Exon-Intron Structure and Evolution of the Lipocalin Gene Family

Diego Sánchez,[1] María D. Ganfornina,*[1] Gabriel Gutiérrez,† and Antonio Marín†*

*Departamento de Bioquímica y Fisiología y Genética Molecular-IBGM, Universidad de Valladolid-CSIC, Valladolid, Spain; and †Departamento de Genética, Universidad de Sevilla, Sevilla, Spain

The Lipocalins are an ancient protein family whose expression is currently confirmed in bacteria, protoctists, plants, arthropods, and chordates. The evolution of this protein family has been assessed previously using amino acid sequence phylogenies. In this report we use an independent set of characters derived from the gene structure (exon-intron arrangement) to infer a new lipocalin phylogeny. We also present the novel gene structure of three insect lipocalins. The position and phase of introns are well preserved among lipocalin clades when mapped onto a protein sequence alignment, suggesting the homologous nature of these introns. Because of this homology, we use the intron position and phase of 23 lipocalin genes to reconstruct a phylogeny by maximum parsimony and distance methods. These phylogenies are very similar to the phylogenies derived from protein sequence. This result is confirmed by congruence analysis, and a consensus tree shows the commonalities between the two source trees. Interestingly, the intron arrangement phylogeny shows that metazoan lipocalins have more introns than other eukaryotic lipocalins, and that intron gains have occurred in the C-termini of chordate lipocalins. We also analyze the relationship of intron arrangement and protein tertiary structure, as well as the relationship of lipocalins with members of the proposed structural superfamily of calycins. Our congruence analysis validates the gene structure data as a source of phylogenetic information and helps to further refine our hypothesis on the evolutionary history of lipocalins.

Introduction

Ever since the rise of Gram-negative bacteria the lipocalins have been functioning and evolving in these organisms and in their eukaryotic symbionts, which possibly acquired the primordial lipocalin gene through a horizontal transmission event (Bishop 2000). A protein family developed through the standard evolutionary mechanisms of gene duplication and divergence (Ohno 1999), giving rise to at least 10 different genes that are currently recognized in the most recently evolved organismal taxa. Previous studies of three-dimensional and sequence similarity grouped lipocalins in kernel or outlier subfamilies based on the presence of three structurally conserved protein regions (SCRs) (Flower, North, and Attwood 1993) in the β barrel-based structural fold of lipocalins (fig. 1*A*).

Several phylogenetic reconstructions have been built upon the alignment of amino acid residues of lipocalin sequences (e.g., Igarashi et al. 1992; Toh et al. 1996). We have performed comprehensive phylogenetic analyses of the lipocalin family using protein sequence alignments with guidance based on protein structure data, and tree-building methods based on maximum likelihood (ML) (Ganfornina et al. 2000; Gutiérrez, Ganfornina, and Sánchez 2000). These studies group lipocalins in well-supported clades. When rooted with the bacterial lipocalins, the topology of the tree and the organismal distribution of lipocalins suggest that these proteins tend to increase the rate of sequence divergence and of gene duplication during evolution. Also, their internal pocket appears to have evolved toward binding smaller hydrophobic ligands with more efficiency.

An extensive literature supports the conservation of exon-intron structure in clades of orthologous genes (COGs) (Rokas, Kathirithamby, and Holland 1999; Wada et al. 2002), as well as in families of paralogous genes (Krem and Di Cera 2001) and protein superfamilies (Betts et al. 2001). These findings support the use of gene features as sources for phylogenetic inference (Rokas and Holland 2000; Krem and Di Cera 2001). In a previous report Salier (2000) proposed a scenario for the evolution of the lipocalin gene family by studying the gene structure and chromosomal location of 15 lipocalins. However, this view of lipocalin evolution is very dependent on the concepts of kernel versus outlier subfamilies, and it conflicts with our proposed evolutionary history of lipocalins (Gutiérrez, Ganfornina, and Sánchez 2000). To reassess our hypothesis of lipocalin evolution, we have used gene structure features as characters to build phylogenetic trees through different tree-reconstruction methods.

In this report we present the gene structure data of three insect lipocalins that we have been studying for their role in nervous system development (Ganfornina, Sánchez, and Bastiani 1995; Sánchez, Ganfornina, and Bastiani 1995; Sánchez et al. 2000*b*). The position and phase of introns in a number of lipocalins are used to reconstruct a phylogeny by maximum parsimony methods and by a distance matrix built with a measure of gene structure similarity (Betts et al. 2001). We also analyze the variability in introns present in the C-termini of lipocalins belonging to different COGs, and compare lipocalin intron arrangement with tertiary structure. We test the conservation of intron arrangement within the calycins, a proposed structural superfamily (reviewed by Flower, North, and Sansom 2000). Finally, we analyze the congruence of phylogenies based on protein sequence and gene structure, and build a consensus tree to refine our hypothesis on lipocalin evolution.

Materials and Methods

Genomic PCR Amplification of the Lazarillo Gene

Genomic DNA was purified from grasshoppers (*Schistocerca americana*). Brain tissue was lysed in 25 mM EDTA, 0.5% SDS, and 0.1 mg/ml proteinase K. The

[1] Contributed equally to this work.

Key words: lipocalin, calycin, molecular evolution, gene phylogeny, exon-intron, intron evolution.

E-mail: opabinia@ibgm.uva.es.

Mol. Biol. Evol. 20(5):775–783. 2003
DOI: 10.1093/molbev/msg079

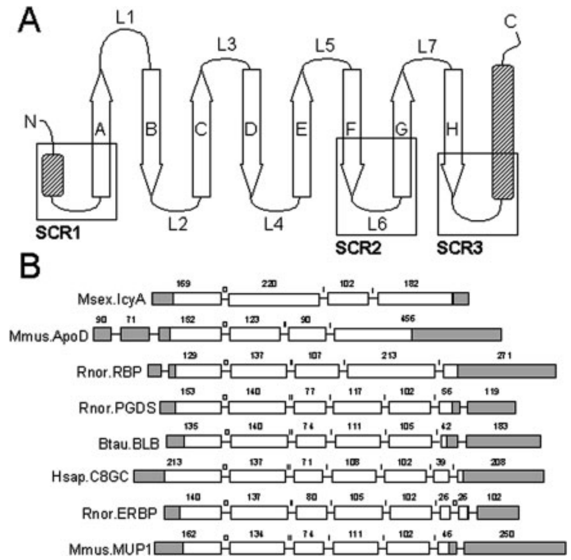
A
L1
L3
L5
L7
N
C
A
B
C
D
E
F
G
H
L2
L4
L6
SCR1
SCR2
SCR3
B
Msex.IcyA 169 0 220 I 102 I 182
Mmus.ApoD 90 71 152 0 123 II 90 I 456
Rnor.RBP 129 0 137 II 107 I 213 I 271
Rnor.PGDS 153 0 140 II 77 I 117 I 102 I 56 119
Btau.BLB 135 0 140 II 74 I 111 I 105 I 42 183
Hsap.C8GC 213 0 137 II 71 I 108 I 102 I 39 I 208
Rnor.ERBP 140 0 137 II 80 I 105 I 102 I 26 0 26 102
Mmus.MUP1 162 0 134 II 74 I 111 I 102 I 46 250

FIG. 1.—*A*, Schematic diagram of the topology of the lipocalin structural fold. β strands are represented by white arrows, lettered A–H. α helices are shown as barreled cylinders. The boxes outline the three structurally conserved regions (SCRs). *B*, Schematic representation of the position of introns in representatives of lipocalin clades. White boxes correspond to the gene CDS, and gray boxes represent the untranslated regions. Size (in nucleotides) is shown above each exon. Lines represent intron insertions (not drawn to scale), and the phase of each intron is indicated above the line.

DNA was extracted with phenol and RNAse A–treated. We used the Expand Long Template PCR System (Roche Biochemicals) following manufacturer's specifications to perform polymerase chain reaction (PCR) amplifications in a thermal cycler (GeneAmp 9700 Perkin Elmer) using thin-walled plastic tubes (PE Biosystems). Primers were designed from the Lazarillo cDNA sequence (GenBank Accession Number U15656) with the primer3 program (www-genome.wi.mit.edu/cgi-bin/primer/primer3_WWW.cgi), and synthesized by Amersham Pharmacia Biotech. Primer sequences were: A5′ (GTGCTGCTGTCTGTAAGCTG) and A3′ (TGGAGTTGACGACTGTGATG) for amplification of intron A; B5′ (ACGGCAGAGTACTCCATGTCG) and B3′ (AGCTCGCTCGCGAACTCTGC) for testing the presence of intron B; D5′ (CGACAACTACTCCATTGTGTGG) and D3′ (GCTGCAGATTCTTCAGCTCATC) for amplification of intron D; and primers EF5′ (TCCTATTACGATCACGGAAC) and EF3′ (TCATGACTCGCTGACCATAC) for testing the presence of introns E/F. Polymerase chain reaction products were sequenced with an ABI Prism 377 automated DNA sequencer using Taq FS DNA Polymerase.

Sequence Searches and Alignments

We searched for lipocalin genes whose intron-exon structure has been confirmed by the knowledge of their mRNA sequence. No deduced intron-exon arrangement was included in the analysis to avoid ''noise'' produced by poorly predicted splice sites, and to discard pseudogenes. Using the same seeding process and selection criteria previously described (Ganfornina et al., 2000; Gutiérrez, Ganfornina, and Sánchez 2000), a search for lipocalin cDNA and EST sequences was performed using the Blast program (Altschul et al. 1990) in the GenBank database available December 13, 2001. Thirty-seven of the sequences retrieved contained the complete CDS of the lipocalins and had the corresponding genomic sequence available on the databases. These genes are shown in table 1. We evaluated the presence, location, and phase of introns for these genes, and made a selection (asterisks in table 1) based on two criteria: (1) being the representative of a lipocalin COG (to avoid sampling bias), and (2) showing an intron pattern unique in the family. Thus we are accounting for the overall gene structure variation present in the lipocalin family.

Protein sequences were aligned with ClustalX (1.8) (Thompson et al. 1997) using a Gonnet series scoring matrix and a gap penalty mask based on the aligned secondary structures of the lipocalins with known tertiary structure. Based on our knowledge of lipocalin structure and function, we made minor manual corrections to the alignment. Intron positions and phases were then mapped onto the protein sequence alignment. Intron phase was named 0 when the intron splits two consecutive codons; I if an intron locates between the first and second codon nucleotides; and II if an intron locates between the second and third codon nucleotides. We used only the introns intervening the ORF of lipocalin genes, as those located in the 5′- and 3′-UTR can not be mapped onto the protein sequence alignment.

Phylogenetic Analyses

Phylogenetic analyses based on protein sequences were carried out using the maximum likelihood method with the MOLPHY 2.3 software (Adachi and Hasegawa 1996) as previously reported (Ganfornina et al. 2000). Bootstrap support for tree branches was estimated using the resampling log likelihood method (Hasegawa and Kishino 1994) to calculate local bootstrap proportions (LBP).

We have used intron positions of 23 representative lipocalins as phylogenetic characters. We built three input matrices based on three intron character states: (1) the presence/absence of a given intron, (2) the intron phase, and (3) the intron position in the alignment. Two procedures were carried out: The first was a maximum parsimony analysis using the intron presence matrix as input. Characters were considered as unordered. We made heuristic tree searches by the TBR method of PAUP* (Swofford 1998). A majority rule consensus tree was constructed from the most parsimonious trees found in the analysis. The second procedure started with the construction of a distance matrix based on a measure of gene structure similarity (Betts et al. 2001) that uses the presence, location, and phase of intron matrices described above to estimate the exon-intron similarity between two of the aligned proteins.

$$S_G(a,b) = \frac{1}{2N_{\max}} \sum_{i=1}^{N_{\text{equiv}}} \left(\frac{1}{1 + e^{\gamma(d_i - \delta)}} + \varphi(a_i, b_i) \right)$$

Table 1
List of Experimentally Determined Lipocalin Gene Structures

Protein	Species	Abbreviation	Clade	Taxon[a]	Accession Number
Lipocalin	*Dictyostelium discoideum*	Ddis.Lip*	I	P	JC1b154f03
Lipocalin fly neural-Lazarillo	*Arabidopsis taliana*	Atha.OML*	I	Pl	NC_003076
	Drosophila melanogaster	Dmel.Nlaz*	II	A	L81559
Lipocalin fly glial-Lazarillo	*Drosophila melanogaster*	Dmel.Glaz*	II	A	DS01087
Lipocalin Karl	*Drosophila melanogaster*	Dmel.Karl*	II	A	AE003487
Insecticyanin A	*Manduca sexta*	Msex.IcyA*	II	A	X64714
Insecticyanin B	”	Msex.IcyB	II	A	X64715
Lazarillo	*Schistocerca americana*	Same.Laz*	II	A	In process
Apolipoprotein D	*Homo sapiens*	Hsap.ApoD*	II	PM	M16648–9
M16695–6					
”	*Mus musculus*	Mmus.ApoD	II	PM	NW_000107
Retinol binding protein	*Homo sapiens*	Hsap.RBP	III	PM	NT_030084
”	*Rattus norvegicus*	Rnor.RBP*	III	PM	M10610
K03045–6					
Beta-lactoglobulin B	*Bos taurus*	Btau.BLB*	IV	PM	Z48305
”	*Capra hircus*	Chir.BLB	IV	PM	Z33881
Beta-lactoglobulin	*Macropus eugenii*	Meug.BL*	IV	MM	L14954–60
Beta-lactoglobulin B	*Ovis aries*	Oari.BLB	IV	PM	X12817
Beta-lactoglobulin A	”	Oari.BLA	IV	PM	M32232–37
Glycodelin	*Homo sapiens*	Hsap.Glyc*	IV	PM	M34046
Prostaglandin D synthase	*Homo sapiens*	Hsap.PGDS	V	PM	M98537–39
”	*Rattus norvegicus*	Rnor.PGDS*	V	PM	M94134
”	*Mus musculus*	Mmus.PGDS	V	PM	Y10138
Neutrophil gelatinase lipocalin	*Homo sapiens*	Hsap.NGAL*	V	PM	X99133
”	*Mus musculus*	Mmus.NGAL	V	PM	X81627
Quiescence protein-21	”	Ggal.QS-21*	V	B	AF121346
Alpha-1 microglobulin	*Homo sapiens*	Hsap.A1mg	VI	PM	M88165
M88243–47					
M88249					
”	*Mus musculus*	Mmus.A1mg*	VI	PM	AF034692
Complement C8γ subunit	*Homo sapiens*	Hsap.C8GC*	VII	PM	U08198
Major urinary protein 1	*Mus musculus*	Mmus.MUP1*	VIII	PM	X03208
Aphrodisin	*Mus musculus*	Mmus.Aphr	X	PM	NW_042625
Aphrodisin	*Mesocricetus auratus*	Maur.Aphr*	X	PM	AJ225170
Alpha-1 acid glycoprotein 2	”	Hsap.a1G2*	XII	PM	AH007409
von Ebner’s gland protein	*Homo sapiens*	Hsap.VEG	XIII	PM	L14927
”	*Sus scrofa*	Sscr.VEG*	XIII	PM	V96150
von Ebner’s gland protein 1	*Rattus norvegicus*	Rnor.VEG1	XIII	PM	X74805
von Ebner’s gland protein 2	”	Rnor.VEG2	XIII	PM	X74807
Epididymal RA-binding prot.	*Rattus norvegicus*	Rnor.ERBP*	XIV	PM	X59831
Epididymal RA-binding prot.	*Mus musculus*	Mmus.ERBP	XIV	PM	U68381
Odorant binding protein	*Mus musculus*	Mmus.OBP1	X	PM	NW_042625
Odorant binding protein IIa	*Homo sapiens*	Hsap.OBP2a*	X	PM	AJ251029
Odorant binding protein IIb	*Homo sapiens*	Hsap.OBP2b	X	PM	AJ251025

[a] Abbreviations: A, arthropod; Am, amphibian; Bi, bird; F, fish; MM, marsupial mammal; PM, placental mammal; P, protoctist; Pl. plant; R, reptile.

* Lipocalins chosen for gene structure phylogenetic analysis (see criteria in *Materials and Methods*).

S_G is the similarity measure for the two proteins (a and b); $N_{\max}$ is the largest number of introns found in either protein; N_{equiv} is the number of equivalent (homologous) introns; a_i and b_i are the ith equivalent intron positions in the two proteins; d_i is the difference in position of the introns within the two proteins (in amino acids); $\varphi(a_i,b_i)$ is 1 if the intron phases are the same and 0 if they are different; γ and δ are constants (0.2 and 30) optimized such that the sigmoid function is insensitive to small changes in intron positions (±10 residues) (Betts et al. 2001). A matrix of distances calculated by this method was used to reconstruct a tree by the Neighbor-Joining method (Saitou and Nei 1987) implemented in PHYLIP (Felsenstein 1993). The generation of consensus trees and the analysis of congruence were performed with the RadCon program (Thorley and Page 2000). The protein sequence alignments and gene structure data matrices used for the phylogenetic studies are available from the authors upon request.

Results and Discussion

Gene Structure in the Lipocalin Family

Lipocalin genes have been found to contain four to eight exons (Salier 2000). A schematic representation of the position of introns is shown in figure 1*B* for representative lipocalins. Most introns interrupt the ORF, and only a few appear to be located in the 5′- and 3′-UTR of lipocalins. The position and phase of introns intervening the lipocalins ORF appeared to be fairly conserved. These similarities in exon-intron organization provide strong support for a common origin of the lipocalin genes and therefore make gene-structure information suitable for phylogenetic inference.

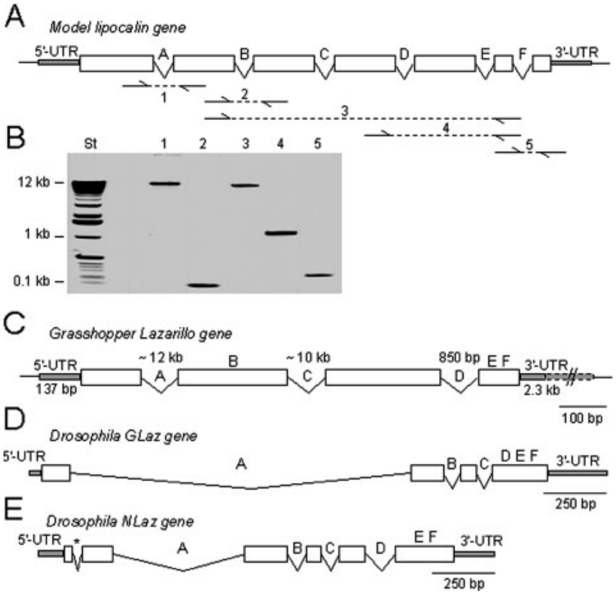

FIG. 2.—Exon-intron arrangement of the insect Lazarillo genes. *A*, Diagram of the gene structure of a model lipocalin. Introns in the ORF region are named A–F. Arrows and numbers below them show the primer sets designed to amplify specific introns from the genomic DNA. *B*, Photograph of an ethidium bromide gel showing the results of PCR amplifications from grasshopper genomic DNA with the primer sets shown in *A*. *C–E*, Diagram of the Lazarillo genes in grasshopper and Drosophila. Introns size are shown by numbers in *C*, and scaled in *D* and *E* as represented by the scale bars. The asterisk in *E* shows a unique intron in the 5′ region of the DNLaz ORF.

Before the present analysis, sampling of metazoan lipocalin genes was strongly biased toward the chordate phylum. Only one gene from arthropods was reported (Li and Riddiford 1992). Therefore, we set out to study the gene structure of other known arthropodan lipocalins.

Exon-Intron Arrangement of the Lazarillo Genes in Schistocerca and Drosophila

The gene structure of Lazarillo, a lipocalin found in the grasshopper *Schistocerca americana* (reviewed by Sánchez, Ganfornina, and Bastiani 2000*a*), would be of great value in providing insight into lipocalin evolution because of the ancestral position of orthopteroids within the arthropod lineage (Caterino, Cho, and Sperling 2000).

The ORF of lipocalin genes is interrupted by 6 introns at the most. These introns (named A–F) are represented in a model lipocalin depicted in figure 2*A*. The predicted location of the six introns in the grasshopper Lazarillo gene was deduced by locating intron positions in a multiple protein sequence alignment of Lazarillo with other lipocalins of known gene structure. We then designed Lazarillo primers that would PCR amplify specific introns from genomic DNA. The primer sets are shown numbered under the lipocalin model in figure 2*A*. The PCR amplifications using grasshopper DNA appear in the ethidium bromide gel shown in figure 2*B*. Each numbered lane refers to the set of primers used. These amplifications revealed the presence of three introns in the CDS of the Lazarillo gene (fig. 2*C*) that corresponded to introns A, C, and D of the model lipocalin gene. Intron size was estimated by band size for introns A and C, and by complete sequencing for the short intron D. Sequencing the PCR products defined the exact location and phase of the Lazarillo introns (see table 2). These intronic sequences are deposited in GenBank (Accession Numbers: AY197702, AY197703, AY197704, and AY197705).

The availability of the Drosophila genome sequence has made possible to locate the introns present in NLaz and GLaz, the two fruit fly lipocalins homologous to Lazarillo (Sánchez et al. 2000*b*). The intron location and size are represented schematically in figure 2*D–E*, and their sequence boundaries are shown in table 2. Three and four introns are present respectively in the GLaz and NLaz genes that are common to other lipocalins (see below). A unique intron located in the signal peptide (pointed with an asterisk in fig. 2*E*) is present in the N-terminal region of NLaz.

Phylogenetic Analysis of Lipocalins Based on Gene Structure

In addition to the already characterized lipocalin genes (Salier 2000) and the arthropodan Lazarillo genes reported above, we searched for other lipocalin genes whose intron-exon structure was confirmed by the knowledge of their mRNA sequence. All the lipocalin genes found are listed in table 1, with genes selected for the analysis marked with asterisks (23 representatives; see *Materials and Methods*).

We found a protoctist gene (from *Dictyostelium discoideum*, EST #C24642*)*, a plant gene (from *Arabidopsis thaliana*, mRNA Acc. Number AY062789), and another Drosophila gene (Karl, EST # NM_132520). The Dictyostelium and Arabidopsis genes are of singular value for our evolutionary analysis because they are the only representatives of lipocalins from unicellular eukaryotes and plants.

Alignment of Lipocalin Gene Structures

The intronic architecture of the selected lipocalin genes was mapped onto a multiple protein sequence alignment in the context of the overall secondary structure of an archetypal lipocalin (fig. 3). Noteworthy, there is a strong conservation of the location and phase of introns, a finding also reported in other gene structure analyses (Igarashi et al. 1992; Holzfeind and Redl 1994; Toh et al. 1996; Lindqvist et al. 1999; Salier 2000). This conservation is evident among COG members, but also among paralogous lipocalins. Some intron positions and phases are very well conserved (e.g., intron A), while others show slight variations (e.g., B and C). Some introns are present in most lipocalins (e.g., introns A and C) while others are present only in a subset of them (e.g., introns D, E, and F).

An important assumption of our analysis is the homology of each intron (A–F) found in the ORF of lipocalins. We accept that some variation in intron position could be due to ambiguities in the alignment of paralogous genes, where nearby insertion/deletions can cause apparent displacement of intron positions (Stoltzfus et al. 1997). A systematic examination of orthologous sequences would be needed to evaluate the presence and relevance of

Table 2
Exon-Intron Boundaries Present in the CDS of Drosophila and Schistocerca Lipocalin Genes

Lipocalin Gene	Intron	Splice Donor		Splice Acceptor		Codon Phase
Dmel.DNLaz	α	CAC TCG AG His Ser Se	gtaagcgcca	atccccacag	T TCG CAC r Ser His	2
	A	GCG GAA GCG Ala Glu Ala	gtgagttctg	aatacttcag	TAT ATG GGC Tyr Met Gly	0
	B	AAT CGA TT Asn Arg Le	gtgagtatca	gatgaaaaag	C ACC GGA u Thr Gln	2
	C	CCG ACG C Pro Thr G	gtgagtaatg	tacattttag	AG CCA TTG ln Pro Leu	1
	D	AAT TTC A Asn Phe L	gtgagttaat	ttaattgcag	AA ATT GTT ys Ile Val	1
Dmel.DGLaz	A	ATG AGT CGG Met Ser Arg	gtaagttagt	tatcttgtag	GTC CTT GGA Val Leu Gly	0
	B	AAT CGC AT Asn Arg Il	gtatgattaa	tcctttttag	A ACT GGT e Thr Gly	2
	C	GAT TTT AAG Asp Phe Lys	gtatctacaa	tttttcctag	TTT ACC ACC Phe Thr Thr	0
Same.Laz	A	GCC ACG CTG Ala Thr Leu	Unsequenced	gatttcgtag	TAC ATG GGG Tyr Met Gly	0
	C	AGT GTT G Ser Val G	gtgagtttac	aatgttgcag	GT AAC TAC Ly Asn Tyr	1
	D	TCT ACA G Ser Thr G	gtcagtcagt	ctctgtgcag	AA ATC TCA lu Ile Ser	1

intron-sliding as an additional source of variation. Our selection of genes, most of them paralogous to each other, precludes us from answering this question. Nevertheless, independently of its source, the intron position variation is incorporated in our distance measure (see *Materials and Methods*) and is used to assess the evolutionary history of lipocalins.

Taking into account the presence, position and phase of introns, we performed both maximum parsimony and distance-based phylogenetic reconstructions. The resulting trees (fig. 4) are rooted with the Dictyostelium lipocalin for its presence in an ancient organismal lineage (whose origin predates the arrival of metazoans), and because of the ancestral character of this protein sequence as judged by its similarity to bacterial lipocalins.

Maximum Parsimony Analysis

This analysis recovered six equally parsimonious trees (minimum step number 8). The majority rule consensus tree is shown in figure 4*A*. This tree (that computes the presence or absence of introns A–F as discrete character states) resolves five gene structure-related groups: (1) the Dictyostelium and plant lipocalins, (2) two arthropodan lipocalins (Laz and IcyA), (3) two Drosophila lipocalins plus ApoD and RBP, (4) a numerous group of lipocalins that belong to the clades IV-XIII (defined in our protein phylogeny, Gutiérrez, Ganfornina, and Sánchez, 2000; see table 1 for details), and (5) the three lipocalins bearing six introns (C8GC, a1mg and ERBP). The fruit fly Nlaz gene sets apart, although grouped with the remainder arthropodan lipocalins, due to its unusual set of introns.

Distances Phylogeny

A distance-based phylogenetic reconstruction was carried out by computing a distance matrix with gene structure data (intron presence, location, and phase). These data were combined to produce a quantitative measure of gene structure similarity (Betts et al., 2001; see *Materials and Methods*). The Neighbor-Joining (NJ) tree (Saitou and Nei 1987) rooted with the Dictyostelium lipocalin is shown in figure 4*B*. Similar to the parsimony tree, the NJ tree relates monophyletically most arthropodan lipocalins with ApoD, and segregates the Drosophila NLaz and RBP as genes with unique exon-intron structures. The Arabidopsis and Dictyostelium lipocalins remain at the base of the tree, and the set of lipocalins belonging to clades IV-XIII are forming a monophyletic group, also related to the 6-intron C8GC, a1mg and ERBP. Despite displaying short branch lengths, this tree also establishes relationships among different lipocalin COGs, as can be seen in the cladogram shown in figure 4*B*.

Gene Structure versus Protein Sequence Phylogenies

The gene structure-inferred view of lipocalin evolution shares basic topological features with the protein sequence-based phylogeny (see Gutiérrez, Ganfornina, and Sánchez 2000). Although in principle there are no reasons to expect congruence between these two trees, it is clear that in both phylogenetic reconstructions the arthropodan lipocalins are related to ApoDs, and they appear related to protoctist and plant lipocalins; RBPs form a separate group, related to some insect lipocalins; and the rest of lipocalins form a well supported monophyletic group. To further test this, we built a ML tree using the protein sequence alignment from which the gene structure matrices were derived, and rooted this tree with the Dictyostelium lipocalin (fig. 5*A*). We used the program RadCon (Thorley and Page 2000) to evaluate the congruence of the protein sequence ML and the gene structure NJ trees. Both source trees are well resolved:

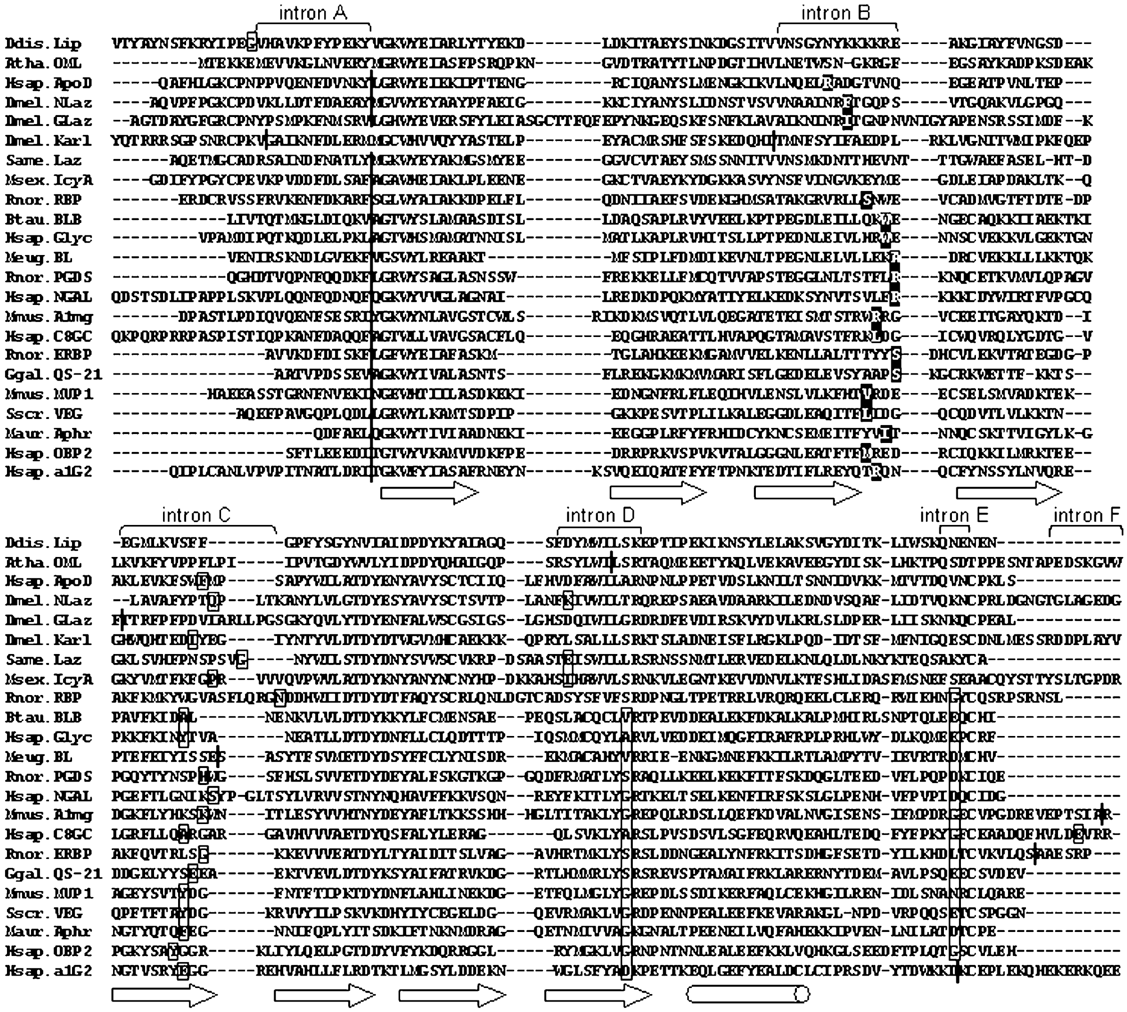

FIG. 3.—Alignment of the mature proteins of lipocalin representatives with known gene structure. The position and phase of the introns are mapped onto the alignment in the context of the overall secondary structure of an archetypal lipocalin (β strands are represented by white arrows, and α helices by cylinders). Intron phase 0 is shown as a line between the split codons; phase 1 or 2 introns as open or shaded boxes around the amino acids presenting the split codon.

their cladistic information content, a normalized measure of how much a tree reduces uncertainty regarding phylogenetic relationships (Thorley, Wilkinson, and Charleston 1998), is 0.98 for the gene structure NJ tree, and 1.00 for the protein sequence ML tree.

We also analyzed the positional congruence of each lipocalin COG in the two source trees. The normalized congruence measure, called ‘‘explicitly agree’’ (EA) similarity (Estabrook 1992), is shown in figure 5*A* for each lipocalin, and the average EA similarity (the EA similarity of the trees) is 0.789. This measure reflects the high congruence of the topology of both trees, and suggests that both phylogenetic reconstructions are good estimates of the evolutionary history experienced by lipocalins. Following the strict nesting method (Adams 1972), we built a consensus tree (fig. 5*B*) that shows the commonalities between the two source trees. The consensus tree further corroborates the orthology of ApoDs to the arthropodan lipocalins, and the monophyletic relationship of other chordate lipocalins.

Thus, two sets of independent characters have produced the same phylogenetic relationships between extant lipocalins.

Phylogenetic Distribution of Intron Numbers Within the Lipocalin Family

Another finding revealed by the intron arrangement phylogeny is that lipocalins that have originated more recently contain more introns in their CDS. In figure 5*C* we mapped the number of exons onto an updated version of the ML-based lipocalin protein phylogeny (see Gutiérrez, Ganfornina, and Sánchez [2000] for clade ascription). The ancient unicellular eukaryotic and plant lipocalins are encoded by 2 exons; the arthropodan lipocalins by 4–5 exons; and the chordate lipocalins by 4–7 exons. Introns E and F are absent in nonchordate lipocalins, whereas introns A–D show much wider phylogenetic distributions.

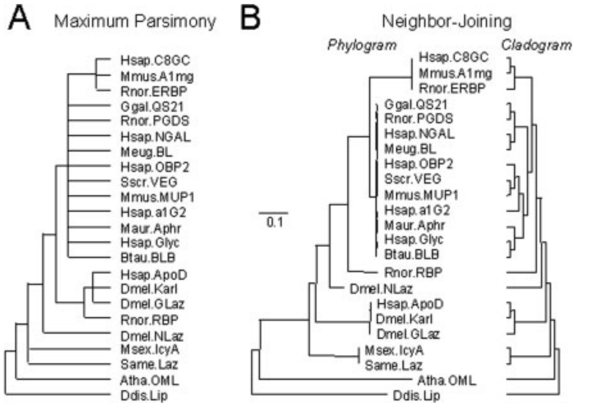

FIG. 4.—Phylogenetic trees derived from lipocalin gene structure information. *A*, Maximum parsimony analysis based on the presence-absence of introns in the gene ORF. *B*, Neighbor-Joining (NJ) phylogenetic reconstruction based on a distance matrix with gene structure data (intron presence, location, and phase) combined in a measure of gene structure similarity (see *Materials and Methods*). The NJ tree is shown both as a phylogram (left) and as a cladogram (right). All trees are rooted with the Dictyostelium lipocalin. The scale bar in the phylogram represents branch length (number of amino acid substitutions/site).

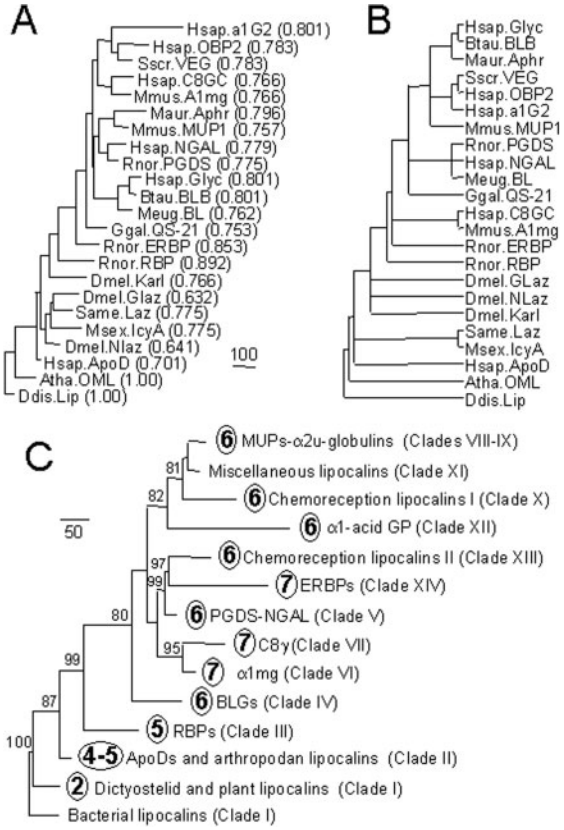

FIG. 5.—*A–B*, Comparisons of gene structure versus protein sequence phylogenies. *A*, Phylogenetic ML tree built upon the protein sequence alignment used to derive the gene structure matrices (see fig. 3), and rooted with the dictyostelid lipocalin. Explicitly agreed similarity values are shown for each lipocalin, according to the RadCon program (see *Materials and Methods*). *B*, Consensus tree obtained by a strict nesting method (Adams 1972), representing the commonality between the gene structure and the protein sequence trees. *C*, Updated (March 2002) maximum-likelihood tree based on the protein sequence of 148 lipocalins (see Ganfornina et al. [2000] for details on tree building) showing the number of introns present in the ORF of each lipocalin clade. LBP values are indicated in each node (see *Materials and Methods*). The tree was rooted with bacterial lipocalins. The scale bars represent branch length (number of amino acid substitutions/100 residues).

A first look at these data might suggest an evolutionary trend to gain introns. In this hypothesis, the origin of introns A and D could be placed early in eukaryotic evolution, introns B and C originated at the base of the metazoan lineage, and introns E and F appeared later during early chordate radiation. The acquisition of introns was accompanied by diverse intron losses in different branches, giving rise to the pattern observed today.

However, we have to be critical when interpreting these observations in the context of lipocalin introns origin and evolution. First, the current insufficient sampling of lipocalins outside the metazoan kingdom generates uncertainty about the very assumption of homology of introns A and D in Dictyostelium and Arabidopsis, respectively. Second, the set of metazoan lipocalins available encompasses only two phyla within the kingdom; any proposal about which set of introns was present in the common ancestor of all metazoans awaits confirmation coming from other phyla. A scenario with a set of four ancient introns and subsequent losses in different lineages (Fedorov et al. 2001; Roy et al. 2002) would be as probable as a scenario with fewer or no ancient introns and a prevalence of intron gain at preferred ‘‘hot spots’’ (or proto-splice sites; Dibb and Newman 1989; Logsdon 1998).

Nevertheless, the extensive sampling of lipocalins in the chordate phylum allows us to make a stronger case for the acquisition of introns E and F during early chordate radiation. Both introns are absent in all arthropod lipocalins and in ApoD, the lipocalin COG that branches off at the base of the chordate lipocalin subtree in our two independent phylogenetic reconstructions (figs. 4*B* and 5*C*). Therefore, intron gain within the chordate lineage is the most parsimonious explanation for the current distribution of introns E and F.

In summary, although many questions about the origin of lipocalin introns and their subsequent evolution remain unanswered, a combination of ancient and recent introns is the most plausible scenario. Our results show that, independent of their origin, the variations in gene structure can be used to reconstruct the history of descent of lipocalin genes.

N-termini Conservation versus C-Termini Variability?

It is remarkable that the introns specific to chordate lipocalins are located in the C-termini of the proteins, whereas introns in their N-terminal region are the most conserved in the family (see fig. 3). This polarity, also noticed by other researchers (Salier 2000), is related neither to a particular distribution of lipocalins length nor to a C-terminal–specific protein sequence variability. Rather, we propose it might be related to a propensity for intron gain/loss in this gene region. The analysis of the 3′ region

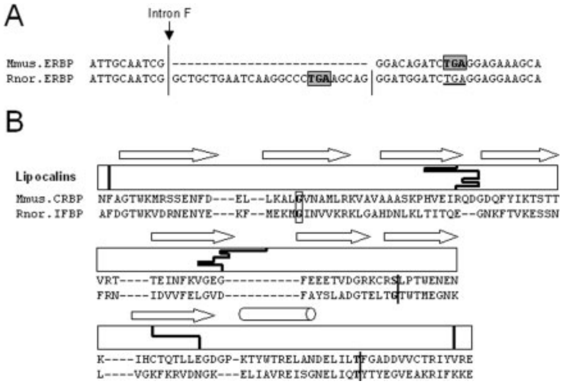
A
Intron F
Mmus.ERBP ATTGCAATCG ------------------------- GGACAGATCTGAGGAGAAAGCA
Rnor.ERBP ATTGCAATCG GCTGCTGAATCAAGGCCCTGAAGCAG GGATGGATCTGAGGAGGAAGCA
B
Lipocalins
Mmus.CRBP NFAGTWKMRSSENFD---EL--LKALGVNAMLRKVAVAAASKPHVEIRQDGDQFYIKTSTT
Rnor.IFBP AFDGTWKVDRNENYE---KF--MEKMGINVVKRKLGAHDNLKLTITQE--GNKFTVKESSN
VRT----TEINFKVGEG----------FEEETVDGRKCRSLPTWENEN
FRN----IDVVFELGVD----------FAYSLADGTELTGTWTMEGNK
K----IHCTQTLLEGDGP-KTYWTRELANDELILTFGADDVVCTRIYVRE
L----VGKFKRVDNGK---ELIAVREISGNELIQTYTYEGVEAKRIFKKE

FIG. 6.—*A*, Alignment of the nucleotide sequences coding for the C-terminal region of the ERBP genes from mouse (Mmus.ERBP) and rat (Rnor.ERBP). Stop codons are marked by gray boxes. Intron insertions are shown by vertical lines. A conserved TGA codon is present (underlined) in the 3′-UTR of the rat gene. *B*, Alignment of the protein sequences of two representative FABP and CRBP proteins with the lipocalin alignment of figure 3. Intron insertions of FABP and CRBP (intron phase shown as in fig. 3) are shown in the context of lipocalin introns and secondary structure. For the sake of simplicity, the lipocalins are represented by a white box and the intron positions (not considering their phase) are highlighted with a thick black line.

of lipocalins bearing 5–6 introns reveals that several lipocalins of particular chordate lineages (e.g., PGDS, VEG, NGAL; data not shown) show introns in the 3′-UTR that are located 0–7 nucleotides away from the stop codon. These introns would be equivalent to intron F if they happened to be in the CDS. Any form of intron sliding (Stoltzfus et al. 1997), or any frameshifting mutation that moves the stop codon in the 3′ direction, could include/exclude a given intron in the gene CDS, generating an apparent intron gain/loss. Moreover, a puzzling case of C-terminus variability comes from the comparison of mouse and rat ERBP (fig. 6*A*). Intron F of mouse ERBP locates 9 nucleotides away from the stop codon. The rat ERBP gene has an insertion that accommodates a short exon and another intron (alternatively, the mouse ERBP could have experienced an equivalent deletion). Were it not for the existence of an in-frame stop codon in the short exon present in the rat ERBP gene, we would have a unique lipocalin with 7 introns.

In summary, the C-termini of chordate lipocalins show genomic plasticity, accommodating introns and mutations that modify the protein length. It is not known whether this genomic plasticity is causally related to a higher probability of intron gain, loss, or sliding in the 3′ end of lipocalins, but this possibility is worth investigating.

Gene Structure and the Three-Dimensional Structure of Lipocalins and Calycins

We mapped the location of exon boundaries in the tertiary structure of lipocalins that belong to different phylogenetic clades (IcyA, RBP, BLB, NGAL, MUP, and ERBP). Most lipocalin introns are located in the boundaries of β strands (see fig. 3). In spite of a certain variability in number and position, introns A–D seem to demarcate the lipocalin β barrel, whereas introns E and F are present in the C-terminal flexible region.

A way of testing a relationship between lipocalin intron-exon boundaries and tertiary structure would be to analyze the gene structure of proteins with a tertiary structure like that of lipocalins. A similar structure and a marginal sequence similarity have been used to propose a structural superfamily, the calycins, that relates lipocalins to proteins such as FABP, CRBP, avidin, and a group of protease inhibitors (Flower, North, and Sansom 2000). We find no gene structure similarity after aligning representatives of these proteins with the lipocalins and comparing intron positions (fig. 6*B*). This finding suggests that (1) we do not have compelling evidence for a relationship between intron-exon arrangement and the tertiary structure of these β barrel–based proteins; (2) the evolutionary relationship of lipocalins with the other proposed calycins, already questioned after analyzing their protein sequence (Ganfornina et al. 2000), remains to be demonstrated; and (3) the homology of introns A–F in lipocalins, the foundation for the phylogenetic inferences that we present in this work, is a reasonable assumption: the pattern and properties of intron-exon boundaries are good markers of the lipocalins history of descent.

Concluding Remarks: Evolutionary Hypothesis for the Lipocalin Gene Family

In conclusion, gene structure is well preserved among lipocalins, and our results validate its use for the reconstruction of lipocalin evolution. The congruence of phylogenetic trees built from two independent sets of data (protein sequence and gene structure) increases the verisimilitude of both reconstructions of the lipocalins history. Furthermore, in the future we can use gene structure data to assay the lipocalin nature of novel proteins whose amino acid sequence and/or protein structure show similarity to lipocalins.

Our results give support to the following hypothesis about the evolutionary history of lipocalins: Bacterial lipocalins were inherited by unicellular eukaryotes and passed on to both plants and metazoans. The primitive metazoans spread a low number of ancient lipocalins into some of their successors, the arthropods and chordates, although these proteins might have been unexploited and subsequently lost in other phyla. The primordial arthropod and chordate lipocalins were likely similar to the Lazarillo and ApoD lipocalins now present in these phyla. Alongside the chordate radiation, the ApoD-like ancestral lipocalin suffered duplications. On the one hand, it gave rise to the ancestor of RBPs, and on the other hand, to one or more ancestors of all other paralogous groups of lipocalins that diverged into the current diverse catalog of chordate lipocalins.

Acknowledgments

We thank C. González for allowing us to use his lab resources to collect the Lazarillo gene structure data needed for this work, and M. Bastiani for the grasshopper genomic DNA. We thank J. P. Salier for helpful discussion and comments on the manuscript. This work was supported by a grant (DGICYT BIO-1999-065-C02-02) to A. M., and a visiting professor grant to D.S. We are grateful to R. Escalante for alerting us to the sequence of the Dictyostelium gene, and to B. Akerstrom for alerting us to the Drosophila gene Karl. We thank C. Santiago at the Centro de Investigaciones Biológicas sequencing facility.

Literature Cited

Adachi, J., and M. Hasegawa. 1996. MOLPHY version 2.3: programs in molecular phylogenetics based on maximum likelihood. The Institute of Statistical Mathematics, Tokyo.

Adams, E. N. 1972. Consensus techniques and the comparison of taxonomic trees. Syst. Zool. **21**:390–397.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403–410.

Betts, M. J., R. Guigo, P. Agarwal, and R. B. Russell. 2001. Exon structure conservation despite low sequence similarity: a relic of dramatic events in evolution? EMBO J. **20**:5354–5360.

Bishop, R. E. 2000. The bacterial lipocalins. Biochim. Biophys. Acta **1482**:73–83.

Caterino, M. S., S. Cho, and F. A. Sperling. 2000. The current state of insect molecular systematics: a thriving tower of Babel. Annu. Rev. Entomol. **45**:1–54.

Dibb N. J., and A. J. Newman. 1989. Evidence that introns arose at proto-splice sites. EMBO J. **8**:2015–2021.

Estabrook, G. F. 1992. Evaluating undirected positional congruence of individual taxa between two estimates of the phylogenetic tree for a group of taxa. Syst. Biol. **41**:172–177.

Fedorov, A., X. Cao, S. Saxonov, S. J. de Souza, S. W. Roy, and W. Gilbert. 2001. Intron distribution difference for 276 ancient and 131 modern genes suggests the existence of ancient introns. **98**:13177–13182.

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

Flower, D. R., A. C. T. North, and T. K. Attwood. 1993. Structure and sequence relationships in the lipocalins and related proteins. Protein Sci. **2**:753–761.

Flower D. R., A. C. North, and C. E. Sansom. 2000. The lipocalin protein family: structural and sequence overview. Biochim. Biophys. Acta **1482**:9–24.

Ganfornina, M. D., G. Gutiérrez, M. J. Bastiani, and D. Sánchez. 2000. A phylogenetic analysis of the lipocalin protein family. Mol. Biol. Evol. **17**:114–126.

Ganfornina, M. D., D. Sánchez, and M. J. Bastiani. 1995. Lazarillo, a new GPI-linked surface lipocalin, is restricted to a subset of neurons in the grasshopper embryo. Development **121**:123–134.

Gutiérrez, G., M. D. Ganfornina, and D. Sánchez. 2000. Evolution of the lipocalin family as inferred from a protein sequence phylogeny. Biochim. Biophys. Acta **1482**:35–45.

Hasegawa, M., and H. Kishino. 1994. Accuracies of the simple methods for estimating the bootstrap probability of a maximum likelihood tree. Mol. Biol. Evol. **11**:142–145.

Holzfeind, P., and B. Redl. 1994. Structural organization of the gene encoding the human lipocalin tear prealbumin and synthesis of the recombinant protein in Escherichia coli. Gene **139**:177–183.

Igarashi, M., A. Nagata, H. Toh, Y. Urade, and O. Hayaishi. 1992. Structural organization of the gene for prostaglandin D synthase in the rat brain. Proc. Natl. Acad. Sci. USA **89**:5376–5380.

Krem, M. M., and E. Di Cera. 2001. Molecular markers of serine protease evolution. EMBO J. **20**:3036–3045.

Li, W., and L. M. Riddiford. 1992. Two distinct genes encode two major isoelectric forms of insecticyanin in the tobacco hornworm, *Manduca sexta*. Eur. J. Biochem. **205**:491–499.

Lindqvist, A., P. Rouet, J.-P. Salier, and B. Akerstrom. 1999. The alpha1-microglobulin/bikunin gene: characterization in mouse and evolution. Gene **234**:329–336.

Logsdon, J. M. Jr. 1998. The recent origins of spliceosomal introns revisited. Curr. Opin. Genet. Dev. **8**:637–648.

Ohno, S. 1999. Gene duplication and the uniqueness of vertebrate genomes circa 1970–1999. Semin. Cell Dev. Biol. **10**:517–522.

Rokas, A., and P. W. Holland. 2000. Rare genomic changes as a tool for phylogenetics. Trends Ecol. Evol. **15**:454–459.

Rokas, A., J. Kathirithamby, and P. W. Holland. 1999. Intron insertion as a phylogenetic character: the engrailed homeobox of Strepsiptera does not indicate affinity with Diptera. Insect Mol. Biol. **8**:527–530.

Roy, S. W., A. Fedorov, and W. Gilbert. 2002. The signal of ancient introns is obscured by intron density and homolog number. Proc. Natl. Acad. Sci. USA **99**:15513–15517

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

Salier, J.-P. 2000. Chromosomal location, exon/intron organization and evolution of lipocalin genes. Biochim. Biophys. Acta **1482**:25–34.

Sánchez, D., M. D. Ganfornina, and M. J. Bastiani. 1995. Developmental expression of the lipocalin Lazarillo and its role in axonal pathfinding in the grasshopper embryo. Development **121**:135–147.

———. 2000*a*. Lazarillo, a neuronal lipocalin in grasshoppers with a role in axon guidance. Biochim. Biophys. Acta **1482**:102–109.

Sánchez, D., M. D. Ganfornina, S. Torres-Schumann, S. D. Speese, J. M. Lora, and M. J. Bastiani. 2000*b*. Characterization of two novel lipocalins expressed in the Drosophila embryonic nervous system. Int. J. Dev. Biol. **44**:349–359.

Stoltzfus, A., J. M. Logsdon, Jr., J. D. Palmer, and W. F. Doolittle. 1997. Intron ''sliding'' and the diversity of intron positions. Proc. Natl. Acad. Sci. USA **94**:10739–10744.

Swofford, D. L. 1998. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Suderland, Mass.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The ClustalX Windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tool. Nucleic Acids Res. **24**:4876–4882.

Thorley J. L., and R. D. M. Page. 2000. RadCon: phylogenetic tree comparison and consensus. Bioinformatics **16**:486–487.

Thorley, J. L., M. Wilkinson, and M. A. Charleston. 1998. The information content of consensus trees. Pp. 91–98 *in* A. Rizzi, M. Vichi, and H.-H. Bock, eds. Advances in data science and classification. Springer-Verlag, Berlin.

Toh, H., H. Kubodera, N. Nakajima, T. Sekiya, N. Eguchi, T. Tanaka, Y. Urade, and O. Hayaishi. 1996. Glutathione-independent prostaglandin D synthase as a lead molecule for designing new functional proteins. Protein Eng. **9**:1067–1082.

Wada, H., M. Kobayashi, R. Sato, N. Satoh, H. Miyasaka, and Y. Shirayama. 2002. Dynamic insertion-deletion of introns in deuterostome EF-1 alpha genes. J. Mol. Evol. **54**:118–128.

Peer Bork, Associate Editor

Accepted December 23, 2002