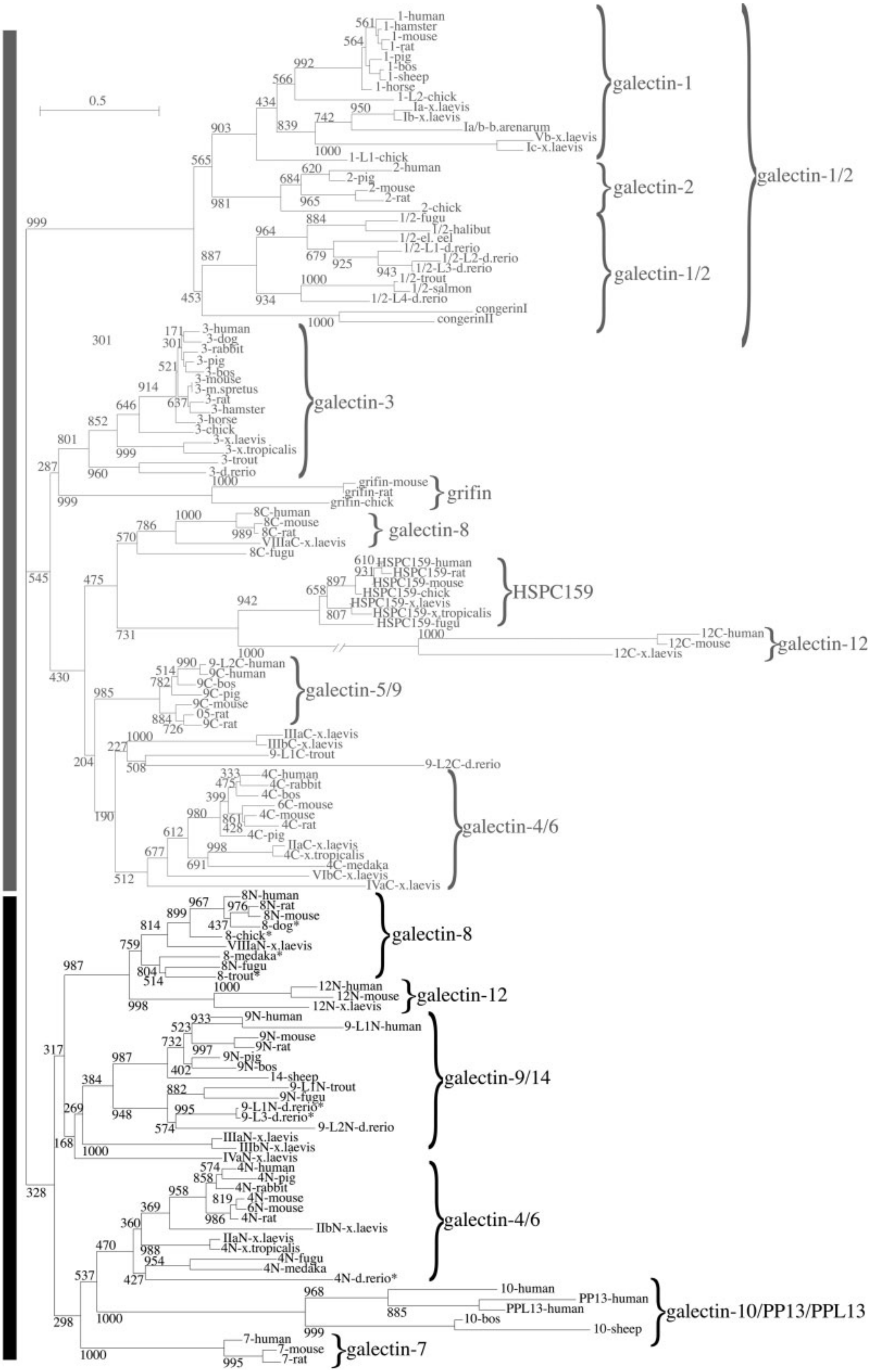

FIG. 3.—Phylogenetic tree connecting the amino acid sequences of the known chordate galectins. The tree was constructed with the distance method using the BIONJ program. The substitution rate for variable sites was determined by a gamma distribution. Pairwise maximum likelihood distances were computed with the Müller and Vingron model of substitution using Tree-Puzzle. Bootstrap values calculated using 1,000 replicates are indicated above each branch. Asterisks correspond to partial sequences. Ambystoma: *Ambystoma mexicanum*; b.arenarum: *Bufo.arenarum*; bos: *Bos taurus*; chick: *Gallus gallus*; ciona: *Ciona intestinalis*; conger: *Conger myriaster*; d.rerio: *Danio rerio*; dog: *Canis familiaris*; el.eel: *Electrophorus electricus*; fugu: *Takifugu rubripes*; halibut: *Paralichthys olivaceus*; hamster: *Cricetulus griseus*; horse: *Equus caballus*; human: *Homo sapiens*;

2002), appears to derive from the Nterm-CRD of galectin-9. Similarly, galectin-5 has only reported in rat and appears to come from the Cterm-CRD of galectin-9. In these cases, a mono-CRD galectin seems to have arisen by partial duplication of a bi-CRD galectin.

We also note that galectin-7 on one hand, and galectin-10 and galectin-10-related sequences (PP13 and PPL13) on the other, group together with the Nterm-CRD of galectin-4. Although the bootstrap values are very low, this result is found in several trees constructed by different methods. This result suggests that galectin-10, PP13, PPL13, and galectin-7 might have all arisen by duplication of the Nterm-CRD of galectin-4. This hypothesis is also supported (i) by the fact that they are all of the F4-CRD subtype; (ii) by their organization in a tight cluster on human chromosome 19 suggestive of a tandem duplication (see fig. 1); and (iii) by the similarity of the S3-S4 loop in galectin-4, galectin-10, PP13, PPL13, and galectin-7 (see alignment in online Supplementary Material).

Relationships Between Sequence, Phylogeny, and Functional Properties

The alignment of sequences available for chordate galectins shows a very conserved organization with no major event of insertion or deletion (see alignment in online Supplementary Material). Most of the part of the molecule directly involved in carbohydrate binding, including the S4, S5, and S6 β-strands as well as the loops between S5 and S6 and between S6 and F3, is very conserved. In contrast, the length of the S3-S4 and S4-S5 loops varies significantly between different subgroups, indicating that these loops might be involved in the fine-tuning of carbohydrate specificity.

The seven conserved amino acids known to interact with bound β-galactosides are present in most of the CRDs analysed. For instance, the key tryptophan in the W exon is found in 139 out of the 142 sequences. Interestingly, the C-term CRD of galectin-12, as well as HSPC159 and galectin-10, lack several of these residues. Therefore, they may have little or no β-galactoside-binding activity as has already been shown for galectin-10 (Dyer and Rosenberg 1996). As a consequence, although galectin-12 is a bi-CRD galectin, whether it acts as a bivalent galectin or not needs to be experimentally tested. Similarly, in *C. intestinalis* galectin-b, five out of the seven key residues are present in the N-terminal CRD but only three out of seven in its C-terminal domain. The ability of the latter to bind galactose is therefore very questionable. In *C. intestinalis* galectin-a, all seven key residues are conserved in the F3-CRD. In the F4-CRD, one of the key residues is missing but it is a conservative substitution (R to K). This suggests that both domains of this bi-CRD galectin are likely to bind β-galactosides.

Many, but not all, of the galectins in the main galectin-1/2 subgroup, including galectins from fish, amphibians, chickens, and mammals, form dimers (see Cooper 2002). As a consequence, these proteins are functionally bivalent. Structural analysis has shown a site of dimerization involving the S1 and F1 β-strands for two fish galectins (congerins; Shirai et al. 2002), a toad galectin (Bianchet et al. 2000), a chick galectin (Varela et al. 1999), and galectins-1 and -2 in mammals (Rini and Lobsanov 1999). It is clear that this potential to form dimers is not restricted to the galectin-1/2 subgroup, because grifin has also been identified as dimers in rat (Ogden et al. 1998).

In contrast, most other mono-CRD galectins do not form non-covalent dimers in biological conditions. In two cases, galectin-7 (Leonidas et al. 1998) and galectin-3 CRD (Seetharaman et al. 1998), dimers have been observed, but in each case the situation is entirely different from the one observed in the galectin-1 like group: a convex side of the β-sandwich (near strands F5 and S2) is involved and dimerization at the S1-F1 interface is prevented by a conserved "hook" sequence immediately preceding S1 (Lobsanov et al. 1993; Seetharaman et al. 1998). This sequence, valine-proline-tyrosine or similar hydrophobic-proline-aromatic triplet, is found in many F4- and F3-CRDs, but not those in the galectin-1 or grifin subgroups (see alignment in online Supplementary Material).

Discussion

The Chordate Galectins Share a Common Ancestor

In figure 4, we propose a model of galectin phylogeny based on the analysis of human galectin chromosomal localization, exon-intron organization, and CRD sequences.

We have shown here that all the CRDs identified in vertebrates are encoded by three exons with very similar borders. The first exon encodes the F1 and S2 β-strands, The second exon, with a 5′ boundary systematically interrupting a codon 5′ to the S3 β-strand, encodes most of the amino acids involved in carbohydrate binding and the third exon encodes the last four or five β-strands. This conserved organization, as well as obvious sequence similarities, strongly suggests that all CRDs from chordate mono- and bi-CRD galectins evolved by duplication and divergence from an ancestral mono-CRD galectin. Therefore, the most parsimonious hypothesis would be that a tandem duplication of this ancestral mono-CRD galectin gave rise to a first bi-CRD galectin.

Two subtypes of CRDs can be distinguished depending on the position of the intron 3′ to the W exon—the N-terminal CRDs with a W exon ending within the sequence encoding the F4 β–strand and the C-terminal CRDs with a W exon ending within the sequence encoding the F3 β–strand. Because an F4-CRD–linker–F3-CRD gene structure is shared between all vertebrate bi-CRD galectins and one bi-CRD galectin from *C. intestinalis*, the ancestral bi-CRD galectin was presumably present in the ancestor common to all chordates. More precise timing of these events, such as whether they occurred before or after the split between protostomes and deuterostomes, is still beyond reach due to the divergence both in sequence and exon-intron organization between these phyla. The presence of the genes encoding bi-CRD galectins-4, -8, -9, and -12 on paralogous

←

medaka: *Oryzias latipes*; mouse: *Mus musculus*; m.spretus: *Mus spretus*; pig: *Sus scrofa*; podarcis: *Podarcis hispanica*; rabbit: *Oryctolagus cuniculus*; rat: *Rattus norvegicus*; salmon: *Salmo sala*; sheep: *Ovis aries*; trout: *Oncorhynchus mykiss*; x.laevis: *Xenopus laevis*; x.tropicalis: *Xenopus tropicalis*.

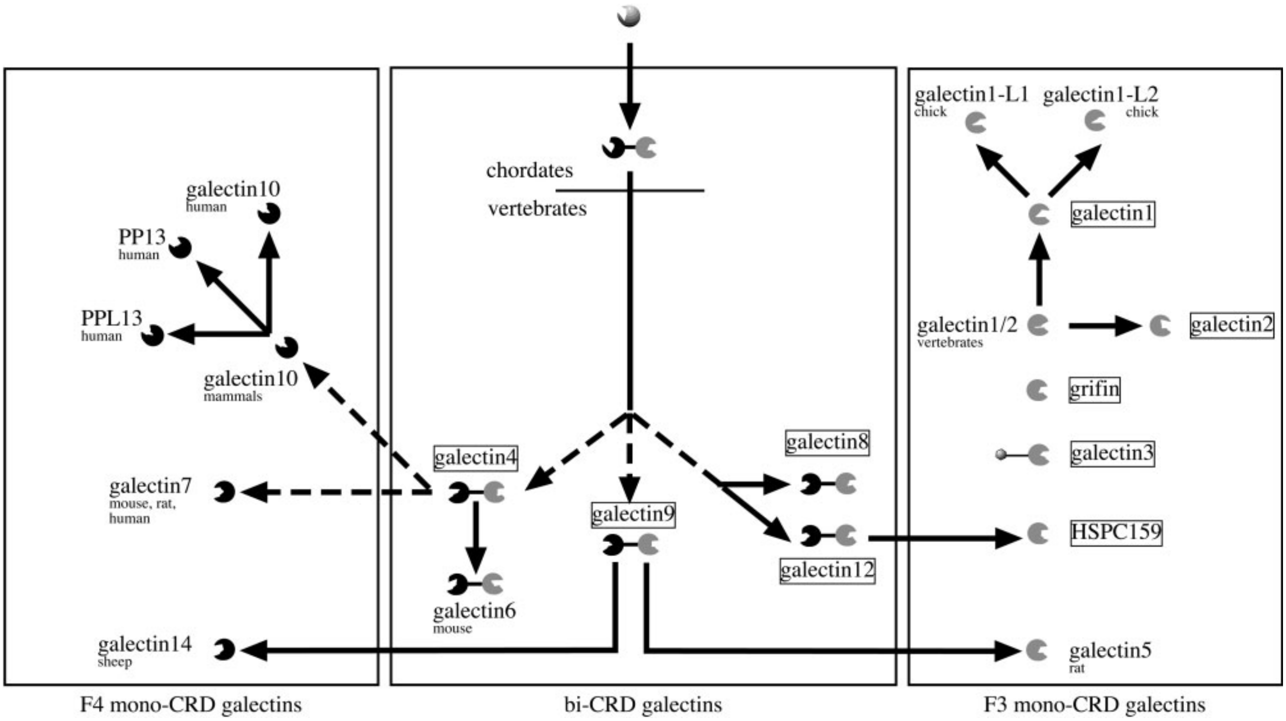

FIG. 4.—Proposed phylogenetic relationships between chordate galectins based on gene location, exon-intron structure and sequence based phylogenetic tree. Dotted arrows indicate uncertainty in the sequence of divergence. The two types of CRDs are shown in black (F4-CRDs) or grey (F3-CRDs). Boxed galectins have been identified in a large variety of species. Unboxed galectins have only been identified in a subset of species (indicated under the name of the galectin).

regions of human chromosomes 1, 11, 17, and 19 indicates a subsequent "en bloc" duplication of this ancestral bi-CRD galectin, maybe as part of the block duplications that occurred in the vertebrate genome (Abi-Rached et al. 2002) after divergence from urochordates and before divergence of teleost fishes and tetrapods.

Gene structure as well as sequence comparison strongly suggests that several genes encoding mono-CRD galectins subsequently arose from either partial duplication or partial deletion of the bi-CRD galectins. For example, HSPC159 is likely to have diverged from the C-terminal domain of Galectin-12 before the divergence of teleost fishes and tetrapods. As a consequence, it is present in a wide variety of species from *D. rerio* to *H. sapiens*. In contrast, some galectins are present in a very limited number of species and are likely to be the consequence of recent duplications or deletions. An interesting case is that of galectin-10, PP13, PPL13, and maybe galectin-7, which presumably arose from the N-terminal domain of galectin-4 as suggested by genome localization (suggestive of a tandem duplication), gene structure (F4-CRD), and phylogenetic tree topology. These clustered galectins have been identified in a limited number of species, which suggests that they result from a recent tandem duplication, and yet they have divergent sequences, which is not compatible with this hypothesis. Several explanations for this situation can be proposed. (i) It is possible that these galectins arose early and are present in many species but have not yet been identified because of their very limited patterns of expression. For example, grifin expression is restricted to the cornea in rat (Ogden et al. 1998) and, to date, only one *D. rerio* and three chick *grifin* ESTs have been identified. Similarly, Galectin-7 is specific to stratified epithelia (Madsen et al. 1995; Magnaldo, Bernerd, and Darmon 1995; Magnaldo, Fowlis, and Darmon 1998; Timmons et al. 1999) and galectin-10 seems to be mainly expressed in pathological situations (Ackerman et al. 1993; Dunphy et al. 2000). (ii) These galectins might have arisen "early" but they have not been identified in other species because they were not beneficial in most lineages and lost. (iii) Another possibility is that these galectins appeared "late" but they diverged quickly before being fixed by natural selection. From this point of view, it is noteworthy that the Ka/Ks ratios for the galectin-10 subgroup are suggestive of a family that is not yet under any selective constraint (data not shown).

The nature of the original mono-CRD galectin remains uncertain. Because F4-CRD galectins are present in both protostomes (*ceLec7* and *ceLec11* in *C. elegans*) and deuterostomes, it is possible that the ancestral mono-CRD galectin, common to both lineages, was of the F4-type. A tandem duplication of this F4-CRD galectin would have given rise to the ancestral F4-CRD–linker–F3-CRD organization found in chordate galectins. If this were the case, then the F3-CRD containing galectins -1/2, -3, and grifin might have subsequently originated by partial duplication of the C-terminal domain of this bi-CRD galectin. This hypothesis is the most parsimonious one, as it postulates a single event leading to the F3-CRDs of both mono- and bi-CRD galectins. However, confirmations will

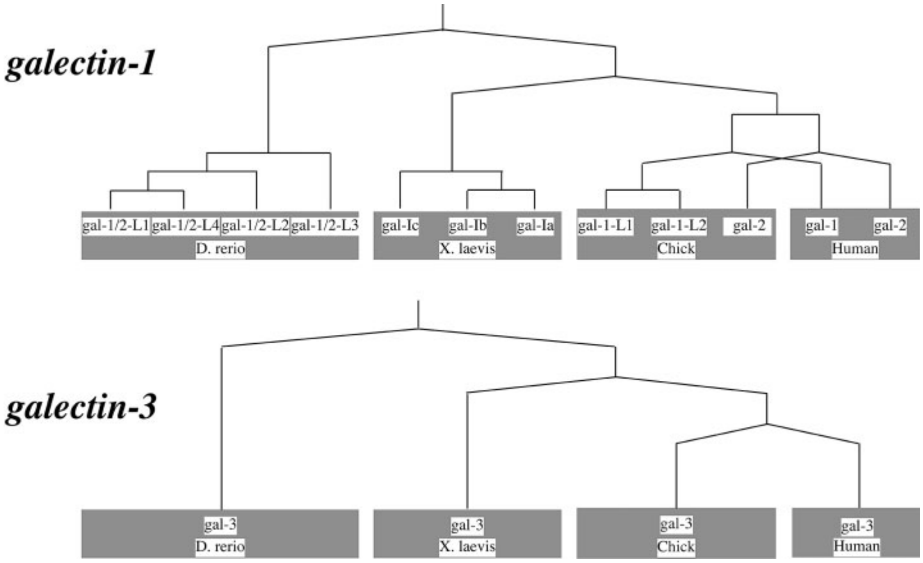

FIG. 5.—Duplication and divergence events in the galectin-1/2 (top) and the galectin-3 (bottom) subgroups. Horizontal lines indicate gene duplications; oblique lines indicate speciation events.

probably only come from the isolation of members of the galectin family from intermediate lineages, such as echinoderms and cyclostomes (e.g., lampreys and hagfish).

Functions of the Different Galectins

Over the years, the comparison of different animal models has allowed biologists to unravel many gene functions, but a prerequisite for this approach is to identify equivalent genes in different species. Refining the relationships of orthology and paralogy within the galectin family was one of the goals of this study. The situation is different depending on the galectin subgroup, as illustrated in figure 5 using two extreme examples. Four galectin-1/2-like genes have been identified in *D. rerio*, three in *X. laevis*, three in *G. gallus*, two in mammals, and, in each case, they result from lineage-specific duplication events. Therefore, human galectin-1 is equally orthologous to the four *D. rerio* or to the two chicken galectin-1-like genes and transposing results on galectin-1 function from one model organism to another might be difficult. The galectin-3 subgroup is different, because a single member has been identified per species (from fish to human). We can thus reasonably assume that they are orthologues and that function(s) of galectin-3 genes might be retained during evolution. It is not clear whether the differences between the evolutionary histories of the genes orthologous to galectin-1 and galectin-3 result from differences in protein function, such as toxicity of an extra-copy, or if they reflect differences in the genes themselves, such as localization to chromosomal regions that are more or less likely to be duplicated.

Our results show that some galectins have been duplicated very recently in several lineages, which raises the issue of functional redundancy. For example, there are four closely related galectins-1 in *D. rerio*, two closely related bi-CRD galectins (4 and 6) in mice, and a large group of galectin-10-related proteins in humans. In the case of galectin-4 and -6 that have very similar expression patterns in the mouse digestive tract (Gitt et al. 1998*a*), it is possible that they have at least partly overlapping functions.

Because some mono-CRD galectins arose by partial duplication of some bi-CRD galectins, there is a possibility of agonistic or antagonistic action between galectins if they are expressed in the same cell. There is a precedent for this situation because a CRD-containing fragment of galectin-3 has been shown to inhibit the cell signalling activity of intact galectin-3 (Sano et al. 2000) and to have a therapeutic effect in a model of breast cancer (John et al. 2003).

Conclusions

On the basis of their domain organization, galectins have often been divided into three groups: prototype, chimera type and tandem-repeat type (Hirabayashi and Kasai 1993). This nomenclature describes the organization of the galectin proteins, but it does not accurately reflect the evolutionary relationships between them as revealed by the present analysis. The prototype group was originally defined by similarity to galectin-1, the first family member discovered, but it is now generally used to refer to all mono-CRD galectins except galectin-3. However, it appears from our study that galectin-1 is not the evolutionary prototype and that the majority of the other mono-CRD galectins are not closely related to galectin-1. Rather, there are two subgroups of mono-CRD galectins, the F4- and F3-types, which are evolutionarily quite distinct from each other, and the chimeric galectin-3 is actually a member of the F3 subgroup. We have also shown that bi-CRD galectins derived from an ancestral tandem-duplication of a mono-CRD galectin before or early in chordate evolution. Since then, their N-term and C-term CRDs have independently evolved so that bi-CRD galectins cannot be strictly considered as tandem-repeat any longer.

Supplementary Material

Sequence alignment from which the phylogenetic tree shown in figure 3 was built.

Tables indicating the references of all the sequences that have been used.

Sequences of the galectins used in figure 2 on which the position of the different β-strands is indicated.

Sequences of fish galectins.

Acknowledgments

We would like to thank Pierre Netter for helpful discussions during the realization of this work. This work was supported by the CNRS (Centre National de la Recherche Scientifique), Universities Paris 6 and 7, GEFLUC (Groupement des Entreprises Françaises dans la Lutte contre le Cancer), ARC (Association pour la Recherche contre le Cancer, grants number 4672 and 4680) and la Ligue contre le comité de Paris. S.S. was a recipient of an FRM (Fondation pour la Recherche Médicale) fellowship.

Literature Cited

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

Ackerman, S. J., S. E. Corrette, H. F. Rosenberg, J. C. Bennett, D. M. Mastrianni, A. Nicholson-Weller, P. F. Weller, D. T. Chin, and D. G. Tenen. 1993. Molecular cloning and characterization of human eosinophil Charcot-Leyden crystal protein (lysophospholipase). Similarities to IgE binding proteins and the S-type animal lectin superfamily. J. Immunol. **150**:456–468.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped Blast and PSI-Blast: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

Andre, S., S. Kojima, N. Yamazaki, C. Fink, H. Kaltner, K. Kayser, and H. J. Gabius. 1999. Galectins-1 and -3 and their ligands in tumor biology. Non-uniform properties in cell-surface presentation and modulation of adhesion to matrix glycoproteins for various tumor cell lines, in biodistribution of free and liposome-bound galectins and in their expression by breast and colorectal carcinomas with/without metastatic propensity. J. Cancer Res. Clin. Oncol. **125**:461–474.

Aparicio, S. 2000. Vertebrate evolution: recent perspectives from fish. Trends Genet. **16**:54–56.

Barondes, S. H., V. Castronovo, D. N. Cooper, et al. (21 co-authors) 1994*a*. Galectins: a family of animal beta-galactoside-binding lectins. Cell **76**:597–598.

Barondes, S. H., D. N. Cooper, M. A. Gitt, and H. Leffler. 1994*b*. Galectins. Structure and function of a large family of animal lectins. J. Biol. Chem. **269**:20807–20810.

Bianchet, M. A., H. Ahmed, G. R. Vasta, and L. M. Amzel. 2000. Soluble beta-galactosyl-binding lectin (galectin) from toad ovary: crystallographic studies of two protein-sugar complexes. Proteins **40**:378–388.

Colnot, C., D. Fowlis, M. A. Ripoche, I. Bouchaert, and F. Poirier. 1998. Embryonic implantation in galectin 1/galectin 3 double mutant mice. Dev. Dyn. **211**:306–313.

Cooper, D. 2002. Galectinomics: finding themes in complexity. Biochim. Biophys. Acta **1572**:209–231.

Danguy, A., I. Camby, and R. Kiss. 2002. Galectins and cancer. Biochim. Biophys. Acta **1572**:285–293.

Dunphy, J. L., A. Balic, G. J. Barcham, A. J. Horvath, A. D. Nash, and E. N. Meeusen. 2000. Isolation and characterization of a novel inducible mammalian galectin. J. Biol. Chem. **275**:32106–32113.

Dunphy, J. L., G. J. Barcham, R. J. Bischof, A. R. Young, A. Nash, and E. N. Meeusen. 2002. Isolation and characterization of a novel eosinophil-specific galectin released into the lungs in response to allergen challenge. J. Biol. Chem. **277**:14916–14924.

Dyer, K. D., and H. F. Rosenberg. 1996. Eosinophil Charcot-Leyden crystal protein binds to beta-galactoside sugars. Life Sci. **58**:2073–2082.

Felsenstein, J. 1993. PHYLIP (phylogeny inference package). Version 3.5c. Distributed by the author, Department of Genetics, University of Washington, Seattle.

Gabius, H. 2002. Animal lectins and life: a guided tour into the realm of the sugar code. Biochim. Biophys. Acta **1572**: 163–164.

Gabius, H., S. Andre, H. Kaltner, and H. Siebert. 2002. The sugar code: functional lectinomics. Biochim. Biophys. Acta **1572**: 165–177.

Galtier, N., M. Gouy, and C. Gautier. 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput. Appl. Biosci. **12**:543–548.

Gascuel, O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. **14**:685–695.

Gitt, M. A., and S. H. Barondes. 1991. Genomic sequence and organization of two members of a human lectin gene family. Biochemistry **30**:82–89.

Gitt, M. A., C. Colnot, F. Poirier, K. J. Nani, S. H. Barondes, and H. Leffler. 1998*a*. Galectin-4 and galectin-6 are two closely related lectins expressed in mouse gastrointestinal tract. J. Biol. Chem. **273**:2954–2960.

Gitt, M. A., Y. R. Xia, R. E. Atchison, A. J. Lusis, S. H. Barondes, and H. Leffler. 1998*b*. Sequence, structure, and chromosomal mapping of the mouse Lgals6 gene, encoding galectin-6. J. Biol. Chem. **273**:2961–2970.

Hirabayashi, J., and K. Kasai. 1993. The family of metazoan metal-independent beta-galactoside-binding lectins: structure, function and molecular evolution. Glycobiology **3**:297–304.

Holland, P. W., J. Garcia-Fernandez, N. A. Williams, and A. Sidow. 1994. Gene duplications and the origins of vertebrate development. **Dev Suppl**:125–133.

Honjo, Y., P. Nangia-Makker, H. Inohara, and A. Raz. 2001. Down-regulation of galectin-3 suppresses tumorigenicity of human breast carcinoma cells. Clin. Cancer Res. **7**:661–668.

Hughes, A. L. 2001. Evolution of the integrin alpha and beta protein families. J. Mol. Evol. **52**:63–72.

John, C. M., H. Leffler, B. Kahl-Knutsson, I. Svensson, and G. A. Jarvis. 2003. Truncated galectin-3 inhibits tumor growth and metastasis in orthotopic nude mouse model of human breast cancer. Clin. Cancer Res. **9**:2374–2383.

Kilpatrick, D. 2002. Animal lectins: a historical introduction and overview. Biochim. Biophys. Acta **1572**:187–197.

Kim, H. R., H. M. Lin, H. Biliran, and A. Raz. 1999. Cell cycle arrest and inhibition of anoikis by galectin-3 in human breast epithelial cells. Cancer Res. **59**:4148–4154.

Kopitz, J., S. Andre, C. Von Reitzenstein et al. (12 co-authors). 2003. Homodimeric galectin-7 (p53-induced gene 1) is a negative growth regulator for human neuroblastoma cells. Oncogene **22**:6277–6288.

Larhammar, D., L. G. Lundin, and F. Hallbook. 2002. The human Hox-bearing chromosome regions did arise by block or chromosome (or even genome) duplications. Genome Res. **12**:1910–1920.

Leffler, H. 2001. Galectins structure and function—a synopsis. Results Probl. Cell Differ. **33**:57–83.

Leonidas, D. D., E. H. Vatzaki, H. Vorum, J. E. Celis, P. Madsen, and K. R. Acharya. 1998. Structural basis for the recognition of carbohydrates by human galectin-7. Biochemistry **37**:13930–13940.

Liu, F., R. Patterson, and J. Wang. 2002. Intracellular functions of galectins. Biochim. Biophys. Acta **1572**:263–273.

Lobsanov, Y. D., M. A. Gitt, H. Leffler, S. H. Barondes, and J. M. Rini. 1993. X-ray crystal structure of the human dimeric S-Lac lectin, L-14-II, in complex with lactose at 2.9-A resolution. J. Biol. Chem. **268**:27034–27038.

Loris, R. 2002. Principles of structures of animal and plant lectins. Biochim. Biophys. Acta **1572**:198–208.

Madsen, P., H. H. Rasmussen, T. Flint, P. Gromov, T. A. Kruse, B. Honore, H. Vorum, and J. E. Celis. 1995. Cloning, expression, and chromosome mapping of human galectin-7. J. Biol. Chem. **270**:5823–5829.

Magnaldo, T., F. Bernerd, and M. Darmon. 1995. Galectin-7, a human 14-kDa S-lectin, specifically expressed in keratinocytes and sensitive to retinoic acid. Dev. Biol. **168**:259–271.

Magnaldo, T., D. Fowlis, and M. Darmon. 1998. Galectin-7, a marker of all types of stratified epithelia. Differentiation **63**:159–168.

Makalowski, W. 2001. Are we polyploids? A brief history of one hypothesis. Genome Res. **11**:667–670.

Martins-Wess, F., R. Voss-Nemitz, C. Drogemuller, B. Brenig, and T. Leeb. 2002. Construction of a 1.2-Mb BAC/PAC contig of the porcine gene RYR1 region on SSC 6q1.2 and comparative analysis with HSA 19q13.13. Genomics **80**:416–422.

Muller, T., and M. Vingron. 2000. Modeling amino acid replacement. J. Comput. Biol. **7**:761–776.

Nangia-Makker, P., Y. Honjo, R. Sarvis, S. Akahani, V. Hogan, K. J. Pienta, and A. Raz. 2000. Galectin-3 induces endothelial cell morphogenesis and angiogenesis. Am. J. Pathol. **156**:899–909.

Ogden, A. T., I. Nunes, K. Ko, S. Wu, C. S. Hines, A. F. Wang, R. S. Hegde, and R. A. Lang. 1998. Grifin, a novel lens-specific protein related to the galectin family. J. Biol. Chem. **273**:28889–28896.

Pace, K. E., T. Lebestky, T. Hummel, P. Arnoux, K. Kwan, and L. G. Baum. 2002. Characterization of a novel *Drosophila melanogaster* galectin. Expression in developing immune, neural, and muscle tissues. J. Biol. Chem. **277**:13091–13098.

Poirier, F., and E. J. Robertson. 1993. Normal development of mice carrying a null mutation in the gene encoding the L14 S-type lectin. Development **119**:1229–1236.

Prince, V. 2002. The Hox paradox: more complex(es) than imagined. Dev. Biol. **249**:1–15.

Rabinovich, G., N. Rubinstein, and M. Toscano. 2002. Role of galectins in inflammatory and immunomodulatory processes. Biochim. Biophys. Acta **1572**:274–284.

Rini, J. M., and Y. D. Lobsanov. 1999. New animal lectin structures. Curr. Opin. Struct. Biol. **9**:578–584.

Sacchettini, J. C., L. G. Baum, and C. F. Brewer. 2001. Multivalent protein-carbohydrate interactions. a new paradigm for supermolecular assembly and signal transduction. Biochemistry **40**:3009–3015.

Sano, H., D. K. Hsu, L. Yu, J. R. Apgar, I. Kuwabara, T. Yamanaka, M. Hirashima, and F. T. Liu. 2000. Human galectin-3 is a novel chemoattractant for monocytes and macrophages. J. Immunol. **165**:2156–2164.

Schmidt, H. A., K. Strimmer, M. Vingron, and A. von Haeseler. 2002. Tree-Puzzle: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics **18**:502–504.

Seetharaman, J., A. Kanigsberg, R. Slaaby, H. Leffler, S. H. Barondes, and J. M. Rini. 1998. X-ray crystal structure of the human galectin-3 carbohydrate recognition domain at 2.1-A resolution. J. Biol. Chem. **273**:13047–13052.

Shirai, T., Y. Matsui, C. Shionyu-Mitsuyama, T. Yamane, H. Kamiya, C. Ishii, T. Ogawa, and K. Muramoto. 2002. Crystal structure of a conger eel galectin (congerin II) at 1.45A resolution: implication for the accelerated evolution of a new ligand-binding site following gene duplication. J. Mol. Biol. **321**:879–889.

Shoji, H., N. Nishi, M. Hirashima, and T. Nakamura. 2003. Characterization of the Xenopus galectin family. Three structurally different types as in mammals and regulated expression during embryogenesis. J. Biol. Chem. **278**:12285–12293.

Spring, J. 2002. Genome duplication strikes back. Nat. Genet. **31**:128–129.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

Timmons, P. M., C. Colnot, I. Cail, F. Poirier, and T. Magnaldo. 1999. Expression of galectin-7 during epithelial development coincides with the onset of stratification. Int. J. Dev. Biol. **43**:229–235.

Varela, P. F., D. Solis, T. Diaz-Maurino, H. Kaltner, H. J. Gabius, and A. Romero. 1999. The 2.15 A crystal structure of CG-16, the developmentally regulated homodimeric chicken galectin. J. Mol. Biol. **294**:537–549.

David Irwin, Associate Editor

Accepted December 15, 2003