J Mol Evol (2000) 51:363–373
DOI: 10.1007/s002390010098

JOURNAL OF MOLECULAR EVOLUTION

Biglycan-Like Extracellular Matrix Genes of Agnathans and Teleosts

Seikou Shintani,* Akie Sato, Satoru Toyosawa,† Colm O'hUigin, Jan Klein

Max-Planck-Institut für Biologie, Abteilung Immungenetik, Corrensstrasse 42, D-72076, Tübingen, Germany

Received: 13 April 2000 / Accepted: 4 July 2000

Abstract. Biglycan and decorin are two members of a family of small extracellular matrix proteoglycans characterized by the presence of 10 leucine-rich repeats and one or two attachment sites for glucosaminoglycans. Both have thus far been described only from tetrapod species, mainly mammals. Because the extracellular matrix has played an important part in the evolution of Metazoa, the phylogeny of its components is of considerable interest. In this study, biglycan-like (BGL) cDNA sequences have been obtained from two teleost (*Oreochromis* cichlid and zebrafish) and two lamprey species. The analysis of the sequences suggests that, like tetrapods, the lampreys possess two types of proteoglycans, both of which are biglycan-like; decorin-like proteoglycans could not be identified in these species. The genes specifying these two types apparently arose by duplication in the lamprey lineage after its divergence from gnathostomes. The two teleost species possess a BGL proteoglycan and a bona fide decorin. The BGL proteoglycan is highly divergent from the tetrapod biglycan and related to the BGL proteoglycans of the lamprey. Hence, although the duplication generating the ancestors of biglycan and decorin genes occurred after the divergence of agnathans but before the emergence of teleosts, only decorin acquired its characteristic properties in the bony fishes. The *BGL* gene presumably turned into a typical biglycan only in the tetrapod lineages. The presumed acquisitions of new functions appear to have been accompanied by changes in the evolutionary rate.

Key words: *Petromyzon* — *Oreochromis* — Zebrafish — Biglycan — Decorin — Proteoglycans

Introduction

An essential innovation in the evolution of metazoan multicellularity was the appearance of the extracellular matrix, an intricate network of macromolecules deposited between cells, especially well developed in connective tissues (Gerhart and Kirschner 1997). The extracellular matrix isolates cells from one another, creates a medium through which cells can move and be guided to specific locations, participates in signal transduction, provides conditions for cell differentiation, and fulfills a host of other functions in the development and maintenance of the body (Alberts et al. 1989). Among the main macromolecular components of the noncollagenous extracellular matrix are proteoglycans composed of a core protein and a varying number of covalently linked glycosaminoglycan (GAG) side chains (Kresse et al. 1994; Gehron Robey 1996; Iozzo and Murdoch 1996). Proteoglycans are the main constituents of cartilage, in which they contain only a few such side chains. The two main species in this latter class of proteoglycans are biglycan (also called PG I; Fisher et al. 1983, 1989; Neame et al. 1989; Wegrowski et al. 1995) and decorin (PG II; Krusuis and Ruoslahti 1986; Day et al. 1986, 1987; Li et al. 1992; Scholzen et al. 1994). Biglycan has two and decorin a single potential GAG attachment site containing the Ser-Gly sequence preceded by one or more acidic amino acid residues (Bourdon et al. 1987). In some of the biglycan molecules, however, one of the two sites may be unoccupied.

* *Present address:* Department of Pedodontics, Osaka University, Faculty of Dentistry, 1-8, Yamadaoka, Suita, Osaka 585-0871, Japan
† *Present address:* Department of Oral Pathology, Osaka University, Faculty of Dentistry, 1-8, Yamadaoka, Suita, Osaka 585-0871, Japan
Correspondence to: J. Klein

Biglycan and decorin are members of the small leucine-rich proteoglycan (SLRP) family characterized by the leucine-rich repeat (LRR), usually present in a tandem array (Iozzo and Murdoch 1996). The LRR domain has also been found in other proteins that together with SLRPs form the LRR superfamily. The LRR domain of the SLRP family is, however, distinguished from that of the other members of the superfamily in that it is flanked by cysteine clusters and in that each LRR consists of a β-strand LxxLxLxxNyL running parallel to an α-helix xaxxyayyyyayyxyy (where x is any amino acid residue; L leucine or isoleucine; N asparagine, which can, however, be replaced by cysteine or threonine; a an aliphatic residue; and y indicates a loose amino acid requirement; Kobe and Deisenhofer 1993). The cysteine clusters at the N- and C-termini of the LRR domain are composed of four and two similarly spaced cysteine residues, respectively, the former in a stretch of 20 amino acid residues. The cysteines are believed to be involved in the formation of intrachain disulfide bonds. Based on the organization of the cysteine clusters, the SLRP family can be divided into three subfamilies, one of which is comprised of biglycan and decorin. The N-terminal cysteine cluster of these two proteoglycans has the structure Cys-X_3-Cys-X-Cys-X_6-Cys, where X is any amino acid residue (Hocking et al. 1998).

The human and mouse biglycan and decorin core proteins each consist of 10 LRRs, they share 55–57% amino acid identity (Fisher et al. 1989; Scholzen et al. 1994; Wegrowski et al. 1995), and their encoding genes have nearly identical exon-intron organizations (Fisher et al. 1991; Vetter et al. 1993; Scholzen et al. 1994). Despite these similarities, however, the two proteoglycans appear to have different functions, as suggested by their distinct patterns of temporal and spatial expression (Hocking et al. 1998). The human biglycan and decorin genes have been mapped to the Xq27–q28 (Traupe et al. 1992) and 12q21–q22 (Pulkkinen et al. 1992) chromosomal regions, respectively. The two genes (cDNAs) have been cloned from human (Fisher et al. 1989; Krusius and Ruoslahti 1986), mouse (Wegrowski et al. 1995; Scholzen et al. 1994), rat (Dreher et al. 1990; Kokenyesi and Woessner 1989), rabbit (Boykiw et al. 1998; accession code U03394), cattle (Marcum et al. 1993; Day et al. 1987), sheep (accession codes AF034842 and AF125041), and dog (accession codes U83140 and U83141); the biglycan gene (cDNA) has also been isolated from the African clawed toad (*Xenopus laevis;* accession code AB037269), and the decorin sequence is also available from the pig (accession code AF125537) and the domestic fowl (Li et al. 1992).

The emergence of extracellular matrix-rich cartilage, bone, dentin, and other forms of connective tissue represents a major evolutionary innovation associated with the appearance and rise of vertebrates. It is therefore of considerable interest to characterize the molecular basis of this innovation. With this goal in mind, we initiated a search for lower vertebrate homologs of genes that are known to play an important part in the manufacture of the extracellular matrix in tetrapods (Toyosawa et al. 1998, 1999a, 1999b, 2000). The aim of the present study was to trace the evolutionary origin of the biglycan and decorin genes in jawless vertebrates and bony fish.

Materials and Methods

Animals. Adults of the cichlid fish *Oreochromis niloticus* were obtained from a dealer (Limnotherm, Bergheim-Niederaussen, Germany). Stocks of *O. niloticus* were maintained at the Max-Planck-Institut für Biologie, Tübingen. Ammocoetes larvae and adults of the sea lamprey *Petromyzon marinus* were from Lakes Huron and Champlain, USA.

Isolation of DNA (RNA) and cDNA Library Construction. Total RNA was isolated from the gut of *P. marinus* larvae from Lake Huron, and a cDNA library was constructed as described in an earlier publication (Mayer and Tichy 1995). The library had a postamplification titer of 1.0×10^{10} plaque-forming units per ml (pfu/ml). Jaws from adult *O. niloticus* were frozen in liquid nitrogen immediately after excision and kept at −70°C until their use. Total RNA was extracted from the powder of the frozen tissue samples. Poly (A^+) RNA was isolated, and cDNA was synthesized with the help of the mRNA purification kit (Amersham Pharmacia Biotech, Freiburg, Germany) and the TimeSaver cDNA synthesis kit (Amersham Pharmacia Biotech), respectively. The cDNA was inserted into the *Eco* RI-digested λgt10 vector (Stratagene, Heidelberg, Germany), in vitro-packaged with the help of the Gigapack® cloning kit (Stratagene), and used to transform competent *E. coli* MN514 bacteria. The initial titer of the library was 3.2×10^5 pfu/ml; the library was amplified once to a titer of 2.6×10^{11} pfu/ml.

Polymerase Chain Reaction (PCR) Amplification. The cDNA lysates were PCR-amplified in the PTC-100 Thermal Cycler (Biozym, Oldendorf, Germany). One microliter of a solution containing 1.0×10^7 or 2.6×10^8 pfu of the cDNA library was added to a reaction mixture consisting of the PCR buffer (50 m*M* KCl, 1.5 m*M* $MgCl_2$, 10 m*M* Tris buffer, pH 8.5, 0.001% gelatin), 0.2 m*M* of each of the four deoxynucleoside triphosphates (Amersham Pharmacia Biotech), 1 μ*M* of each of the sense and antisense primers, and 2.5 units of *Taq* polymerase (Amersham Pharmacia Biotech). The initial DNA denaturation for 3 min at 95°C was followed by 35 cycles, each consisting of 1 min denaturation at 95°C, 1 min annealing at the appropriate annealing temperature, and 2 min extension at 72°C; the final extension was for 10 min at 72°C.

Isolation of cDNA 5′- and 3′-Ends by Rapid Amplification of cDNA Ends (RACE). Reverse transcriptase (RT)-PCR was carried out using 5′- and 3′-RACE kits (Gibco BRL, Eggenstein, Germany). One microgram of total RNA prepared from *O. niloticus* jaws or *P. marinus* bodies was used for the first-strand cDNA synthesis in 5′ RACE with *O. niloticus* decorin-specific and *P. marinus* biglycan- and decorin-specific primers; oligo (dT) primers with adapters were used in the 3′ RACE. An intermediate terminal deoxynucleotidyl transferase tailing reaction was carried out prior to amplification in 5′ RACE. The first-strand cDNA products were then amplified by PCR with other inner primers. For zebrafish, a cDNA preparation was made from the whole body of a single fish using the SMART cDNA kit (Clonetech, Palo Alto, CA). Short fragments of biglycan-like and decorin genes were obtained using the primers shown in Table 1.

Table 1. Oligonucleotide primers used in this study

Designation	Sequence	Orientation	Location
Lamprey BGL1			
G2	5′-GAT CTT GTT GTT RTG AAG RTA GAC-3′	A	E8
Dec-L7	5′-ATC GTT GCC ATT GAG GAT GAA GA3′	S	E6
Dec-L13	5′-AGG CTG CAC TGG CAA CCG AAG-3′	A	E2
Dec-L14	5′-AAG ATC GGA GCA CTG CAC CAC A-3′	A	E2
Dec-L16	5′-ATG GCA CAG ACT CGG GGT ATT T-3′	S	E2
Lamprey BGL2			
G2	5′-GAT CTT GTT GTT RTG AAG RTA GAC-3′	A	E8
Big-L1	5′-TCT CCG TGA TCT TGT TGT TCT GC-3′	A	E3
Big-L6	5′-CAG GTA CAG CGC ATA TAG TTT ATT-3′	A	E3←E4
Big-L7	5′-GC AAG AGG ATC TGA TCC GAT A-3′	S	E6
Big-L14	5′-TCG TGC AGT GCT CAG ACC TG-3′	S	E2
Oreochromis BGL3			
AM10	5′-CCT GGG TAT GTG AAC TTC AGT TAT GA-3′	A	E8
Big-F1	5′-TCT CAT ACG GTT GTT GTC CAG GT-3′	A	E7
Big-F2	5′-AGT GTG GGA ATC AAC GAC TTC TG-3′	S	E8
Zebrafish BGL3			
Big-F3	5′-AGT GCT CTG ACC AGG GTC TGA-3′	S	E2→E3
Big-F4	5′-GTC GAC AGG CCA TTA AAA GCT C-3′	A	E5
Zebrafish decorin			
BDG1	3′-CAG TGC TCT GAY CTV GGT CTG-3′	S	E2→E3
G2	5′-GAT CTT GTT GTT RTG AAG RTA GAC-3′	A	E8
Oreochromis decorin			
G2	5′-GAT CTT GTT GTT RTG AAG RTA GAC-3′	A	E8
Dec-F2	5′-GAT CTT GTT GTT RTG AAG RTA GAC-3′	A	E6
Dec-F4	5′-TTG ACA ACA ATG CTC TGA CCA G-3′	S	E8
Dec-F5	5′-GGC ATT TCC TTT AGC AAG TTC TT-3′	A	E4
Dec-F6	5′-CCT TTT TAA TCT TGG TGA TCT CAT T-3′	A	E4

A, antisense; S, sense; E, exon.

Cloning and Sequencing. Twenty microliters of the PCR amplification product were purified by 1.5% low-melting-point agarose (Gibco BRL) gel electrophoresis, and the bands were identified with ethidium bromide staining, excised, and isolated using the QIAEX II gel extraction kit (QIAGEN GmbH, Hilden, Germany). The isolated DNA was blunt-ended, phosphorylated, and ligated to *Sma*I-digested pUC18 plasmid vector with the SureClone ligation kit (Amersham Pharmacia Biotech). The reaction products were transformed into competent *E. coli* XL-1 blue bacteria by standard methods and plated on LB agar containing ampicillin (50 μl/ml) and X-GAL. Transformants were grown overnight in LB-ampicillin broth, and minipreps were prepared according to the standard protocol. Two to five micrograms of DNA were used in the dideoxy sequencing reactions with the AutoRead Sequencing kit (Amersham Pharmacia Biotech). The reactions were processed by the Automated Laser Fluorescent (ALF) sequencer (Amersham Pharmacia Biotech).

Southern Blot Hybridization. Seven micrograms of lamprey (*P. marinus*) genomic DNA were digested with 100 units of the restriction enzymes *Eco*RI, *Hin*dIII, *Bam*HI, and *Msp*I (Boehringer Mannheim) overnight. The recovered DNA was loaded onto 0.8% agarose gel and run overnight. Alkali blots were prepared using the Hybond™ N^+ nylon membrane (Amersham Pharmacia Biotech). Prehybridization, hybridization, and probe labeling were carried out using the AlkPhos DIRECT kit (Amersham Pharmacia Biotech). One hundred nanograms of DNA were used for the labeling of the probe. After the overnight hybridizations, the DNA was washed according to the AlkPhos DIRECT protocol. Following the application of the chemiluminescent detection reagent CDP-Star (from the kit), Hyperfilm™ ECL (Amersham Pharmacia Biotech) was exposed to the blot for 2 h and developed.

Data Analysis. The nucleotide sequences were aligned with the aid of the SeqPup computer program (Gilbert 1995a). Sequence similarities were evaluated with the aid of the DottyPlot computer program (Gilbert 1995b). Substitution rates were estimated by the method of Li et al. (1985) from sequences aligned using the ClustalX program (Thompson et al. 1997). Phylogenetic reconstructions were made using the alignment of the protein segment downstream of the first conserved cysteine cluster. The neighbor joining algorithm of the MEGA program (Kumar et al. 1993) was used to make phylograms, which were bootstrapped to estimate reliability. Distances were estimated from the proportion of amino acid identity following the removal of gapped sites. Parsimony trees were made based on amino acid sequences using the PAUP program (Swofford 1998). A heuristic search method was used with a starting tree obtained by stepwise additions of sequences with one tree held per addition. Ten replications of the addition procedure were made. Optimization was by branch-swapping using tree bisection and reconnection.

Results and Discussion

Characteristics of Biglycan-Like cDNA Sequences

Two *Oreochromis,* two zebrafish, and four lamprey sequences homologous to tetrapod biglycan and decorin cDNA sequences were obtained by screening the corresponding cDNA libraries or cDNA preparations with degenerate primers based on the conserved regions of the

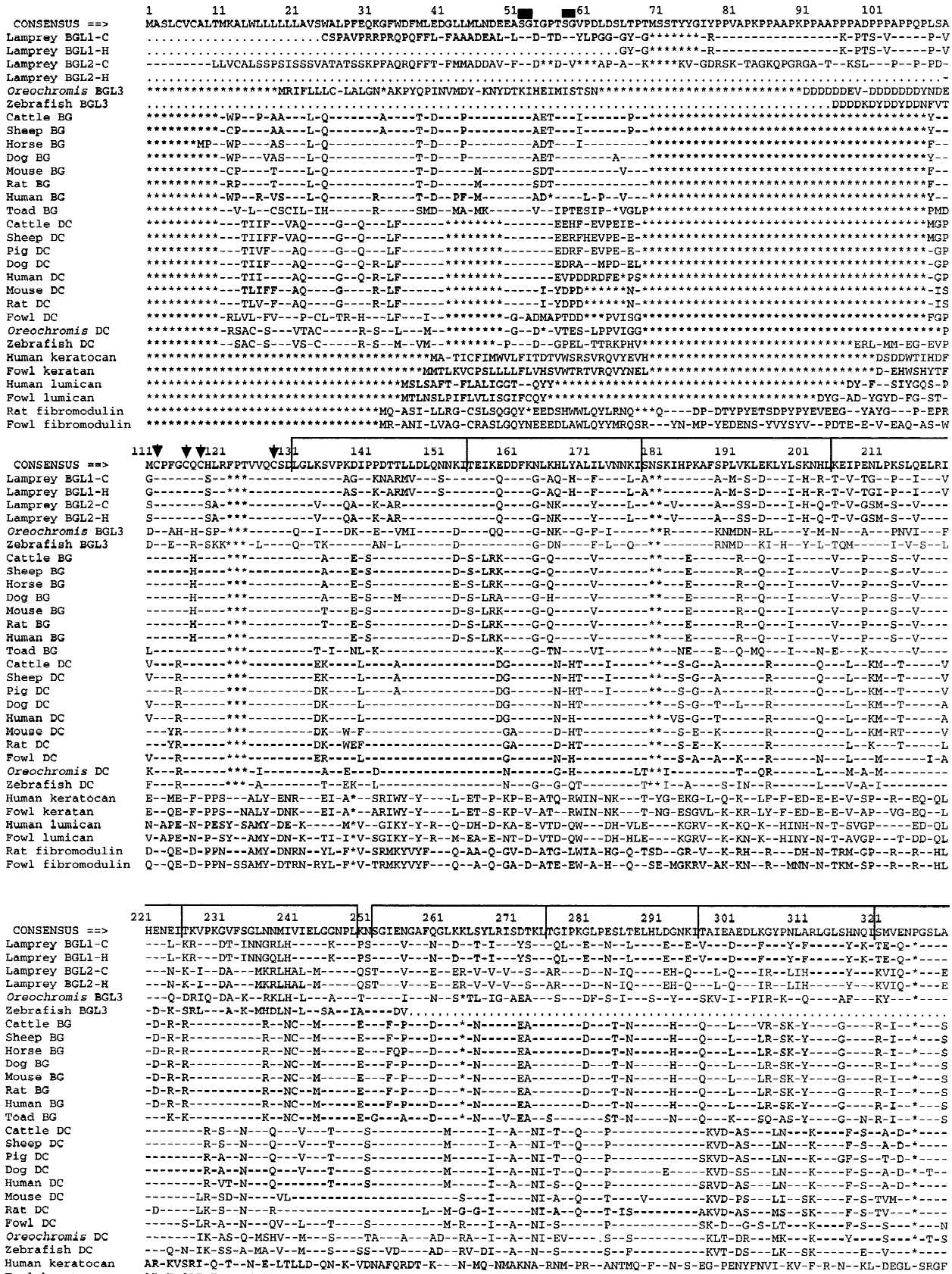

Fig. 1. Legend appears on p. 367.

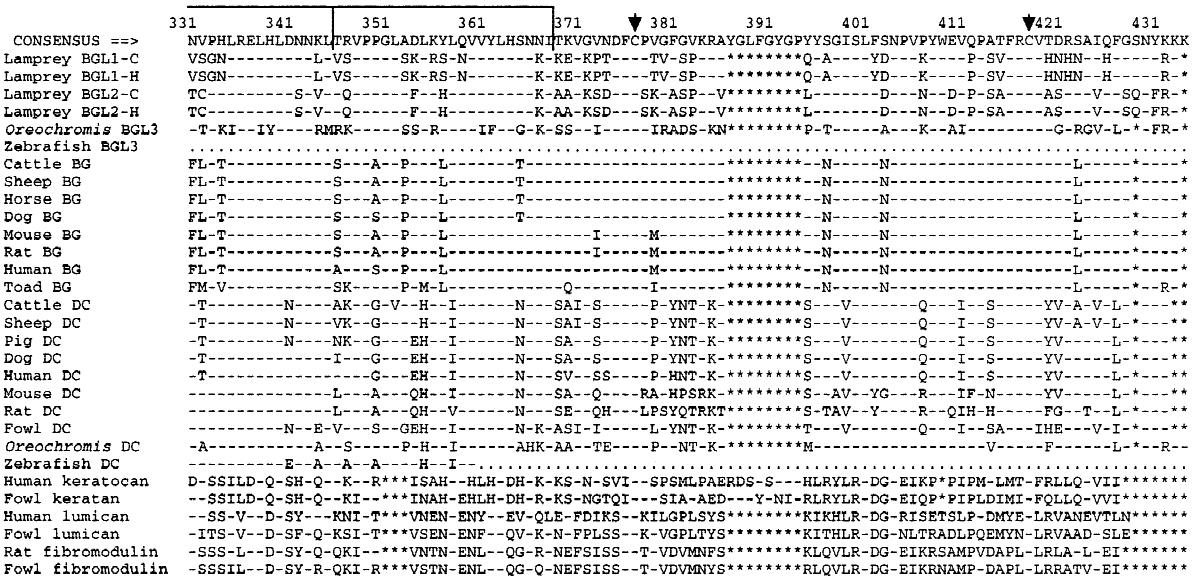

Fig. 1. Amino acid sequence alignment of SLRP family members. The fish and lamprey sequences are from this study; sources for the additional sequences are referenced in the text. The alignment of the initial 110 amino acids was adjusted to bring potential GAG attachment sites at position 52–53 into register. The conserved Ser-Gly residues are indicated with closed rectangles (■). Closed arrows indicate the positions of conserved cysteine residues. The horizontal line with vertical dividers indicates the positions of 10 LRR repeats in the human biglycan (BG) and decorin (DC) sequences, as described by Hocking et al. (1998). A consensus sequence based on simple majority is shown. A dash indicates identity with the consensus, an asterisk (*) the position of an indel, and a dot (.) unavailability of sequence information.

Table 2. Characteristics of cichlid and lamprey proteoglycan cDNA sequences

Species	Proteoglycan	Length			Position of			GenBank accession number
		Total	ORF†	3′ UTR	Translation start site	Stop codon	PAS	
P. marinus	BGL1 C	2240 (P)*	388	1073	?	1165	2207	AF247826
	BGL1 H	2535 (P)	347	1491	?	1042	2518	AF247825
	BGL2 C	3114 (F)	410	1791	91	1321	3068	AF247828
	BGL2 H	2558 (P)	309	1625	?	931	2502	AF247827
O. niloticus	BGL3	1634 (F)	370	136	385	1496	1592	AF247821
	Decorin	2072 (F)	359	865	128	1025	2048	AF247822

* F and P indicate full and partial length cDNA sequences, respectively; the length is given in base pairs (bp)

† Open reading frame given in the number of encoded amino acid residues

C and H indicate the source of *P. marinus* from Lakes Champlain and Huron, respectively; PAS, polyadenylation signals; UTR, untranslated region

tetrapod genes (Table 1). Amino acid sequences translated from nucleotide sequences, which can be accessed in the GenBank database under the numbers AF247821–AF247828, are shown in Fig. 1, together with the known tetrapod sequences. In each case, the initial clone isolated from anchored PCR amplification products provided a partial cDNA sequence. In the case of *Oreochromis* and lampreys, full-length or nearly full-length cDNA sequences were then obtained in three to four steps by anchored PCR using primers based on the partial sequences generated in the preceding step (Table 1). Blast search analyses, sequence comparisons (Fig. 1; Tables 2 and 3), and phylogenetic analyses (Fig. 2) indicate that the eight sequences are very closely related to the biglycan group, which also includes, in addition to biglycan itself, decorin. As discussed in the next section, however, only two of the eight sequences correspond unambiguously to tetrapod decorins and are therefore so designated. The remaining six sequences are designated as biglycan-like (BGL) and distinguished by numbers (see below). Like other members of the biglycan group, the eight sequences have 10 LRRs; potential GAG-attachment sites in the N-terminal part of the protein; and two clusters of cysteine residues flanking the LRR region (Fig. 1). The N-terminal cluster consists of four closely spaced cysteine residues; by contrast, the two cysteine residues in the C-terminal region are separated by 33 amino acid residues. The cysteine residues are invariant in all the thus far identified members of the biglycan group.

Table 3. Amino acid similarities (average % identity) of *Oreochromis* and lamprey proteoglycans with LRR family members

	Lamprey BGL2	*Oreochromis* BGL3	*Oreochromis* DC	Tetrapod BG	Tetrapod DC	Outgroup
Lamprey BGL1	66.7	55.4	55.4	59.6	54.0	30.3
Lamprey BGL2		55.7	57.3	62.1	54.1	28.3
Oreochromis BGL3			53.6	56.3	53.6	27.7
Oreochromis DC				60.4	70.9	28.3
Tetrapod BG					58.9	29.7
Tetrapod DC						29.2

The percentage identity was measured in the conserved portion of the protein, following the omission of the N-terminal sequence up to the conserved cysteine residues. Tetrapod biglycan (BG) includes human, cattle, sheep, horse, dog, mouse, rat, and toad sequences. Tetrapod decorin (DC) includes human, cattle, sheep, pig, dog, mouse, rat, and fowl sequences. The outgroup consists of human keratocan, human and fowl lumican, as well as rat and fowl fibromodulin sequences

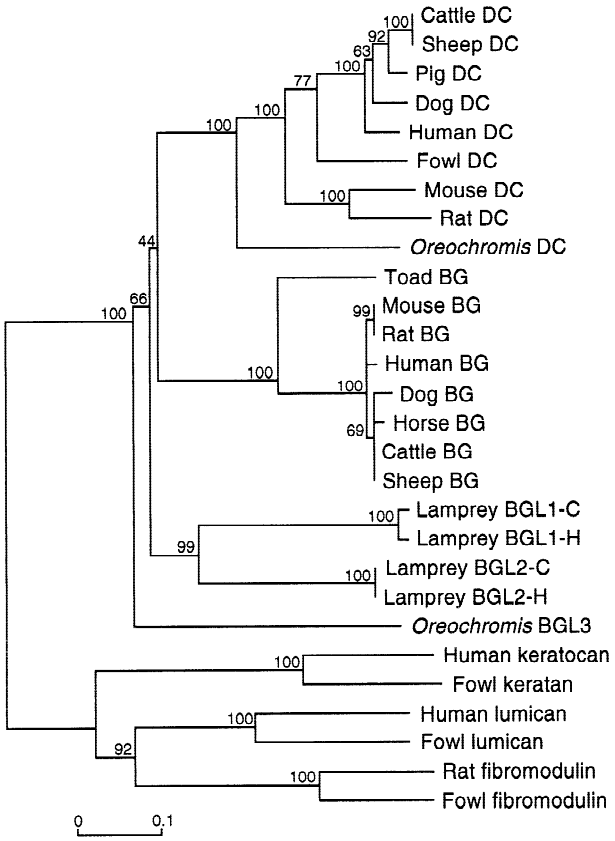

Fig. 2. A neighbor-joining tree of SLRP family protein sequences. The protein alignment shown in Fig. 1 was used. The poorly conserved N-terminal part (initial 110 residues of the alignment) as well as sites containing indels were omitted. Distances were estimated from the proportion of amino acid identity between sequences and the dendrogram drawn using the MEGA program. Numbers on the nodes indicate the percentage recovery of that node in 500 bootstrap replications.

Identities of the Biglycan-Like cDNA Sequences

For convenience, we distinguish the biglycan group (which includes biglycans and decorins of tetrapods, and the BGL sequences of agnathans and teleosts described here) from what we refer to as the "outgroup." The latter consists of the closest known relatives of the biglycan group identified by phylogenetic analysis; it includes human keratocan (Tasheva et al. 1999) and lumican (Grover et al. 1995), fowl keratan (Dunlevy et al. 1998) and lumican (Blochberger et al. 1992), as well as rat (accession code X82152) and fowl (Nurminskaya and Birk 1996) fibromodulin. Furthermore, for the same reason we divide the biglycan coding sequence into an N-terminal part extending from the N-terminus to the first cysteine cluster, the LRR region, and a short C-terminal part. The latter two parts are not difficult to align; the alignment of the LRR regions does not require the introduction of indels at all and that of the C-terminal part requires only one to three indels, depending on the sequence (Fig. 1). There is also no difficulty in extending the alignment of these two parts to the outgroup sequences, because at the amino acid level many residues are invariant not only throughout the biglycan group but also in the outgroup.

In contrast to these two parts, the sequences of the N-terminal part are alignable among tetrapods, but virtually unalignable between agnathans/teleosts and tetrapods (Fig. 1). The part is of different lengths in the various sequences so that long indels must be introduced to obtain alignments based on similarity rather than identity of amino acid residues. The variation in length is concentrated in a segment immediately upstream from the cluster of the four conserved cysteine residues. This segment appears to be absent in the tetrapod sequences, but in the agnathan and teleost sequences it is dominated by a particular amino acid residue that varies from sequence to sequence. Thus, in the *Oreochromis* BGL3, the segment of ~20 residues contains 16 Asp residues, whereas in the lamprey BGL1 and BGL2 sequences, the segment is dominated by proline residues whose number varies among the sequences. Proline-rich segments are also present in some of the outgroup sequences. Outside of this segment, the N-terminal part sequences fall into four classes: the tetrapod biglycans, the tetrapod decorins, the teleost BGL3, and the lamprey BGL1 and BGL2. The outgroup sequences form classes of their own. Within each class, it is possible to align the sequences on the basis of amino acid identity or similarity; alignment between classes is usually equivocal.

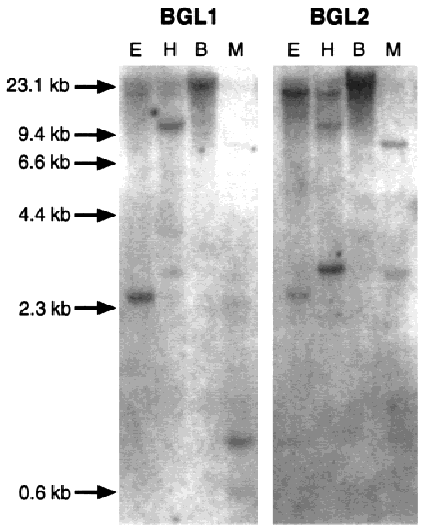

Fig. 3. Southern blot hybridization with the probe of zebrafish *BGL1* and *BGL2*. Zebrafish genomic DNA was digested by *Eco*RI (E), *Hin*dIII (H), *Bam*HI (B), and *Msp*I (M) restriction enzymes.

The canonical GAG attachment site consists of a Ser-Gly pair preceded by acidic amino acid residues. The number of Ser-Gly pairs in the N-terminal part ranges from zero to three. Mammalian biglycans each contain two such pairs, but the toad biglycan contains only one pair, as does the lamprey BGL2. The mammalian decorins each contain one Ser-Gly pair, as does the fowl decorin, whereas the *Oreochromis* decorin contains two pairs. The fowl pair and one of the *Oreochromis* pairs are not preceded by acidic amino acids, however. The lamprey BGL1 contains three Ser-Gly pairs, of which two are preceded by acidic residues, whereas the *Oreochromis* BGL3 contains none, as do some of the outgroup sequences (human keratocan, fowl keratan, rat and fowl fibromodulins). Other outgroup sequences (human and fowl lumicans) contain one pair. The position of the potential GAG-attachment sites is conserved in tetrapod biglycans and decorins (with the exception of the fowl decorin), but not in the outgroup sequences. The number and positions of the GAG-attachment sites cannot, therefore, be taken for a diagnostic feature of the different proteoglycan classes; the position shifts among the sequences of both the biglycan group and the outgroup.

Pairwise comparisons with known sequences reveal that all the new sequences resemble the biglycan group more closely than the outgroup sequences (Table 3) in terms of amino acid similarity. This observation, in combination with the results of the phylogenetic analysis described in the next section (see Fig. 2), establish the membership of the agnathan and teleost sequences in the biglycan group. The comparison furthermore identifies one of the two *Oreochromis* sequences as decorin (~71% similarity with tetrapod decorin) rather than biglycan (~60% similarity with tetrapod biglycans). Here again, the assignment is supported by the phylogenetic analysis (Fig. 2). The second *Oreochromis* sequence (as well as the short zebrafish sequence) has slightly higher sequence similarity with tetrapod biglycans (~56%) than with tetrapod decorins (~54%). On this basis, we have designated it as biglycan-like, BGL3 (but see below). The same is true for the two main types of lamprey sequences (Table 3) and so these, too, are denoted as biglycan-like, specifically BGL1 and BGL2.

Phylogenetic Relationships

In the neighbor-joining tree in Fig. 2, the tetrapod decorins form one clade, which is joined by the *Oreochromis* decorin sequence, and the tetrapod biglycans form a second major clade. The *Oreochromis* BGL3 and the lamprey BGL1 and BGL2 sequences lie outside of these two clades. The *Oreochromis* BGL3 sequence is inexplicably found in a sister-group position to this entire cluster when the tree is rooted by the outgroup sequences. A tree drawn by the maximum parsimony method has a similar topology to the neighbor-joining tree (data not shown). The two issues that arise from this phylogenetic reconstruction are, first, the relationship of the BGL3 sequence of the biglycan group of sequences, and second, the relationship of the lamprey sequences to the biglycan and decorin clades.

There are at least three explanations that may account for the outer position of the BGL3 sequence. The first explanation is that BGL3 is really a biglycan that, however, evolves faster than the tetrapod biglycans and that the accelerated evolutionary rate is responsible for its anomalous position on the tree. Affinity of BGL3 to tetrapod biglycan sequences is suggested not only by its overall closer sequence similarity with biglycans but also by shared conserved sites. In the entire BGL3 sequence there are 34 substitutions and 1 single-codon indel that BGL3 shares with the tetrapod biglycans, compared to 29 substitutions it shares with tetrapod decorins (Fig. 1). The second possibility is that BGL3 is a new member of the biglycan group that is neither biglycan itself nor decorin. If this were the case, the question would then have to be answered regarding the fate of the biglycan gene in teleosts. A search for a more orthodox version of cichilid fish biglycan failed to produce any candidates. When the search was extended to the zebrafish, only an apparent ortholog of the *Oreochromis* BGL3 sequence was found (Fig. 1). This result would then seem to suggest that either the “true” biglycan has been lost in at least some teleosts or that biglycans and decorins diverged after the divergence of teleosts and tetrapods. The latter possibility is, however, contradicted by the presence of a bona fide decorin in *Oreochromis.* The third possibility is that BGL3 is encoded in a pseudogene and so evolves anomalously from the presumably functional biglycans. The only observation that might possibly be interpreted as favoring this possibility is the apparent absence of a GAG-attachment site in BGL3, but we note

A

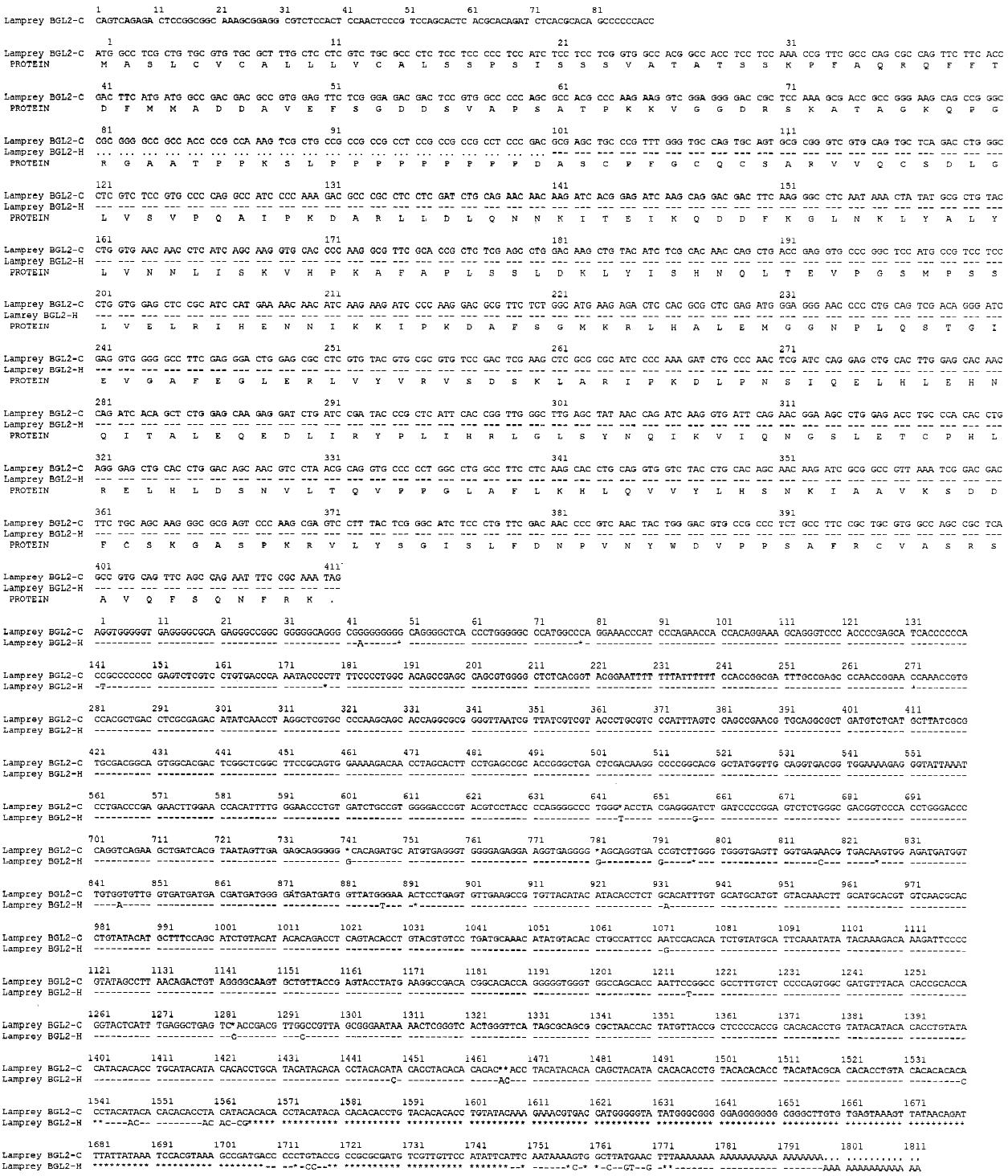

Fig. 4. Alignment of **(A)** *BGL1* and **(B)** *BGL2* nucleotide sequences obtained from *P. marinus* species from Lakes Huron and Champlain. A dash (–) indicates identity with the upper sequence, an asterisk (*) an alignment gap, and a dot (.) unavailability of sequence information. Coding regions are presented and numbered by codon beginning at the first nucleotide. 5′ and 3′ noncoding regions are numbered separately in blocks of 10 nucleotides. A translation in the single letter amino acid code is given below the sequences.

that there are other, apparently functional proteoglycan genes that similarly lack this site. The contraindications to the pseudogene status of *BGL3* are, first, the fact that the gene is expressed in both *Oreochromis* and zebrafish and has no obvious inactivating defects, and second, that the ratio of synonymous to nonsynonymous substitutions in comparisons of zebrafish and *Oreochromis BGL3* is at least 4. (Synonymous substitutions are saturated, giving Ks values of >2, while nonsynonymous substitutions give a Ka value of 0.24.) The fact that synonymous sub-

B

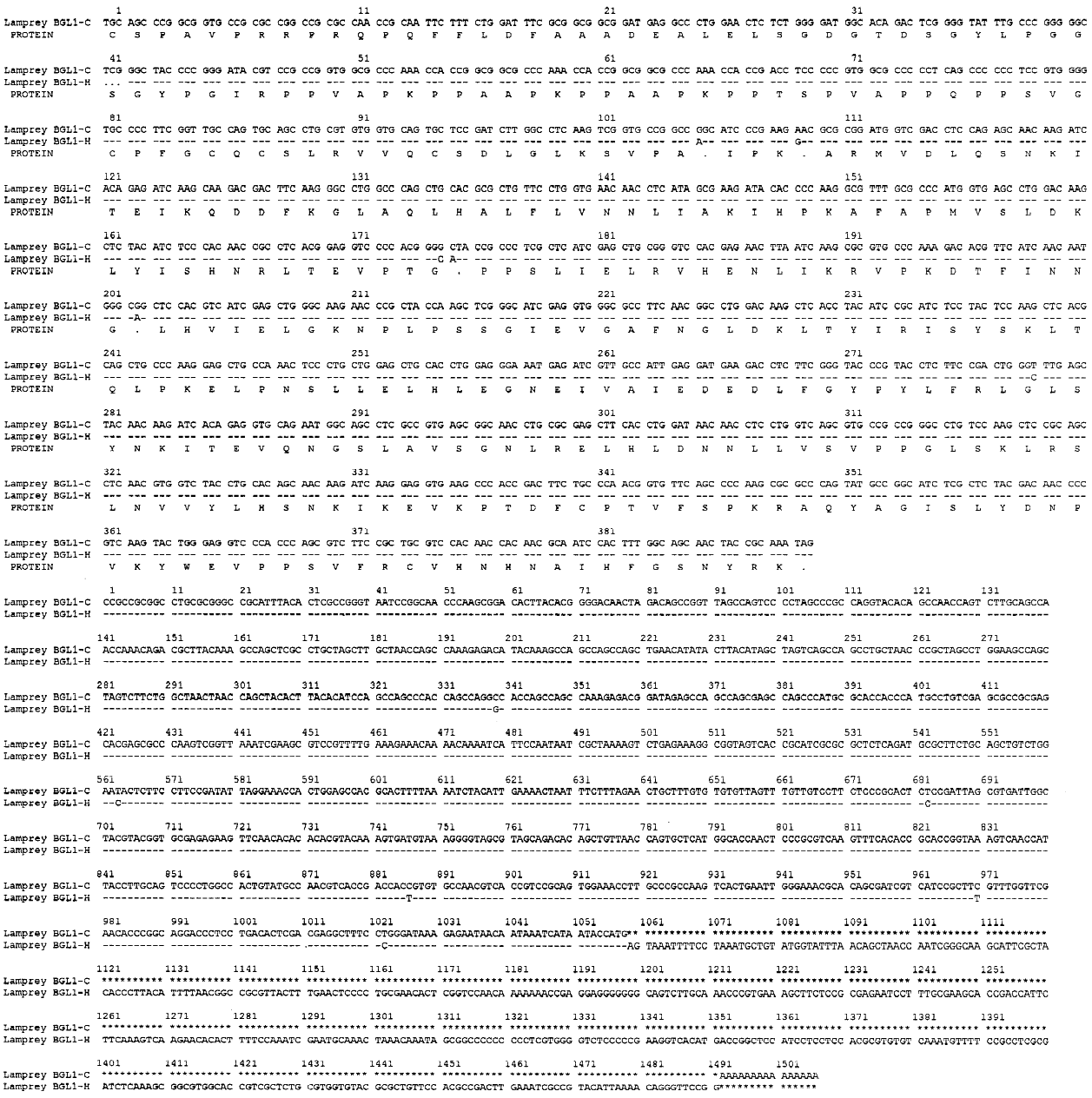

Fig. 4 Continued.

stitutions predominate indicates that the gene has been functional for much if not all of the time since the divergence of zebrafish and *Oreochromis*. Taken together, the most likely of the three possibilities is, at present, the first one.

As for the relationship of the lamprey *BGL1* and *BGL2* sequences to biglycans and decorins, we note that all efforts to find a decorin-like gene in the lamprey have failed. To determine whether any other biglycan- or decorin-related genes might be present in the lamprey genome, a Southern blot was made under conditions of reduced stringency. The probes used were >300-nucleotide-long fragments from conserved parts of either the *BGL1* or *BGL2* genes (Fig. 1; positions 125–242 for *BGL1* and 125–237 for *BGL2*). The Southern blot (Fig. 3) shows between one (*Bam*HI digestion) and five (*Msp*I digestion) hybridizing bands of varying intensity to be present in the lamprey genome, regardless of which probe was used. An internal *Msp*I site is found in both *BGL1* and *BGL2* probes, producing the weaker additional bands seen in the *Msp*I digestion. For both probes with any of the enzymes, the hybridization pattern of bands and their positions are strikingly similar—the major difference appears to be the strength of the hybridization signal rather than the band positions. The relative strength of the bands is also antithetical—bands that hy-

bridize strongly to the *BGL1* probe appear to do so weakly to the *BGL2* probe and vice versa. We interpret this to indicate that each probe hybridizes to both *BGL1* and *BGL2* under the conditions used and further that any similarly related genes should give a hybridization signal. The absence of any bands additional to *BGL1* and *BGL2* indicates that additional loci do not exist in the lamprey or, alternatively, that any such loci are very dissimilar to *BGL1* and *BGL2*. Using the enzyme *Bam*HI (Fig. 3), only one strongly hybridizing band of large size appears to be present in the digest. One possible explanation of this is that the *BGL1* and *BGL2* loci are linked on a single digestion fragment.

At present, therefore, it seems reasonable to conclude that the gene duplication leading to the emergence of biglycans and decorins occurred after the divergence of Agnatha and Gnathostomata and that *BGL1* and *BGL2* are more closely related to the ancestor of biglycan and decorin than to either of the two. The slightly higher similarity of lamprey *BGL* sequences to biglycan than to decorin might be the result of *BGL1* and *BGL2* having biglycan-like (rather than decorin-like) functions. As for the origin of *BGL1* and *BGL2*, it seems that these two genes are the result of an independent duplication that occurred in the agnathan (lamprey) lineage after its divergence from gnathostomes.

Sequence Variation Within Lampreys

Sequences of both *BGL1* and *BGL2* determined from *P. marinus* species collected in Lakes Champlain and Huron were not identical. Figure 4 shows nucleotide alignments of sequences obtained from both sources. In the case of *BGL1*, two synonymous and four nonsynonymous differences are found in the 348 codons of shared coding sequence. A further six differences occur in more than 1 kb of 3′ untranslated (UT) sequence. This corresponds to an overall nucleotide distance of 0.0057 substitutions per site. Furthermore, the polyadenylation site appears to differ between the sequences. In the case of *BGL2* no differences are found in the 310 shared codons. The 3′ UT sequence of *BGL2* appears to contain an extended segment of CA repeats, as well as runs of single-nucleotide repeats. Polyadenylation also appears to occur in different positions in sequences obtained from Lakes Champlain and Huron, and this makes alignment of the 3′ end difficult. In the approximately 1.5 kb of clearly alignable 3′ UT sequence, 12 substitutions as well as many indels are found. This corresponds to an overall distance of 0.0049 substitutions per site, similar to that found for *BGL1*. Although little is known about rates of molecular evolution in lampreys, the degree of difference in sequence between lamprey from Lakes Champlain and Huron is surprisingly high and presumably has taken many years to accumulate. The population from Lake Huron is thought to have been introduced in the 1920s from an unknown source, whereas the Lake Champlain population has probably existed since the last glaciation. The specimens may represent sibling species of *P. marinus*. However, more genes will need to be characterized to determine the true degree of differentiation between the populations.

Interpretation of Biglycan-Decorin Evolution

On the strength of the available evidence it seems reasonable to suggest that agnathans, as represented by the two *Petromyzon* species used in the present study, possess small proteoglycans that may not have as yet differentiated functionally into biglycan and decorin. Instead, the gene encoding the ancestral proteoglycan independently duplicated into two copies, *BGL1* and *BGL2*, which appear to have remained functionally undifferentiated. The duplication of the ancestral gene that ultimately led to biglycan and decorin appears to have taken place after the divergence of agnathans from gnathostomes but before the divergence of teleostei (or at the latest before the divergence of Perciformes here represented by the *Oreochromis* cichid and Cypriniformes represented by the zebrafish). Teleosts have two presumably functionally differentiated proteoglycans, one closely related to the tetrapod decorin and the other (BGL3) allied with the BGLs of the lamprey. The acquisition of the decorin function by one of the duplicated genes was accompanied by a slowdown of its evolutionary rate with the result that this gene can now be identified unambiguously as decorin. The second of the duplicated teleost genes may have kept its original function and a relatively rapid evolutionary rate until the emergence of tetrapods. This supposition may explain why the teleost *BGL3* appears to have diverged earlier than it in fact apparently did. In tetrapods, the *BGL3* acquired a biglycan function and slowed down its evolutionary rate. This slowdown was more dramatic than that of the decorin gene, with the result that the branches in the biglycan clade are shorter than in the decorin clade, especially in mammals.

Supporting this view, the evidence from mammalian sequences suggests that although the overall evolutionary rates are now slow, there is some rate variation. The rat and mouse biglycan have only two residue differences over their entire length, and the sheep-cattle biglycan only three differences. Mammalian decorins appear to evolve more rapidly than mammalian biglycans. The sheep and cattle decorin sequences have five differences, whereas the rat and mouse sequences differ in 44 residues. The mouse and rat decorin sequences appear quite distinctive and are almost as distant as fowl from other mammals even in their conserved regions. The implication is that the evolutionary rate for the sequences can change, perhaps as a consequence of altered functions. In this view, the changing evolutionary rates can

explain the complex relationships found in the gene family in gnathostomes.

References

Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD (1989) Molecular biology of the cell. New York: Garland Publishing

Blochberger TC, Vergen JP, Hempel J, Hassel JR (1992) cDNA to chick lumican (corneal keratan sulfate proteoglycan) reveals homology to the small interstitial proteoglycan gene family and expression in muscle and intestine. J Biol Chem 267:347–352

Bourdon MA, Krusius T, Campbell S, Schwartz NB, Ruoslahti E (1987) Identification and synthesis of a recognition signal for the attachment of glycosaminoglycans to proteins. Proc Natl Acad Sci USA 84:3194–3198

Boykiw RH, Sciore P, Reno C, Marchuk L, Frank C, Hart DA (1998) Altered levels of extracellular matrix molecules mRNA in healing rabbit ligaments. Matrix Biol 171:31–378

Day AA, Ramis CI, Fisher LW, Gehron-Robey P, Termine JD, Young MF (1986) Characterization of bone PG II cDNA and its relationship to PG II mRNA from other connective tissues. Nucleic Acids Res 14:9861–9876

Day AA, McQuillan CI, Termine JD, Young MR (1987) Molecular cloning and sequence analysis of the cDNA for small proteoglycan II of bovine bone. Biochem J 248:801–805

Dreher KL, Asundi V, Matzura D, Cowan K (1990) Vascular smooth muscle biglycan represents a highly conserved proteoglycan within the arterial wall. Eur J Cell Biol 53:296–304

Dunlevy JR, Neame PJ, Vergnes JP, Hassell JR (1998) Identification of the N-linked oligosaccharide sites in chick corneal lumican and keratocan that receive keratan sulfate. J Biol Chem 273:9615–9621

Fisher LW, Termine JD, Dejter Jr SW, Whitson SW, Yanagishita M, Kimura JH, Hascall VC, Kleinman HK, Hassell JR, Nilsson B (1983) Proteoglycans of developing bone. J Biol Chem 258:6588–6594

Fisher LW, Termine JD, Young MF (1989) Deduced protein sequence of bone small proteoglycan I (biglycan) shows homology with proteoglycan II (decorin) and several nonconnective tissue proteins in a variety of species. J Biol Chem 264:4571–4576

Fisher LW, Heegaard A-M, Vetter U, Vogel W, Just W, Termine JD, Young MF (1991) Human biglycan gene. Putative promoter, intron-exon junctions, and chromosomal localization. J Biol Chem 266:14371–14377

Gehron Robey P (1996) Vertebrate mineralized matrix proteins: Structure and function. Connect Tissue Res 35:131–136

Gerhart J, Kirschner M (1997) Cells, embryos, and evolution. Toward a cellular and developmental understanding of phenotypic variation and evolutionary adaptability. Malden, MA: Blackwell Science Publications

Gilbert DG (1995a) SeqPup version 0.4j: a biosequence editor and analysis application. http://iubio.bio.indiana.edu/soft/molbio

Gilbert DG (1995b) DottyPlot, version 1.0c. http://iubio.bio.indiana.edu/soft/molbio

Grover J, Chen XN, Korenberg JR, Roughly PJ (1995) The human lumican gene. Organization, chromosomal location, and expression in articular cartilage. J Biol Chem 270:21942–21949

Hocking AM, Shinomura T, McQuillan DJ (1998) Leucine-rich repeat glycoproteins of the extracellular matrix. Matrix Biol 17:1–19

Iozzo RV, Murdoch AD (1996) Proteoglycans of the extracellular environment: clues from the gene and protein side offer novel perspectives in molecular diversity and function. FASEB J 10:598–614

Kobe B, Deisenhofer J (1994) The leucine-rich repeat: a versatile binding motif. Trends Biochem Sci 19:415–421

Kokenyesi R, Woessner JFJ (1989) Purification and characterization of a small dermatan sulphate proteoglycan implicated in the dilatation of the rat uterine. Biochem J 260:413–419

Kresse H, Hausser H, Schönherr E (1994) Small proteoglycans. In: Jollès P (ed) Proteoglycans. Basel, Switzerland: Birkäuser Verlag, pp 73–100

Krusius T, Ruoslahti E (1986) Primary structure of an extracellular matrix proteoglycan core protein deduced from cloned cDNA. Proc Natl Acad Sci USA 83:7683–7687

Kumar S, Tamura K, Nei M (1993) MEGA: Molecular evolutionary genetic analysis, version 1.01. University Park, PA: The Pennsylvania State University

Li W-H, Wu C-I, Luo C-C (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide and codon changes. Mol Biol Evol 2:150–174

Li W, Vergnes J-P, Cornuet PK, Hassell JR (1992) cDNA clone to chick corneal chondroitin/dermatan sulfate proteoglycan reveals identity to decorin. Arch Biochem Biophys 296:190–197

Marcum JA, Torok M, Evans S (1993) cDNA sequence for bovine biglycan (PGI) protein core. Biochim Biophys Acta 1173:81–84

Mayer WE, Tichy H (1995) A cDNA clone from the sea lamprey *Petromyzon marinus* coding for a scavenger receptor Cys-rich (SRCR) domain protein. Gene 164:267–271

Neame PJ, Choi HU, Rosenberg LC (1989) The primary structure of the core protein of the small, leucine-rich proteoglycan (PGI) from bovine articular cartilage. J Biol Chem 264:8653–8661

Nurminskaya MV, Birk DE (1996) Differential expression of fibromodulin mRNA associated with tendon fibril growth: isolation and characterization of a chicken fibromodulin cDNA. Biochem J 317: 785–789

Pulkkinen L, Alitalo T, Krusius T, Peltonen L (1992) Expression of decorin in human tissues and cell lines and defined chromosomal assignment of the gene locus (DCN). Cytogenet Cell Genet 60: 107–111

Scholzen T, Solursh M, Suzuki S, Reiter R, Morgan JL, Buchberg AM, Siracusa LD, Iozzo RV (1994) The murine decorin. Complete cDNA cloning, genomic organization, chromosomal assignment, and expression during organogenesis and tissue differentiation. J Biol Chem 269:28270–28281

Swofford DL (1998) PAUP: Phylogenetic analysis using parsimony (and other methods), Version 4. Sunderland, MA: Sinauer Associates

Tasheva ES, Funderburgh JL, Funderburgh ML, Corpuz LM, Conrad GW (1999) Structure and sequence of the gene encoding human keratocan. DNA Seq 10:67–74

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) the ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 24:4876–4882

Toyosawa S, O'hUigin C, Figueroa F, Tichy H, Klein J (1998) Identification and characterization of amelogenin genes in monotremes, reptiles, and amphibians. Proc Natl Acad Sci USA 95:13056–13061

Toyosawa S, O'hUigin C, Klein J (1999a) Dentin matrix protein 1 gene of prototherian and metatherian mammals. J Mol Evol 48:160–167

Toyosawa S, O'hUigin C, Tichy H, Klein J (1999b) Characterization of dentin matrix protein 1 gene in crocodilia: implications for the evolution of mineralized tissues. Gene 234:301–314

Toyosawa S, Sato A, O'hUigin C, Tichy H, Klein J (2000) Expression of the dentin matrix protein 1 gene in birds. J Mol Evol 50:31–38

Traupe H, van den Ouweland AMW, van Oost BA, Vogel W, Vetter U, Warren ST, Rocchi M, Darlison MG, Ropers H-H (1992) Fine mapping of the human biglycan (BGN) gene within the Xq28 region employing a hybrid cell panel. Genomics 13:481–483

Vetter U, Vogel W, Just W, Young MF, Fisher LW (1993) Human decorin gene: intron-exon junctions and chromosomal localization. Genomics 15:161–168

Wegrowski Y, Pillarisetti J, Danielson KG, Suzuki S, Iozzo RV (1995) The murine biglycan: complete cDNA cloning, genomic organization, promoter function, and expression. Genomics 30:8–17