Lamprin is a unique structural protein which forms the extracellular matrix of several cartilaginous structures found in the lamprey. Lamprin is noncollagenous in nature but shows sequence similarities to elastins and to insect structural proteins. Here, we characterize the structure and organization of lamprin genes, demonstrating the presence of multiple similar but not identical copies of the lamprin gene in the genome of the lamprey. In at least one species of lamprey, *Lampetra richardsoni*, the multiple gene copies are arranged in tandem in the genome in a head-to-tail orientation. Lamprin genes from *Petromyzon marinus* contain either seven or eight exons, with exon 4 being alternatively spliced in all genes, resulting in a total of six different lamprin transcripts. All exon junctions are of class 1.1. An unusual feature of the lamprin gene structure is the distribution of the 3’ untranslated region sequence among multiple exons. A TATA box and cap sequence have been identified in upstream sequences in close proximity to the transcription start site, but no CAAT box could be identified. Sequence and gene structure comparisons between lamprins, elastins, and insect structural proteins suggest that the regions of sequence similarity are the result of a process of convergent evolution.

### Introduction

The last common ancestor shared by the lamprey, one of the two living jawless vertebrates, and all other extant vertebrate species existed over 500 MYA at the threshold of vertebrate evolution (Morris 1997). As a result, present-day lamprey species have several unique features, one of which is the nature of their endoskeleton. All components of the lamprey endoskeleton are cartilaginous (Parker 1883; Hardisty 1981). Although the lamprey is not the only vertebrate with an absence of bony tissues, the nature of the cartilage found in the lamprey is unique. Whereas the major matrix protein of all other vertebrate cartilages, including those of the sharks and rays, is type II collagen (Moss 1977), cartilage matrix proteins in lampreys are noncollagenous in nature (Wright, Keeley, and Youson 1983; Wright and Youson 1983; Robson et al. 1997).

Two types of noncollagenous cartilage have been described in the lamprey, designated cranial or branchial according to the tissues in which they were first recognized (Robson et al. 1997). While little sequence information is available for the cartilaginous proteins found in branchial cartilage structures, several cDNAs of lamprin, the predominant matrix protein of cranial cartilage, have been cloned (Robson et al. 1993). While it is very different from collagen, lamprin is highly hydrophobic in nature and shares biochemical characteristics with vertebrate elastins (Wright, Keeley, and Youson 1983; Wright and Youson 1983). Furthermore, lamprin cDNAs revealed tandemly repeated sequences with similarities to repeat sequences found not only in mammalian and avian elastins, but also in structural proteins of some invertebrates, including insect chiron proteins and spider silks (Robson et al. 1993). For example, a tandem repeat region of a silkmother chiron class B protein shares a 21/24 amino acid identity with lamprin (Robson et al. 1993). Even more striking is a 28/30 amino acid identity shared between oothecin, an eggshell protein of the cockroach, and the tandem repeat sequence of lamprin (Pau, Brunet, and Williams 1971; Pau 1984). While such sequence similarities might suggest descent from a common ancestral protein, there are several difficulties with arguments based on sequence conservation. For example, outside these regions of identity, the sequences of the proteins show essentially no other similarity. Although these regions of identity could reflect a common exon that has been shuffled between genes over time, a second explanation for the appearance of such isolated sequence identity is a mechanism dependent on sequence convergence.

Sequence convergence, defined as derived sequence similarity between proteins of unrelated origin, is thought to be a rare process, especially for soluble globular proteins (Doolittle 1994). Structural convergence apparently does occur, but similar three-dimensional structures can be achieved by sequences that do not share much linear identity (Russell et al. 1997). Thus, physical contacts in three dimensions generated by the sequence are more important than the precise linear array of amino acids in determining structure. For this reason, the probability of primary sequence convergence in soluble globular proteins is very low.

On the other hand, the likelihood of sequence convergence in simpler, nonglobular proteins may be somewhat greater. For example, a strong case for sequence convergence has been established for two antifreeze gly-
coproteins, where a tripeptide repeat important for inhibiting the growth of an ice crystal lattice has independently appeared in two unrelated groups, or taxa, of fishes, one from the Arctic and the other from the Antarctic (Chen, DeVries, and Cheng 1997). Lamprin and the proteins with which it shares sequence similarity also appear to be relatively simple structurally. All are predominantly hydrophobic in nature and have the ability to spontaneously self-aggregate into stable polymeric matrices (Hamodrakas et al. 1983; Bressan et al. 1986; Robson et al. 1993). Such self-aggregation has been suggested to be based on interdigitation of hydrophobic side chains in short β-sheet/β-turn structures (Marsh, Corey, and Pauling 1955; Hamodrakas et al. 1982, 1983; Robson et al. 1993).

To understand the origins of lamprin, a novel protein found only in the lamprey, and to address the question of sequence convergence in these proteins, we determined the structure and organization of the genes encoding lamprin in two distinct species of lamprey, *Petromyzon marinus* and *Lampetra richardsoni*. We identified the presence of multiple alternately spliced genes which, in at least one of these species, appear to be located in tandem in a head-to-tail orientation in the genome. The 3′ untranslated regions (UTRs) of lamprin transcripts are highly unusual in that they are generated from more than one exon. Characterization of the structures and organization of these lamprin genes provides further evidence that similarities between lamprin and insect structural proteins are a result of sequence convergence.

**Materials and Methods**

*Genomic DNA Isolation*

*Petromyzon marinus* genomic DNA was isolated from annular cartilage of a single adult lamprey captured during its upstream (spawning) migration in Duffs Creek, Ontario. Genomic DNA from *L. richardsoni* was isolated from the kidneys of four animals taken from Hatzic Slough, British Columbia. After dissection, tissue was immediately frozen in liquid nitrogen and stored at −70°C. DNA was purified following the method in *Current Protocols in Molecular Biology* (Ausubel et al. 1998, pp. 2.2.1–2.2.3).

**Southern Blotting For Determination of Gene Copy Number in *P. marinus***

After overnight digestion of *P. marinus* genomic DNA individually with *Hind*III, *Sal*I, *Eco*RI, and EcoRV, restriction fragments were electrophoretically separated (0.6% agarose), transferred, and then UV cross-linked to Hybond-N nylon membrane (Amersham, Ontario, Canada). Membranes were incubated at 67°C for 9 h in prehybridization buffer (6 × SSPE, 5 × Denhardt’s solution, 0.5% SDS, 100 μg/ml denatured herring sperm DNA). This buffer was then replaced by hybridization buffer (6 × SSPE, 0.5% SDS, 100 μg/ml denatured herring sperm DNA) with the appropriate α-32P dCTP random prime-labeled probe, and incubation continued overnight at 67°C. Membranes were washed in 2 × SSC, 0.5% SDS for 5 min at 25°C, followed by 2 × SSC, 0.1% SDS for 15 min at 25°C and 0.1 × SSC, 0.5% SDS for 30 min at 37°C, with a final stringent wash in 0.1 × SSC, 0.5% SDS for 30 min at 68°C. Hybridized DNA was detected by autoradiography.

**Amplification of Introns of Lamprin Genes from *P. marinus***

Individual lamprin genes in *P. marinus* were amplified from gDNA using gene-specific primers in the first and last exons of each gene. The primers were L-1.8-specific forward (5′-CGACACCGAAGACGAAACAAAAATCCCTCC-3′), L-1.8-specific reverse (5′-TTTGGGTGGAATGTGAAAGCGGGAGG-3′), L-0.9-L-0.8 forward (5′-AGCCCTCCTCCTCCAGTGTTCG-3′), L-0.8-specific reverse (5′-CACATGCAGATGATTGGAGGAATTTG-3′), and L-0.9-specific reverse (5′-CAGGAACTGAAACGAGCGCAGGAATAC-3′). The primary amplification reactions consisted of Advantage Taq polymerase mix (Clontech, California), 1 × amplification buffer supplied with the enzyme, 1.1 mM Mg(OAc)₂, 200 μM dNTPs, 200 nM of each primer, and 0.1 μg of gDNA (isolated from a single animal) in a total reaction volume of 100 μl. The cycling parameters for amplification were as follows: denaturation at 94°C for 30 s, annealing at 65°C for 30 s, and extension at 68°C for 8 min for 30 cycles. After electrophoresis, amplification products were purified from agarose gel slices (Genesclean, BIO 101, California) and used as templates for intron-specific amplification reactions.

To determine the location of specific exon/intron junctions and intron sizes, the full-length amplified genes were used as templates at a 1/2,000 dilution in combination with sets of primers designed for neighboring exons. The primers used in these reactions were designed to contain either *Bam*HI or *Eco*RI restriction sites to facilitate cloning of the amplification products into pBluescript plasmid vector (Stratagene, California). Cloned products were sequenced at the HSC Biotechnology Service Centre (Toronto, Ontario, Canada).

**Determination of Alternative Splicing of Lamprin Genes from *P. marinus***

Two primer sets were designed for each of the two *P. marinus* lamprin genes (L-1.8 and L-0.8). Both primer sets for L-1.8 used a reverse primer designed for exon 7 (5′-CGCAATTCGATCGCTAGACATTTCTCCAGTCA-3′; italicized sequence represents restriction site and clamp sequence), an exon found only in this gene. This primer was used in combination with two different forward primers, one to exon 3 (5′-GGGAGATCTCTGGTGGGTGGTCCAC-3′); the other to exons 4–7 (5′-GGGATCCACCCCTTACAGGGTGGATTGATATAC-3′). For L-0.8, the same two forward primers were used, but in combination with the L-0.8-specific reverse primer (described in Amplification of Introns, above), which anneals to exon 8. The reaction mixture contained 0.25 U of Taq polymerase (Life Technologies, Maryland), 1 × PCR buffer supplied with the enzyme, 3 mM MgCl₂, 200 μM dNTPs, 0.5 μM of each primer, and 1 μl of an annular
cartilage cDNA library (Robson et al. 1993) in a 50-μl reaction volume. The cycling parameters were as follows: denaturation at 94°C for 45 s, annealing at 58°C for 45 s, and extension at 72°C for 1 min for 34 cycles. Products were analyzed on a 1.5% agarose gel, subcloned into pBluescript, and sequenced.

**Genome Walking for 5’-Upstream Sequence of Lamprin Genes from *P. marinus***

The Universal GenomeWalker kit (Clontech) was used to amplify 5’-upstream sequences. Briefly, *P. marinus* genomic DNA was digested to completion separately with *Dra*I, *EcoRV*, *PvuII*, *Sal*I, and *Stu*I. Adapters were ligated to the ends of the restriction fragments, and these “libraries” were used as templates to amplify the 5’-upstream sequences. For the primary amplification reactions, reverse primers specific to exon 1 of either L-1.8 (5’-GCTTGCA-TGGTGGCGGCTATTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
Fig. 1.—Lamprin gene copy number in *Petromyzon marinus*. Southern blots of lamprey *P. marinus* genomic DNA digested with *Hind*III (H), *Bam*HI (B), *Sal*I (S), *Eco*RI (EI), and *Eco*RV (EV). Blot A was probed with the full-length intron 1 sequence of L-1.8, blot B with ~600 bp of intron 1 sequence from L-0.9, and blot C with the full-length intron 6 sequence of L-0.8. A single band in each lane of blot A indicates a single copy of the L-1.8 gene. Two or three bands per lane in blots B and C indicate multiple copies of the L-0.8/L-0.9 genes. Large single bands in the *Bam*HI-digested lanes in blots A and B and the *Sal*I-digested lane in blot C are likely due to incomplete digestion of the genomic DNA. Positions of molecular size markers (kb) are indicated.

To distinguish between the two or more copies of L-0.9, we focused on the normally less well conserved 3′ UTR regions to identify differences in sequence between the respective mRNAs. The full-length 3′ UTR of L-0.9 was amplified from genomic DNA using primers that would not recognize L-1.8. Two bands resulted, one of the size expected if there were no introns present between the primers, and a second, smaller, band. The smaller band was not the result of a spurious amplification product from an L-0.9 template, as an L-0.9 cDNA yielded only the larger product. Sequencing of the smaller band demonstrated that it was very similar to the 3′ UTR of the L-0.9-12 transcript, except for a 184 bp deletion (from position 693 to position 877 in GenBank accession L05925). The similarity between these two products was much greater than that to the corresponding region of L-1.8 (97.2% vs. 81.9% identity). These results therefore indicated the presence of a third gene for lamprin, which was very similar, but not identical, to L-0.9, differing mainly in the length of its 3′ UTR. This third gene was thereafter designated L-0.8. Sequence differences between L-0.9 and L-0.8 3′ UTRs were utilized to design specific primers for each, which were then used to amplify intronic regions (see Exon/Intron Structure, below).

Sequence comparisons between intron 1 of L-0.9 and intron 1 of L-0.8 indicated 96.5% identity over the
first 184 bp and 95.2% identity over the last 166 bp, much greater identities than to the corresponding regions of intron 1 of L-1.8 (see above). With this degree of sequence identity, cross-hybridization would be expected even under the stringent washing conditions used. Therefore, analysis of the Southern blot in figure 1B suggests that the darker band in the HindIII lane corresponded to the L-0.9 gene, and the less strongly hybridizing band represented the L-0.8 gene. The third band in the EcoRV lane is the result of an EcoRV site within intron 1 of L-0.8, 581 bp from the 3′ end. A similar Southern blot using intron 6 of the L-0.8 gene as a probe showed two bands in both the HindIII and EcoRI lanes of the Southern blot, although with differing intensities (fig. 1C). The same two bands were seen using intron 6 from L-0.9 as a probe, but with their respective intensities reversed (data not shown). There is 93% identity between introns 6 of L-0.8 and L-0.9, and there is 77.5% identity between introns 6 of L-0.8 and L-1.8. Neither L-0.8 intron 6 nor L-0.9 intron 6 contains internal HindIII or EcoRI sites. Therefore, the results of the Southern blots suggested that these probes were each recognizing both L-0.9 and L-0.8 genes, but each was hybridizing more strongly to itself than to the other gene. The lack of a third band using these probes argues against an additional L-0.9-like gene. Together with L-1.8, this means that lamprin is coded for by a total of three genes in P. marinus.

Exon/Intron Structure of Lamprin Genes from P. marinus

Although screening of a genomic library would have been the preferred method of cloning the three P. marinus lamprin genes, difficulty in the construction of this genomic library resulted in an alternative PCR approach. The first round of amplification used three different primer sets, each specific to one of the lamprin genes. Forward primers were designed for the 5′ UTRs, and reverse primers were designed for the 3′ UTRs. For L-0.8 and L-0.9, the same forward primer was used. This forward primer differed from the corresponding region in L-1.8 by 221 positions. The reverse primer for L-0.8 amplification spanned the point of deletion in the 3′ UTR that distinguished L-0.8 from L-0.9. The sequence inserted at this site in L-0.9 was used to construct the L-0.9 reverse primer. For L-1.8, the forward primer differed from the corresponding region in the other two genes at 11/27 positions. The reverse primer had 21/25 and 20/25 sequence similarity to L-0.9 and L-0.8, respectively. The sequence differences between these primer sets, combined with the high annealing temperature used in the amplification reaction, were sufficient for gene-specific amplification. The three PCR products generated from these primer sets ranged from 8 to 9.5 kb in size. These products were used as templates for the identification of the individual exon/intron boundaries. The amplification products were subcloned and sequenced to determine precise exon/intron junctions. The exon boundaries are summarized in figure 2.

Eight exons were identified, a relatively large number given the small sizes of the lamprin gene products (10 and 12 kDa). The first exon contains the 5′ UTR, the signal peptide, and the first amino acid of the mature protein. The third exon contains the sequence that encodes the GGLGY pentapeptide repeat, a characteristic motif in the lamprin protein sequence. The fourth exon is subject to alternate splicing (see below). The stop codon is located within the sixth exon, which also includes 12 bp of the 3′ UTR. Exon 7 is present only in L-1.8. Exon 8 is the largest exon in all three genes and contains the majority of the 3′ UTR. As indicated above, the 3′ UTR of L-0.8 differs from that of L-0.9 by the deletion of a 184-bp sequence.

Alternatives of exon 4 in P. marinus

The previous characterization of lamprin cDNAs suggested that transcripts from L-0.9 could be subject to alternative splicing. Delineation of the intron/exon boundaries (figs. 2 and 3) confirmed that this insert corresponded to an individual exon, exon 4 of L-0.9. This indicated that the L-0.9-10 transcript could indeed be the result of the splicing out of exon 4 from the mRNA rather than the product of a second lamprin gene, otherwise identical but lacking this exon.

The mRNA for L-1.8-10 lacked the sequence coded for by exon 4, characterization of the gene for L-1.8 clearly demonstrated the presence of this exon (fig. 3). Similarly, sequence for exon 4 was also present.
in L-0.8. In order to determine whether alternative splicing of exon 4 might also take place in transcripts of L-1.8 and L-0.8, primer sets specific for each of the two genes were designed. The forward primer corresponded to exon 3 and recognized both genes. The L-1.8-specific reverse primer corresponded to a sequence in exon 7, an exon present only in L-1.8. The L-0.8-specific reverse primer was the same as that used for amplification of this entire gene.

PCR amplification of the original annular cartilage cDNA library with these primer sets produced two bands per primer set (lane 1 in fig. 4A and B). These differed in size by approximately 60 bp, a size consistent with that of exon 4. The use of a forward primer corresponding to exon 4 rather than exon 3 in this amplification resulted in a single band in each (lane 3 in fig. 4A and B). This confirmed that the pair of bands produced from the first primer set was the result of alternate splicing in exon 4 and not due to splicing differences downstream of this exon. The presence of exon 4 was confirmed by sequencing appropriate amplification products. These results indicated that exon 4 was subject to alternative splicing in all three lamprin genes, resulting in a total of six distinct mRNA products.

Genomic Structure of L. richardsoni

Prior to screening the genomic library for L. richardsoni, we confirmed by Northern blot analysis that lamprin cDNAs from P. marinus cross-hybridized with
the corresponding mRNAs in *L. richardsoni*. Lamprin mRNAs similar in size to those of *P. marinus* (~0.9 and 1.8 kb) were identified (data not shown). These cDNAs were then used to isolate seven unique genomic clones, designated A–G, from the *L. richardsoni* genomic library. Each of these clones was shown to be unique based on *SalI* restriction fragment differences.

To identify *SalI* fragments that were hybridization-positive for lamprin cDNAs, two Southern blots were done on the seven *SalI*-digested genomic clones. One probe was a 177-bp fragment from the 5' end of the L-0.9-12 cDNA. A second probe contained the remainder of the sequence from this cDNA. In five of the seven clones (A, B, D, F, and G in fig. 5A and B), the same two bands hybridized with both the 5' and the 3' probes. In clone C, a single (but different) fragment hybridized to each of the probes. In clone E, one fragment hybridized to the 5'-end probe. This fragment, as well as a second fragment, hybridized to the 3'-end probe in this clone. These data provided the first indication that the lamprin genes in *L. richardsoni* were arranged in tandem.

Mapping the *SalI* fragments from genomic clone D and partial sequence data of exon-containing regions within this clone (fig. 5C) clearly showed that lamprin genes are located in tandem in the *L. richardsoni* genome. Exon 1 and a region of exon 8 for one lamprin gene and exons 1, 2, 3, and 5 from the second gene found in genomic clone D were sequenced. These results indicated that the two lamprin genes in this clone were oriented in a head-to-tail manner and were separated by 5.7–7.0 kb (depending on the length of exon 8 in the first gene) from the 3' end of one gene to the transcription start site of the second. Interestingly, although the intervening region between exons 3 and 5 in the second gene was fully sequenced, there was no evidence for the presence of exon 4, the exon found to be alternatively spliced in the three lamprin genes of *P. marinus*. This observation suggests that additional variants of lamprin genes may be present in the *L. richardsoni* genome.

**Identification of the Transcription Start Site in *P. marinus***

Since a genomic library was not available for *P. marinus*, a PCR-based genome-walking approach was used to isolate 5'-upstream sequence. This approach yielded 909, 350, and 370 bp of sequence from L-1.8, L-0.9, and L-0.8, respectively. Initially, L-0.9 and L-0.8 promoter regions were amplified together, since the reverse gene-specific primers could not distinguish between the two templates. Fortunately, sequence differences became apparent between the newly cloned upstream sequences from L-0.8 and L-0.9 (see fig. 7; positions −190 to −170). Forward primers designed for this region specific for either L-0.8 or L-0.9, combined with a reverse primer for exon 2, allowed us to take advantage of the size differences found in intron 1 between these two genes. The sizes of the amplified products from genomic DNA using these primer sets defined which 5'-upstream sequence corresponded to which gene.

The transcription start site of L-0.9 was determined by primer extension. Two primers, both complementary to regions within the 5' UTR of L-0.9, were hybridized to total RNA isolated from the annular cartilage of a juvenile adult lamprey. Kidney total RNA was used as a negative control. Similar bands found in both the kidney and the cartilage could be ruled out as nonspecific, since lamprin is not expressed in the kidney. Single cartilage-specific bands were identified for each of the two primers. Both primer termination sites mapped to the same position 24 bases upstream of the previously identified 5' end of the cDNA (fig. 6).

Alignment of the 5'-upstream sequences from the three *P. marinus* lamprin genes with those from the two *L. richardsoni* lamprin genes described earlier demonstrated substantial sequence conservation between both paralogous and orthologous genes (fig. 7). These sites of sequence conservation may indicate regions of importance for transcriptional regulation of these genes.

![Image](image-url)
Fig. 5.—Tandem organization of multiple lamprin genes in Lampetra richardsoni. The seven unique genomic clones obtained (A–G) were digested to completion with SalI and probed with either a 177-bp fragment corresponding to the 5’ end of the Petromyzon marinus L-0.9-12 cDNA (A) or a 720-bp fragment corresponding to the 3’ remainder of this cDNA (B). For five of the seven clones (A, B, D, F, G) both SalI fragments hybridized to both the 5’ and the 3’ probes, suggesting that at least two copies of the lamprin gene were present in each of these genomic clones. A summary of the mapping and sequencing data for the SalI fragments of genomic clone D are shown in C. The SalI sites are labeled (S), with the approximate distance between each site indicated (kb). In the case of genomic clone D, both the 6.5- and the 3.6-kb fragments hybridized with both the 5’ and the 3’ cDNA probes (A and B, respectively). Dashed lines represent regions which have been sequenced. The open boxes with corresponding numbers represent the identities and locations of the exons found in this genomic clone. Positions of molecular size markers (kb) are indicated.

Two well-conserved regions contain a putative TATA box and a cap signal. The TATA box is located approximately 29 bp upstream of the transcriptional start site, within the preferred range of this common promoter element (Bucher 1990). The cap signal, another less highly conserved RNA polymerase II promoter element, was identified in its preferred region, just downstream of the transcription start site (Bucher 1990). No regions were found that fit the consensus sequences for either a CCAAT box or a GC box, two other elements commonly found within RNA polymerase II-transcribed genes.

The most divergent of the five upstream/promoter sequences is that of the P. marinus L-1.8 gene (fig. 6). This gene shows strong sequence identity up to position $-196$ in the alignment, but no identifiable similarities with the other four sequences upstream of this position. However, a region much farther upstream in this gene ($-685$ to $-564$) contains significant sequence similarity to the $-298$ to $-174$ region in the alignment of the other genes. Given the location-independent nature of enhancer elements, conservation of this sequence in L-1.8 at another site suggests that the sequence could be important for regulation of lamprin gene transcription. The remaining four genes maintain strong sequence similarity to the end of the known sequences for P. marinus L-0.9 and L-0.8, approximately 400 bp upstream of the transcription start site. Sequence similarity between the two L. richardsoni lamprin genes remains very high up to the point approximately 530 bp upstream from the start site, beyond which sequence similarity is lost.

**Lamprin Protein Sequence Comparisons**

We previously described the protein-coding sequence for three P. marinus lamprin cDNAs encoded by
FIG. 6.—Mapping of the lamprin gene transcription start site by primer extension. Two different primers were designed for regions within the 5' untranslated region of the L-0.9 cDNA of Petromyzon marinus. The nucleotides corresponding to the 3' ends of these primers are indicated by underline and number in the sequence below. These primers were annealed to total RNA from annular cartilage and kidney (control) and extended in the 5' direction with reverse transcriptase. The sizes of the resulting fragments were determined from the gel by comparison with a standard sequencing reaction (marker lane). Both primer 1 and primer 2 terminated extension at the same location, indicated by asterisks on the gel and in the sequence. This location was 24 bases upstream of the 5' end of the L-0.9 cDNA (indicated in bold) and 25 bases downstream of the putative TATA box (open box).

Discussion
Multiple Copies of Lamprin Genes

Previous cloning of cDNAs from P. marinus suggested the existence of at least two genes for lamprin (Robson et al. 1993). Here, we demonstrated the presence of three similar but nonidentical genes, designated L-0.8, L-0.9, and L-1.8, for lamprin in the P. marinus genome. The fact that the sequences of the L-0.8 and L-0.9 genes are more similar to one another than to that of the L-1.8 gene suggests that a gene duplication event leading to the formation of L-0.9 and L-0.8 occurred more recently than that giving rise to L-1.8 and the common precursor of L-0.9/L-0.8.

As has been suggested for other multiple-copy genes (Long and Dawid 1980), including those for histones, rRNAs, and tRNAs, multiple lamprin genes may have evolved as a result of a requirement for synthesis of large amounts of the lamprin gene product. Lamprin is a highly abundant protein in lamprey cranial cartilages, accounting for at least 50% of the dry weight of the tissue (Wright, Keeley, and Youson 1983). Structural proteins of comparable abundance in their respective tissues are known to be single-copy genes (e.g., the 72-kDa elastin of aorta [Indik et al. 1987; Olliver et al. 1987] and the 96-kDa α1(II) collagen chain of mammalian cartilage [Cremer, Rosloniec, and Kang 1998]). However, because of the small size of the lamprin protein (10–12 kDa), comparable rates of gene expression for lamprin, elastin, and collagen would yield only a sixth as much lamprin as elastin or collagen. Therefore, multiple lamprin genes may have evolved to provide the necessary quantities of this abundant but relatively small protein.

Investigation of the genomic structure of lamprin genes from L. richardsoni provided clear evidence for organization of these genes in tandem in the genome of this species. Five of the seven unique genomic clones isolated from L. richardsoni contained two lamprin genes, and sequence data for clone D revealed that the two genes located in this clone were oriented in a head-to-tail fashion and separated by no more than 7 kb of intervening sequence. Although 7 kb is technically well within the size range for PCR amplification, attempts to amplify intervening sequence between P. marinus genes were unsuccessful. This suggests that if lamprin genes
Fig. 7.—Alignment of 5’-upstream sequences from five lamrin genes. Here, 5’-upstream sequences from the three *Petromyzon marinus* lamrin genes, designated L-0.8, L-0.9, and L-1.8, were aligned with 5’-upstream sequences from the two lamrin genes present in genomic clone D from *L. richardsoni*. The alignment begins 570 bp upstream of the transcription start site (+1) and ends 30 bp into exon 1. *Petromyzon marinus* L-0.9 and L-0.8 genes were sequenced upstream only to position 2403 in the alignment. A further upstream region of the L-1.8 gene from *P. marinus* (2685 to 2564) showing sequence similarity to a downstream region of the other lamrin genes (2298 to 2174) was also included in the alignment. Positions agreeing with the top sequence in each block are indicated by dots; nucleotides are indicated where they differ from the top sequence. Deletions are represented by dashes. Both the putative TATA box and the cap signal are boxed.
Genomic Organization of Lamprin 1749

Fig. 8.—Alignment of *Petromyzon marinus* and *Lampetra richardsoni* lamprin protein-coding exons. The amino acids encoded by all exons of the three *P. marinus* lamprin genes (*P.m.* ) and for the available *L. richardsoni* exons (*L.r.*) are aligned. Sequences are aligned relative to the complete sequence of *P.m.* 0.8. Amino acid differences are indicated. Dots indicate identical sequence, dashes indicate deletions. Tildes ( ~ ) indicate amino acids which could not be identified because the downstream exon encoding the last two bases of the codon had not been sequenced. An asterisk indicates the site of an additional valine resulting from a 3-bp insertion in L-0.9. The location of the signal peptide cleavage site is indicated.

<table>
<thead>
<tr>
<th>Exon</th>
<th><em>P.m.</em> 0.8</th>
<th><em>P.m.</em> 0.9</th>
<th><em>P.m.</em> 1.8</th>
<th><em>L.r.</em> B2</th>
<th><em>L.r.</em> C2</th>
<th><em>L.r.</em> D3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exon 1</td>
<td>MAAAIQALLV LALLHLATAT P</td>
<td>............... ...............</td>
<td>...............</td>
<td>S ~</td>
<td>I ...............</td>
<td>I ...............</td>
</tr>
<tr>
<td>Exon 2</td>
<td>VIKGTKTVSTL STGYLGHPV ~ G</td>
<td>............... ...............</td>
<td>V*</td>
<td>...............</td>
<td>...............</td>
<td>~</td>
</tr>
<tr>
<td>Exon 3</td>
<td>GLGGYGLGGY GLGVAGLGA GLGGYGLGYP GAALGGVYTH HAA</td>
<td>...............</td>
<td>...............</td>
<td>...............</td>
<td>A ~</td>
<td>...............</td>
</tr>
<tr>
<td>Exon 4</td>
<td>GLVHPYGGGLG YHAAPYHSA</td>
<td>...............</td>
<td>...............</td>
<td>H</td>
<td>...............</td>
<td>...............</td>
</tr>
<tr>
<td>Exon 5</td>
<td>LGGLGYPLGI GAGVVAPHV KGKVA</td>
<td>...............</td>
<td>...............</td>
<td>I</td>
<td>...............</td>
<td>...............</td>
</tr>
<tr>
<td>Exon 6</td>
<td>APLAPVVAAL</td>
<td>...............</td>
<td>...............</td>
<td>...............</td>
<td>...............</td>
<td>...............</td>
</tr>
</tbody>
</table>

are situated in tandem in *P. marinus*, the intervening sequences may be much larger.

### Alternative Splicing

The results presented here demonstrate that all three genes of *P. marinus* can be alternatively spliced in exon 4, yielding a total of six possible protein products. The functional significance of these alternatively spliced transcripts is unknown, but the preservation of alternative splicing in all three genes suggests some importance. The amino acid sequence encoded in the alternatively spliced exon is not unlike that of the remainder of the protein and contains one copy of the GGLGY pentapeptide, which we have speculated (Robinson et al. 1993) is important for the extracellular assembly of the insoluble lamprin matrix. The difference in the size of the protein created by the additional exon, rather than the primary sequence of exon 4, may be the functionally significant feature of this alternative splicing, perhaps leading to different fiber-forming properties of lamprin. Other extracellular matrix proteins, including fibronectin (Gehris et al. 1996) and elastin (Rosenbloom et al. 1991), are known to undergo alternative
splicing. At least in the case of fibronectin, this alternative splicing is both temporally and spatially regulated. Although we have shown that forms of lamprin both containing and lacking exon 4 are found in annular cartilage at the juvenile adult stage of development, differential expression of these alternatively spliced variants could potentially be important at other stages of development.

Structure of the 3' UTR of Lamprin Genes

An unusual feature of the exon structure of lamprin genes is the presence of multiple exons encoding the 3' UTR (fig. 2). In contrast to the case for all three lamprin genes described here, the translational stop codon in the majority of genes is located within the 3'-most exon. A recent exhaustive survey of stop codon position relative to exon structure found that only 7% of characterized genes have the translational termination signal within their penultimate exon (Nagy and Maquat 1998), as is the case for lamprin genes L-0.8 and L-0.9. Location of the stop codon in the third to the last exon, as for L-1.8, was seen in only 0.4% of all genes.

In the same study, the normal termination codon position was found to be <50 bases upstream of the 3'-most exon in 98% of the genes which had one or more 3' untranslated exons. This is thought to be related to a nonsense-mediated decay mechanism (Hentze and Kulozik 1999) whereby a termination codon located more than 50–55 nt upstream of the 3'-most exon-exon junction results in mRNA decay (Nagy and Maquat 1998), protecting the cell from prematurely terminated transcripts. The termination codon for all three lamprin genes falls within this range. For example, according to this termination codon position rule, exon 7 of L-1.8 could be a maximum of 45 nt in size. At 42 nt (the smallest exon found in the lamprin genes), exon 7 fulfills this requirement.

Splicing of exons into a mature mRNA is a complex process (Berget 1995; Tarn and Steitz 1997), suggesting that the presence of this additional exon in the 3' UTR of L-1.8 may have some functional significance. Interestingly, the MFOLD program in GCG predicts a very strong hairpin structure in contiguous exon 6 and 7 sequences for all three genes. However, when exon 7 sequence is included (as it is in the naturally occurring L-1.8 cDNA), this structure is disrupted (data not shown). Band compressions seen during the sequencing of the L-0.9 cDNA also suggested secondary structure in this region. As elements of secondary structure in mRNAs have been reported to be involved in regulating translational efficiency and mRNA stability (Munro and Eisenstein 1989; Klaff, Riesner, and Steger 1996), this unusual organization of the 3' UTR in the lamprin genes may have some functional role in these processes.

Lamprin Promoter Sequences

Both oligonucleotides used in the primer extension studies of L-0.9 mapped to the same transcription start site, and a characteristic TATA box was found just upstream of this location, within the usual range of these elements (Bucher 1990). Although the transcriptional start site was experimentally determined only for the L-0.9 gene, the strong sequence conservation in this region suggests that this site is likely maintained in all three lamprin genes.

Regions of sequence conservation among the three *P. marinus* genes and the two *L. richardsoni* genes may be useful in suggesting the location of functionally important enhancer elements. However, no functionally significant (i.e., cartilage-specific) enhancer elements were recognized in these lamprey genes. The nature of enhancer elements, relatively short in length and somewhat variable in sequence, combined with the remote phylogenetic position of the lamprey, suggests that transcription factor databases may not be particularly useful in identifying potential sites in lamprey genes. Thus, other than the well-conserved TATA box and cap signal, the identification of specialized promoter/enhancer elements in lamprey genes will require further functional promoter studies.

Evidence for Sequence Convergence

Regions of sequence similarity, consisting of shared tandem pentapeptide repeats based on the GGLGY sequence of lamprin, had previously been identified among lamprey cartilage proteins, insect eggshell proteins, and mammalian and avian elastins (fig. 9) (Robson et al. 1993). Several considerations suggest that these similarities in sequence are more likely to be an example of sequence convergence than the result of divergence from a common ancestral protein existing more than 500 MYA.

In the first place, outside of these regions of similarity based on GGLGY repeats, the remainder of the sequences of these proteins are all very distinct. Furthermore, both for elastins and for the large family of silkmoth chorion proteins, sequence similarities in these regions are stronger between different groups of proteins (e.g., lamprin, elastin, chorion proteins) than between members of a family of proteins (Spoerel, Nguyen, and Kafatos 1986; Boyd et al. 1991).

Differences in gene structure between lamprin and chorion genes also support arguments for sequence convergence. We have shown that exon junctions of lamprin genes are all of class 1,1, with introns splitting codons between the first and second bases. This type of junction is common among extracellular matrix proteins (Patthy 1987, 1991), including elastin (Bashir et al. 1989; Yeh et al. 1989) and would allow for exon shuffling and duplication (Patthy 1987), mechanisms which appear to have been prevalent in the evolution of at least some extracellular matrix proteins (Patthy 1991). Although the genomic organization of oothecin is not known, the single intron of the large family of silkmoth chorion proteins always falls between complete codons (Spoerel, Nguyen, and Kafatos 1986; Bucher 1990; Hibner, Burke, and Eickbush 1991). Such a difference in exon classes between lamprin and chorion genes make it less likely that they share exons from a common ancestral gene, since simple exon shuffling using exons of differ-
ent classes would interfere with the reading frame and alter protein sequence. If this is the case, then sequence similarities between lamprin and oothecin, which share a 28/30 amino acid sequence identity, may represent one of the best examples of primary sequence convergence so far identified.

Sequence convergence is presumed to be driven by structural or functional properties imparted to the protein by the shared sequence. All of the proteins showing sequence similarities to lamprin, including elastins and the insect structural proteins, are components of the extracellular matrix, where they are assembled into extensive fibrillar structures. We have suggested elsewhere (Robson et al. 1993) that such sequences may be important for the ability of these predominantly hydrophobic proteins to self-organize into a polymeric, fibrillar matrix. Requirements for such hydrophobic self-aggregation may be an important factor limiting the selection of amino acids used in the tandem repeats in these various structural proteins. Therefore, such simple repetitive sequences may have been independently “reinvented” in unrelated proteins in order to provide such characteristics. Cloning and characterization of morphologically and biochemically similar proteins which form the extracellular matrix of lamprey branchial cartilage (Robson et al. 1997), hagfish cartilages (Robson, Wright, and Keeley 2000), and noncollagenous cartilages of several invertebrates (unpublished data) will shed additional light on the functional and evolutionary relationships among these unusual cartilage structural proteins.

Acknowledgments

This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

LITERATURE CITED


FIG. 9.—Alignment of regions of proteins with sequence similarity to lamprin. Oothecin from the American cockroach (accession number 224080), chorion class B protein from the domestic silkmoth (accession number CAA33566), and elastin from the rat (exon 30; accession number Q99372) all show similarities to the GGLGY tandem repeat found in all forms of lamprin (boxed). Oothecin has the greatest similarity to lamprin, with a 28/30 amino acid identity. Silkmoth chorion class B protein has a 21/24 amino acid identity. Of the known elastin sequences, exon 30 of rat elastin shows the greatest sequence similarity to lamprin, with a sixfold repeat of the sequence GGLGA.