

A Phylogenetic Analysis of the Lipocalin Protein Family

Maña D. Ganfornina,* Gabriel Gutiérrez,† Michael Bastiani,* and Diego Sánchez*

*Department of Biology, University of Utah; and †Departamento de Genética, Universidad de Sevilla, Seville, Spain

The lipocalins are a family of extracellular proteins that bind and transport small hydrophobic molecules. They are found in eubacteria and a great variety of eukaryotic cells, in which they play diverse physiological roles. We report here the detection of two new eukaryotic lipocalins and a phylogenetic analysis of 113 lipocalin family members performed with maximum-likelihood and parsimony methods on their amino acid sequences. Lipocalins segregate into 13 monophyletic clades, some of which are grouped in well-supported superclades. An examination of the G+C content of the bacterial lipocalin genes and the detection of four new conceptual lipocalins in other eubacterial species argue against a recent horizontal transfer as the origin of prokaryotic lipocalins. Therefore, we rooted our lipocalin tree using the clade containing the prokaryotic lipocalins. The topology of the rooted lipocalin tree is in general agreement with the currently accepted view of the organismal phylogeny of arthropods and chordates. The rooted tree allows us to assign polarity to character changes and suggests a plausible scenario for the evolution of important lipocalin properties. More recently evolved lipocalins tend to (1) show greater rates of amino acid substitutions, (2) have more flexible protein structures, (3) bind smaller hydrophobic ligands, and (4) increase the efficiency of their ligand-binding contacts. Finally, we found that the family of fatty-acid-binding proteins originated from the more derived lipocalins and therefore cannot be considered a sister group of the lipocalin family.

Introduction

The transport and storage of hydrophobic molecules in metazoans is generally achieved by proteins that bear a buried structural pocket as a binding site. Among the proteins that display this structural motif are the lipocalins, a heterogeneous group of secreted proteins that bind a wide variety of small hydrophobic ligands (reviewed by Flower 1996). The lipocalins are small proteins of around 200 residues, with molecular masses averaging 20 kDa. Most lipocalins exhibit an N-terminal signal peptide and lack other strong hydrophobic regions, features commonly found in extracellular soluble proteins. Similarly, most lipocalins contain from one to three disulfide bridges that contribute to constraint of the overall structure by stabilizing the N- and C-terminal regions of the protein.

The amino acid sequences of lipocalins are quite divergent, and low levels of sequence identity, even below 20%, are found when comparing the overall sequence among some members of the family. In spite of the low level of sequence similarity, the tertiary structures of lipocalins are strongly preserved. The lipocalin folding motif (fig. 1) (Cowan, Newcomer, and Jones 1990; Flower 1995) is an eight-stranded antiparallel β -barrel with an N-terminal 3_{10} α -helix and a C-terminal α -helix (A1 and A2 in fig. 1, respectively). The barrel is open at one side and encloses a binding pocket. Another characteristic of the lipocalins is their ability to form oligomers, which range from the dimeric state of many lipocalins, such as odorant-binding proteins (Tegoni et al. 1996), to the complex octamers of crustacy-

anins (Keen et al. 1991). There are three conserved sequence motifs called structurally conserved regions (SCRs) that have been proposed as a prerequisite for a protein to be considered a lipocalin (Flower, North, and Atkwood 1993). Flower, North, and Atkwood (1993) propose a separation of kernel versus outlier lipocalins based on the conservation of the SCRs, as well as on the existence of disulfide bridges. The SCRs represent a structural element composed of three loops that are close to each other in the three-dimensional structure and constitute the bottom of the β -barrel. A role for the SCRs as a receptor-binding site has been suggested based on their exposure to the solvent. This proposition, however, remains to be demonstrated. With regard to ligands, a broad set of hydrophobic molecules have been shown to bind to different lipocalins. Some lipocalins have an exquisite specificity for a given ligand, such as the epididymal secretory proteins that bind retinoic acid, but not other retinoids (Newcomer and Ong 1990). Other lipocalins, like β -lactoglobulins, apolipoproteins D (ApoD's), and some chemoreceptor lipocalins, bind a variety of ligands of very different natures (reviewed by Flower 1995).

Abundant information is available about structural, biochemical, and functional aspects of lipocalins. However, few studies have reported on the phylogenetic relationships and evolutionary history of the lipocalin family. Some of these phylogenetic studies have been carried out by choosing exemplars of the functionally different groups of lipocalins (Igarashi et al. 1992), while others have focused on the phylogenetic relationships of restricted lipocalin clades (Ganfornina, Sánchez, and Bastiani 1995; Pottie et al. 1998). Two main features of the lipocalins might account for the lack of a comprehensive phylogenetic analysis on such a prolific family: (1) a strongly divergent protein sequence, which denotes a rapid rate of molecular evolution and makes kinship attributions difficult to resolve, and (2) an evolutionary history rich in gene duplications, which increases the difficulty of obtaining an understanding of orthologous relationships.

Abbreviations: GPI, glycosyl-phosphatidylinositol; LBP, local bootstrap probability; ORF, open reading frame; RT-PCR, reverse transcriptase polymerase chain reaction; SCR, structurally conserved region.

Key words: lipocalin, calycin, molecular evolution, protein phylogeny.

Address for correspondence and reprints: Diego Sánchez, Department of Biology, University of Utah, 257 South 1400 East, Salt Lake City, Utah 84112-0840. E-mail: sanchez@bioscience.utah.edu.

Mol. Biol. Evol. 17(1):114–126. 2000

© 2000 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

Table 1
List of Proteins Used for Amino Acid Sequence Alignments

Protein	Species	Taxon ^a	Abbreviation	Accession No.	Clade
Outer membrane lipoprotein	<i>Citrobacter freundii</i>	B	Cfre.OML	U21727	I
	<i>Escherichia coli</i>	B	Ecol.OML	AE000487	I
	<i>Vibrio cholerae</i>	B	Vcho.Lpro	X64097	I
Putative lipocalin	<i>Dictyostelium discoideum</i>	P	Ddis.Lip	C24642	I
Lipocalin	<i>Drosophila melanogaster</i>	A	Dmel.Lip	AC004439	II
Gallerin	<i>Galleria mellonella</i>	A	Gmel.Gall	L41641	II
Crustacyanin 1	<i>Homarus gammarus</i>	A	Hgam.CRC1	P80029	II
Crustacyanin 2	<i>H. gammarus</i>	A	Hgam.CRC2	P80007	II
Insecticyanin A	<i>Manduca sexta</i>	A	Msex.IcyA	X64714	II
Insecticyanin B	<i>M. sexta</i>	A	Msex.IcyB	X64715	II
Bilin-binding protein	<i>Pieris brassicae</i>	A	Pbra.Bbp	X76568	II
Lazarillo	<i>Schistocerca americana</i>	A	Same.Laz	U15656	II
Apolipoprotein D	<i>Cavia cobaya</i>	PM	Ccob.ApoD	S79402	II
	<i>Homo sapiens</i>	PM	Hsap.ApoD	J02611	II
	<i>Mus musculus</i>	PM	Mmus.ApoD	X82648	II
	<i>Oryctolagus cuniculus</i>	PM	Ocun.ApoD	L42979	II
	<i>Rattus norvegicus</i>	PM	Rnor.ApoD	X55572	II
Retinol-binding protein	<i>Bos taurus</i>	PM	Btau.RBP	S65585	III
	<i>Equus caballus</i>	PM	Ecab.RBP	U21208	III
Purpurin	<i>Gallus gallus</i>	Bi	Ggal.Purp	M17538	III
Retinol-binding protein	<i>G. gallus</i>	Bi	Ggal.RBP	X77960	III
	<i>H. sapiens</i>	PM	Hsap.RBP	X00129	III
	<i>M. musculus</i>	PM	Mmus.RBP	U63146	III
	<i>Ovis aries</i>	PM	Oari.RBP	L11346	III
Retinol-binding protein 1	<i>Oncorhynchus mykiss</i>	F	Omyc.RBP1	P24774	III
Retinol-binding protein 2	<i>O. mykiss</i>	F	Omyc.RBP2	P24775	III
Retinol-binding protein	<i>O. cuniculus</i>	PM	Ocun.RBP	S45958	III
	<i>Sus scrofa</i>	PM	Sscr.RBP	M68860	III
	<i>Xenopus laevis</i>	Am	Xlae.RBP	J02718	III
Beta-lactoglobulin B	<i>Bos bubalus</i>	PM	Bbub.BLB	P02755	IV
	<i>B. taurus</i>	PM	Btau.BLB	X14712	IV
Beta-lactoglobulin A	<i>Canis familiaris</i>	PM	Cfam.BLA	P33685	IV
Beta-lactoglobulin C	<i>C. familiaris</i>	PM	Cfam.BLC	P33686	IV
Beta-lactoglobulin B	<i>Capra hircus</i>	PM	Chir.BLB	Z19569	IV
Beta-lactoglobulin A	<i>Equus asinus</i>	PM	Easi.BLA	P19647	IV
Beta-lactoglobulin B	<i>E. asinus</i>	PM	Easi.BLB	P13613	IV
Beta-lactoglobulin I	<i>Equus caballus</i>	PM	Ecab.BL1	U60356	IV
Beta-lactoglobulin II	<i>E. caballus</i>	PM	Ecab.BL2	U60978	IV
Beta-lactoglobulin A	<i>Felis catus</i>	PM	Fcat.BLA	P21664	IV
Beta-lactoglobulin B	<i>F. catus</i>	PM	Fcat.BLB	P33687	IV
Beta-lactoglobulin C	<i>F. catus</i>	PM	Fcat.BLC	P33688	IV
Beta-lactoglobulin	<i>Macropus giganteus</i>	MM	Mgig.BL	P11944	IV
Beta-lactoglobulin	<i>Macropus eugenii</i>	MM	Meug.BL	L14954	IV
Beta-lactoglobulin B	<i>O. aries</i>	PM	Oari.BLB	X04520	IV
	<i>S. scrofa</i>	PM	Sscr.BLB	X54976	IV
Beta-lactoglobulin	<i>Trichosurus vulpecula</i>	MM	Tvul.BL	U34289	IV
Pregnancy protein 14	<i>H. sapiens</i>	PM	Hsap.PP14	M61886	IV
Prostaglandin D synthase	<i>B. taurus</i>	PM	Btau.PGDS	AB004647	V
	<i>F. catus</i>	PM	Fcat.PGDS	D82048	V
	<i>H. sapiens</i>	PM	Hsap.PGDS	M61900	V
	<i>M. musculus</i>	PM	Mmus.PGDS	X89222	V
	<i>R. norvegicus</i>	PM	Rnor.PGDS	J04488	V
	<i>S. scrofa</i>	PM	Sscr.PGDS	X92979	V
	<i>Ursus arctos</i>	PM	Uarc.PGDS	D82047	V
Neutrophil gelatinase lipocalin	<i>H. sapiens</i>	PM	Hsap.NGAL	X83006	V
	<i>M. musculus</i>	PM	Mmus.NGAL	P11672	V
	<i>R. norvegicus</i>	PM	Rnor.NGAL	X13295	V
Quiescence-specific protein 20	<i>Gallus gallus</i>	Bi	Ggal.QS20	M25784	V
Quiescence-specific protein 21	<i>G. gallus</i>	Bi	Ggal.QS21	M55644	V
Choroid plexus lipocalin	<i>Bufo marinus</i>	Am	Bmar.Lip	Q01584	V
Choroid plexus lipocalin 1	<i>Xenopus laevis</i>	Am	Xlae.Cp11	X84414	V
Alpha-1 microglobulin	<i>B. taurus</i>	PM	Btau.A1mg	U35642	VI
	<i>H. sapiens</i>	PM	Hsap.A1mg	X04494	VI
	<i>Mesocricetus auratus</i>	PM	Maur.A1mg	D31814	VI
	<i>M. musculus</i>	PM	Mmus.A1mg	D28812	VI
	<i>Meriones unguiculatus</i>	PM	Mung.A1mg	D31813	VI
	<i>Pleuronectes platessa</i>	F	Ppla.A1mg	X63762	VI
	<i>R. norvegicus</i>	PM	Rnor.A1mg	S87544	VI
	<i>Salmo salar</i>	F	Ssal.A1mg	L26598	VI
	<i>S. scrofa</i>	PM	Sscr.A1mg	X53685	VI

Table 1
Continued

Protein	Species	Taxon ^a	Abbreviation	Accession No.	Clade
Complement C8γ subunit	<i>X. laevis</i>	Am	Xlae.A1mg	D87752	VI
	<i>H. sapiens</i>	PM	Hsap.C8GC	X06465	VII
	<i>O. cuniculus</i>	PM	Ocun.C8GC	L26979	VII
Major urinary protein 1	<i>M. musculus</i>	PM	Mmus.MUP1	M16355	VIII
Major urinary protein 2	<i>M. musculus</i>	PM	Mmus.MUP2	M16356	VIII
Major urinary protein 4	<i>M. musculus</i>	PM	Mmus.MUP4	M16358	VIII
Major urinary protein 5	<i>M. musculus</i>	PM	Mmus.MUP5	M16360	VIII
Major urinary protein 6	<i>M. musculus</i>	PM	Mmus.MUP6	P02762	VIII
Major urinary protein	<i>M. musculus</i>	PM	Mmus.MUP	M28649	VIII
Minor-major urinary protein	<i>M. musculus</i>	PM	Mmus.mMUP	X03525	VIII
Alpha 2-globulin (L type)	<i>R. norvegicus</i>	PM	Rnor.a2g1	M26836	IX
Alpha 2-globulin (S type)	<i>R. norvegicus</i>	PM	Rnor.a2g2	M26837	IX
Submaxillary alpha 2-globulin	<i>R. norvegicus</i>	PM	Rnor.a2g3	J00738	IX
Prostate protein Probasin	<i>M. musculus</i>	PM	Mmus.Pbas	AF005204	X
	<i>R. norvegicus</i>	PM	Rnor.Pbas	M27156	X
Pheromone carrier Aphrodisin	<i>Cricetus cricetus</i>	PM	Ccri.Aphr	X65238	X
Allergen BDA20	<i>B. taurus</i>	PM	Btau.Alle	L42867	X
Odorant-binding protein	<i>B. taurus</i>	PM	Btau.OBP	P07435	X
Odorant-binding protein 1	<i>R. norvegicus</i>	PM	Rnor.OBP1	J03093	X
Endometrial P19 protein	<i>E. caballus</i>	PM	Ecab.p19p	X98459	XI
Allergen f2	<i>C. familiaris</i>	PM	Cfam.f2p	AF027178	XI
Lactation protein Trichosurin	<i>T. vulpecula</i>	MM	Tvul.Lip	U40376	XI
Alpha-1 acid glycoprotein 1	<i>H. sapiens</i>	PM	Hsap.a1GP	M13692	XII
Alpha-1 acid glycoprotein 2	<i>H. sapiens</i>	PM	Hsap.a1G2	M21540	XII
Alpha-1 acid glycoprotein 2	<i>M. musculus</i>	PM	Mmus.a1GP	M27009	XII
Alpha-1 acid glycoprotein	<i>Mus caroli</i>	PM	Mcar.a1GP	M34649	XII
	<i>O. cuniculus</i>	PM	Ocun.a1GP	X58727	XII
	<i>R. norvegicus</i>	PM	Rnor.a1GP	J00696	XII
Odorant-binding protein 2	<i>R. norvegicus</i>	PM	Rnor.OBP2	A40464	XIII
von Ebner's gland protein	<i>H. sapiens</i>	PM	Hsap.VEG	M90424	XIII
	<i>S. scrofa</i>	PM	Sscr.VEG	S77587	XIII
von Ebner's gland protein 1	<i>R. norvegicus</i>	PM	Rnor.VEG1	P20289	XIII
von Ebner's gland protein 2	<i>R. norvegicus</i>	PM	Rnor.VEG2	X74806	XIII
Vomerolateral secretory protein 1	<i>M. musculus</i>	PM	Mmus.VNS1	D38580	XIII
Vomerolateral secretory protein 2	<i>M. musculus</i>	PM	Mmus.VNS2	D38581	XIII
Allergen f1	<i>C. familiaris</i>	PM	Cfam.f1p	AF027177	XIII
Late lactation protein	<i>M. eugenii</i>	MM	Meug.LLP	X15213	XIII
Late lactation protein	<i>T. vulpecula</i>	MM	Tvul.LLP	U34287	XIII
Allergen IV	<i>Blattella germanica</i>	A	Bger.All4	U40767	—
Olfactory-specific protein	<i>Rana pipiens</i>	Am	Rpip.OP	M15531	—
Epididymal secretory protein	<i>Lacerta vivipara</i>	R	Lviv.ESP	X63151	—
Epididymal secretory protein	<i>R. norvegicus</i>	PM	Rnor.ESP	M12790	—
Allergen C1	<i>E. caballus</i>	PM	Ecab.C1p	U70823	—

^a Abbreviations: A, arthropod; Am, amphibian; B, bacteria; Bi, bird; F, fish; MM, marsupial mammal; PM, placental mammal; P, protoctist; R, reptile.

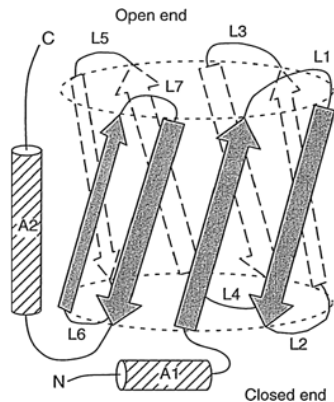


FIG. 1.—Schematic diagram of the topology of the lipocalin structural fold.

In this paper, we report the detection of two novel lipocalins in different organisms and the analysis of the phylogenetic relationships of all the proteins so far inscribed in the lipocalin family.

Materials and Methods

Sequence Searches and Alignments

A search for lipocalins was performed using the BLAST program (Altschul et al. 1990) at the NCBI web site on all the databases available as of February 13, 1998. We started our search from proteins that matched the lipocalin pattern in Prosite (PS00213). Subsequently, the mature protein sequences of randomly chosen representatives of lipocalins belonging to functionally different groups were used as queries to find similar sequences in the databases. A second run of searches was done using the less similar sequence (although with an unambiguous lipocalin character) found on the first run.

After establishing significant local similarities, the entire protein sequences were retrieved (or deduced from the DNA sequence) and subjected to a pairwise comparison with each query lipocalin sequence. The criteria used for a sequence to be included in the lipocalin family were overall sequence identities with its closest relative above 20%, the existence of at least two of the structurally conserved lipocalin motifs (SCRs; see *Introduction*), and a mature protein length in agreement with the previously known lipocalins.

Only the mature protein sequences were used, either deduced from the known mature N-terminal amino acid or predicted by von Heijne's (1990) method. Similarly, the C-terminal peptide cleaved after the attachment of glycosyl-phosphatidylinositol (GPI) to a particular lipocalin was not considered for the alignment. Partial sequence entries were excluded. The most recent entry was considered when several sequences were available for a given protein. Proteins were named using an abbreviated species name followed by a functional label (see table 1).

Protein sequences were aligned with CLUSTAL W, version 1.7 (Thompson, Higgins, and Gibson 1994), using a PAM series scoring matrix and a gap penalty mask based on the aligned secondary structures of the lipocalins whose crystal structures were known. Minor manual corrections were performed on the alignment based on the knowledge of lipocalin structure and function. The same strategy was used for each separate alignment of the different lipocalin clades (see below). The final alignments used for the phylogenetic studies are available from the authors on request.

Phylogenetic Analyses

Phylogenetic analyses based on protein sequences were carried out using the maximum-likelihood method with the MOLPHY software, version 2.3 (Adachi and Hasegawa 1996). Parsimony analyses were performed with the PROTPARS program of the PHYLIP software package, version 3.5 (Felsenstein 1993).

The number of sequences is the most important limiting factor when an exhaustive phylogenetic analysis is attempted under the maximum-likelihood principles. In our study, the 113 sequences of our sample make an exhaustive tree topology search prohibitive. Therefore, we devised a different strategy. First, we reconstructed a global tree based on the protein sequences given in table 1. The steps for obtaining the global tree were: (1) calculation of a maximum-likelihood distance matrix with the PROTML program under the JTT model (Jones, Taylor, and Thornton 1992) normalized to the amino acid composition of the data set (program options D and jf), (2) a neighbor-joining tree reconstruction (Saitou and Nei 1987) with the program NJDIST using the previous distance matrix as an input, (3) the use of the resulting tree topology as a seed to search for a topology with higher likelihood value using the same amino acid substitution model. We used the method of topology search by local rearrangement (option R of the PROTML program) to try to improve the neighbor-joining tree and to calculate a local bootstrap probability

(LBP) under 1,000 replicates (Hasegawa and Kishino 1994).

We then divided the global tree into subtrees following two criteria: an LBP greater than 75%, and literature information on different functional groups of lipocalins. The sequences composing each group were aligned de novo using the lipocalin structural mask described above. We also selected for each group the sequence in the most related subtree with the shortest distance to the node joining both groups. This sequence was used as an outgroup for the purpose of tree representation. Two different methods were used to analyze the defined groups, depending on the number of sequences. Exhaustive topology searches were carried out for groups with eight or less sequences (including the outgroup). This approach renders the maximum-likelihood tree under the JTT model with data frequencies. The LBP under 1,000 replicates was also calculated for each node. In groups of more than eight sequences, we followed the same strategy as that used in estimating the global tree. A majority-rule of $\text{LBP} \geq 50\%$ was established for each node in every subtree; unsupported nodes were excluded, and their branches forced to yield polytomies.

Results and Discussion

Novel Protein Sequences

Two new protein sequences were included in this study, since they met the criteria (outlined above) for belonging to the lipocalin family.

In our searches, we detected a conceptual protein deduced from a cDNA clone obtained at the slug stage of the cellular slime mold *Dictyostelium discoideum*. The DNA sequence came from the *Dictyostelium* cDNA project in Japan, but no information is currently available concerning the properties and role of the protein. This putative lipocalin displays significant sequence similarity in the regions aligned with SCRs 1 and 3 of the kernel lipocalins and shows high overall sequence similarity with the prokaryotic lipocalins (fig. 2).

We found another conceptual protein resembling a lipocalin in the fruitfly *Drosophila melanogaster*. The DNA sequence from which the protein was deduced came from the Berkeley *Drosophila* Genome Project. The transcription of this lipocalin has been demonstrated in our laboratory using RT-PCR. The expression pattern and function of this novel insect lipocalin is currently under study in our laboratory and will be reported elsewhere. The *Drosophila* protein also fulfills our criteria for being a lipocalin: It contains 200 amino acids, its sequence shows a 40% overall identity with its closest relative, and it clearly displays the three SCRs (fig. 2). Specifically, these characteristics make the *Drosophila* protein a kernel lipocalin.

Although not included in this study, we also detected a putative open reading frame (ORF) with a conceptual protein that strongly resembles the α_{2u} -globulins in the rabbit. Interestingly, this ORF is located in the 3' untranslated region of the cytochrome P-450 gene (accession number RABIA11).

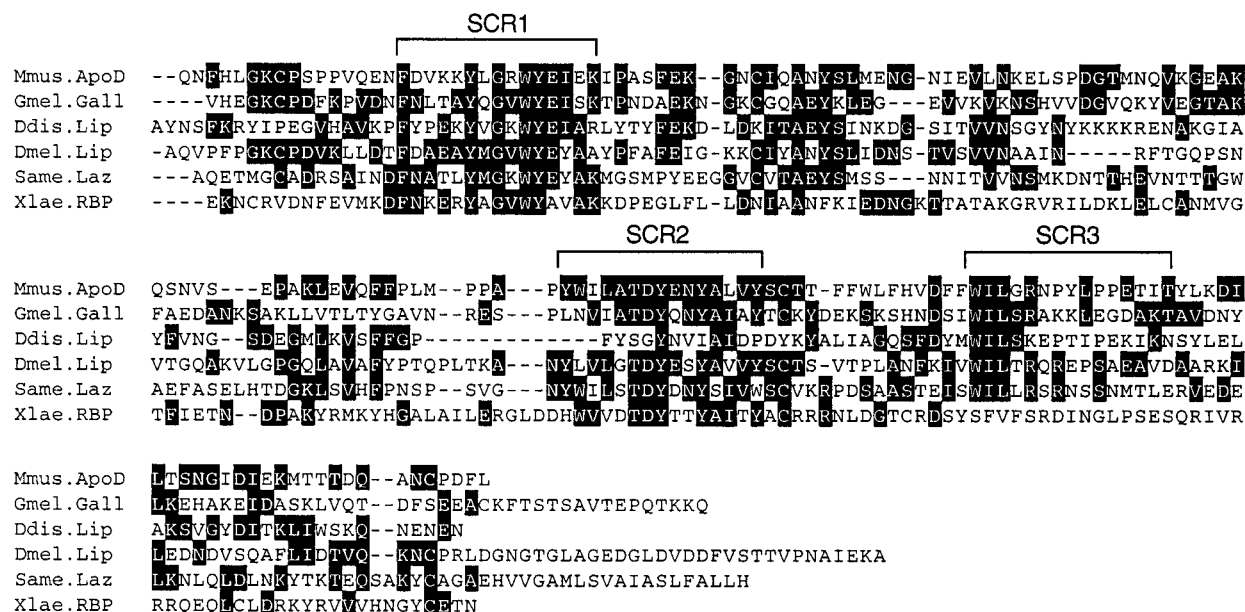


FIG. 2.—Alignment of the predicted mature proteins of the novel dictyostelid and fruitfly lipocalins with selected members of the lipocalin family (see table 1 for protein identification). The sequences were aligned with CLUSTAL W. Black boxes show residue identities present in two or more of the selected proteins. The three SCRs are highlighted in the figure.

Finally, recent work reports on plant enzymes with sequence similarity to lipocalins (Bugos, Hieber, and Yamamoto 1998). These enzymes, involved in photo-protection of the photosynthetic apparatus, are formed by a transit peptide needed for translocation to the thylakoid space of chloroplasts, a cysteine-rich domain, a lipocalin-like polypeptide, and a C-terminal charged region. It is plausible that a fusion of a plant lipocalin to other proteins has occurred during evolution to create the xanthophyll cycle enzymes. The lipocalin α_1 -microglobulin is in fact an example of a mosaic protein that is initially synthesized fused to a proteinase inhibitor (Kaumeyer, Polazzi, and Kotick 1986) and is cleaved thereafter. However, we would rather be cautious with family ascription until more information is gathered that unequivocally establishes the evolutionary origin of these plant enzymes, either from a lipocalin gene fused to other plant genes or as an unrelated protein converging toward the lipocalin fold. Thus, the xanthophyll cycle enzymes were not included in our study.

Alignments

When aligning the lipocalin amino acid sequences, we used structural criteria to guide the inclusion of alignment gaps. A penalty mask was constructed from the known folding pattern of eight lipocalins whose crystal structures had been resolved. The gap penalty mask, included because of the conserved nature of the lipocalin fold, is essential for the proper alignment of some highly divergent lipocalin sequences.

These structural constraints allow gaps in the loop regions of the lipocalin fold (see fig. 1). There are four main gaps in the lipocalin alignment that are placed at loops L1, L4, L5, and L7. Interestingly, L1, L5, and L7 are located at the open end of the barrel, and they form a loop scaffold strongly implicated in ligand binding and

in protein-protein interactions (Flower 1995). The gap at L1 is mainly due to insertions in α_1 -microglobulins (A1mg's) and α_1 -acid glycoproteins (a1GP's). Loop L1 is of particular importance for A1mg's, because they covalently bind to IgA and to a still-unknown chromophore via an intermolecular disulfide bridge located in this loop (Calero et al. 1994). A unique insertion present in the retinol-binding proteins (RBPs) accounts for the gap located in L5, a loop involved in the specific interaction of RBP with the protein transthyretin (Sivaprasadarao, Boudjelal, and Findlay 1993). Similarly, the gap in L7 is generated by an insertion in RBPs. Finally, the gap in L4, a loop located at the closed ends of the calyxes of lipocalins, is formed by insertions in β -lactoglobulins (BLs), the prostaglandin D synthases (PGDSs), and the neutrophil lipocalins (NGALs).

Other alignment regions that deserve comment are the N and C termini of lipocalins. After cleavage of the signal peptide, the N-terminal region is variable in size, ranging from 3 to 20 amino acids up to the beginning of the structurally conserved 3_{10} helix (A1 in fig. 1). The C termini are more variable than the N termini, with substantial differences even between closely related lipocalins. Moreover, there is a very specific C-terminal region in the grasshopper lipocalin Lazarillo that is responsible for the GPI membrane linkage uniquely exhibited by this lipocalin (Ganfornina, Sánchez, and Bastiani 1995). All of this information agrees with a less conserved nature of the C terminus at the structural level.

Phylogenetic Analysis

Global Tree

The resulting global tree is represented unrooted in figure 3. The seed tree showed a negative log-likelihood value of $-27,015.63 \pm 965.09$. The final tree had a

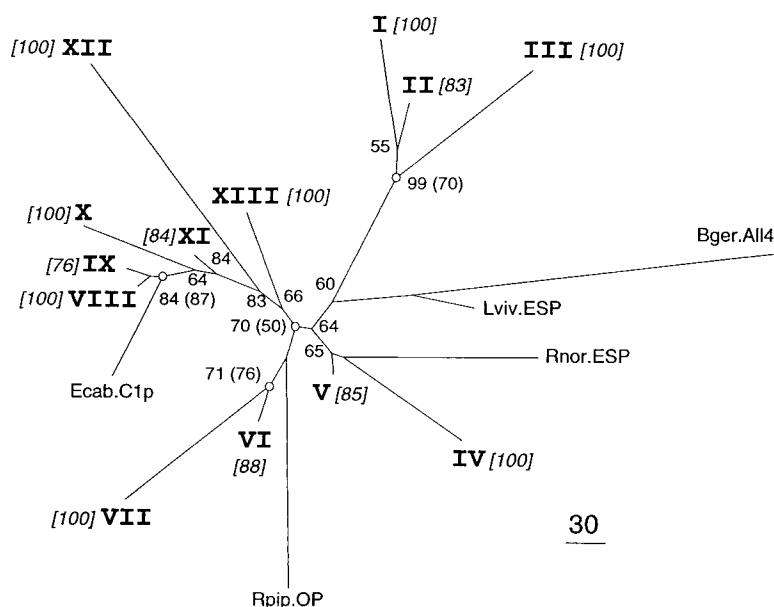


FIG. 3.—Unrooted phylogenetic tree of the lipocalin protein family. Thirteen monophyletic clades are recognized and are labeled with Roman numerals. Values in brackets are bootstrap values for the nodes defining each monophyletic clade of lipocalins. Empty circles point to nodes also supported by parsimony analysis performed with a bootstrap of 200 replicates. Local bootstrap probability (LBP) values are indicated at each node for maximum-likelihood analysis and for maximum parsimony (in parentheses). Ungrouped lipocalin proteins are individually labeled. The scale bar represents branch length (number of amino acid substitutions/100 residues).

value of $-26,918.6 \pm 961.58$. This difference reflects the improvement obtained with the local branch rearrangement method. In order to test the reliability of this tree, we repeated the reconstruction method entering the alignment sequences in different orders. No changes in topology or in LBP were obtained. Thirteen lipocalin clades (labeled with Roman numerals) were identified in the global tree. The sequences used in our analysis are listed in table 1 and assigned to each particular clade.

Clade I represents the eubacterial and dictyostelid lipocalins. The bacterial lipocalins are membrane lipoproteins that are thought to function in adaptation to starvation conditions (Bishop et al. 1995). Clade II includes the arthropodan lipocalins and ApoD's. The arthropodan lipocalins play roles in carapace coloration (Holden et al. 1987) and nervous system development (Sánchez, Ganfornina, and Bastiani 1995), whereas ApoD's are thought to play a multifunctional role in cholesterol metabolism, nerve regeneration, and cell growth and differentiation (Flower 1996). Clade III contains the RBPs, principally involved in retinol transport (Ross 1993). This clade also includes Purpurin, a protein found in the chicken retina that functions in adhesion, differentiation, and survival of the retinal epithelium (Schubert, LaCorbiere, and Esch 1986). Clade IV is composed of the BLs, proteins of unclear function abundantly present in milk whey. Clade V groups the PGDSs, the NGALs, and the quiescence-related lipocalins (QSPs). The PGDSs are crucial for prostaglandin D synthesis in the brain (Nagata et al. 1991), although they have been recently reported to bind retinoic acid (Tanaka et al. 1997), which might confer on them novel functions. The NGALs inhibit neutrophil gelatinase and therefore may have anti-inflammatory properties (Kjeld-

sen et al. 1993). Clade VI clusters the A1mg's, which show immunosuppressive and mitogenic activities (Åkerström and Logdberg 1990). Clade VII contains the complement 8γ lipocalins (C8GCs), which could be involved in the regulation of the complement cascade (Haefliger et al. 1988). The mouse urinary proteins (MUPs; clade VIII) and the rat α_{2u} -globulins (α_2g 's; clade IX) are pheromone transporters in rodent urine (Cavaggioni, Findlay, and Tirindelli 1990). Clades X and XIII group proteins (e.g., OBPs, Pba's, VNSPs, VEG) involved in chemoreception of odorants and gustatory molecules (reviewed by Flower 1996). Clade XI groups three recently discovered lipocalins of unknown function. Finally, clade XII brings together the $\alpha 1GP$'s, lipocalins with multiple immunoregulatory properties (Kremer, Wilting, and Janssen 1988). Five lipocalin sequences remained ungrouped, although they clearly fell into separate superclades. Some of these sequences might belong to an already-described clade, but have a large amino acid sequence divergence. Alternatively, they could belong to new clades for which other related sequences have not been reported.

Overall, the global tree segregates the lipocalins into the same family groups previously described in the literature (Flower 1996). However, there are clades, such as clades II and V, that group together proteins lacking a common expression pattern or function. Moreover, some nodes in the tree receive strong support and suggest the existence of monophyletic lipocalin superclades. These superclades are: (1) the group of clades I, II, and III; (2) the group including clades I–V and three ungrouped sequences; (3) the node clustering clades VI and VII; and (4) the node that groups clade VIII, clade IX, and the ungrouped horse sequence Ecab.C1p. These

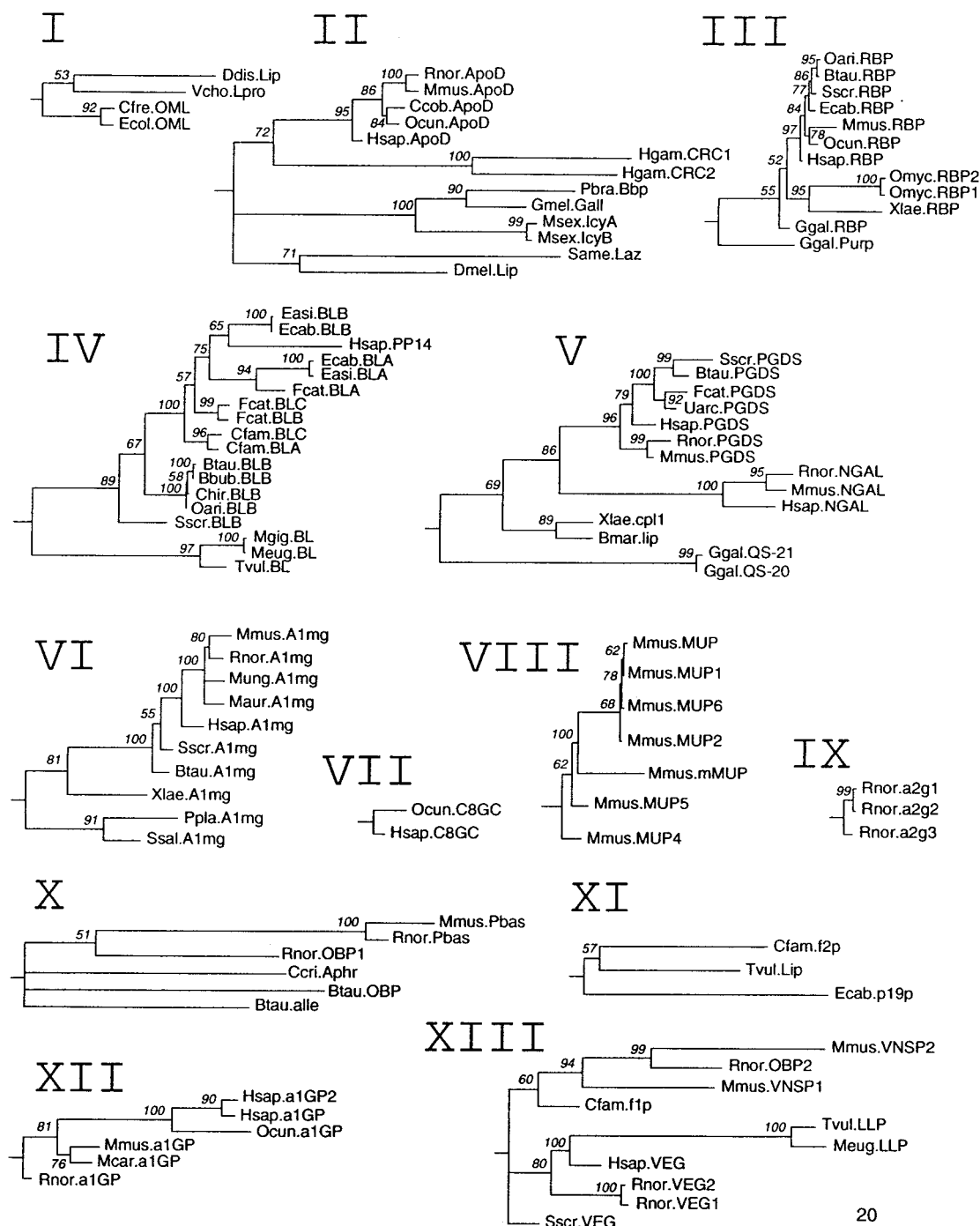


FIG. 4.—Phylogenetic trees derived from maximum-likelihood analyses of the individual clades of lipocalins. Local bootstrap probability (LBP) values are indicated in each node. Polytomies reflect nodes with LBP values below 50%. The scale bar represents branch length (number of amino acid substitutions/100 residues).

superclades are also identified and supported by 200 bootstrap replicates of a maximum-parsimony analysis. A monophyletic group composed of clades VIII–XII is supported with an 83% LBP but is not supported by maximum-parsimony analysis.

Analysis of Subtrees

In analyzing the phylogeny of lipocalins, we want to resolve both the relationships among different lipocalin groups and the relationships among the sequences within each group. For this reason, we ran separate phylogenetic analyses on each clade identified in the global tree. The resulting subtrees are shown in figure 4.

There are some clades (e.g., clades I, VII, and IX) that are composed of small numbers of proteins, some even from a single species. In those cases, it is difficult to obtain meaningful information about the evolutionary history of the proteins.

Downloaded from <http://mbe.oxfordjournals.org/> at Biblioteca Area Biomedica on May 9, 2015

The phylogeny of A1mg (clade VI) clearly reflects orthologous relationships. Both its topology and the branch distances are in good agreement with the species tree of vertebrates. In contrast, other subtrees, such as that of the BLs (clade IV), illustrate a complex history of gene duplications. In some cases, those duplications preceded speciation, as in the case of equine BLs. In other cases, recent duplications appear in a given species (e.g., the dog BLA and BLC, and the cat BLB and BLC). The existence of recurrent duplications in this subtree is supported by the presence of some BL pseudogenes in the goat and cow (Passey and Mackinlay 1995), which have not been included in our analysis. The BL tree reveals an interesting location for the human PP14, a protein secreted by the endometrial and decidual epithelia and abundantly present in uterine and amniotic fluids. PP14 is grouped with the two equine BLBs and does not appear as an ancestral form in the BL tree, as is the case in maximum-parsimony trees, both in the present study (not shown) and in the work by Piotte et al. (1998). The derived position of Hsap.PP14 in our BL subtree suggests a strong divergence from the BL sequence signature. This agrees with the absence of BLs in primate milk and suggests that BLs were recruited for other reproductive roles early in primate evolution. Other lipocalin clades such as RBPs (clade III) or a1GPs (clade XII) mostly show orthologous sequences, although some recent duplications appear in specific species. An interesting duplication is recorded in the RBP subtree involving the retinal protein Purpurin, so far only identified in the chicken. The basal position of Purpurin in the tree suggests two possible evolutionary possibilities. Either an early lineage-specific duplication of RBPs along with divergence of Purpurin occurred in birds, while the rest of vertebrates lack the retinal protein, or, alternatively, Purpurin is indeed present in other lineages but not identified yet.

Two clades (clades II and V) group proteins that are apparently unrelated. Clade V combines the PGDSs, the NGALs, and the QSPs. The amphibian proteins Xlae.cpl1 and Bmar.lip are most probably orthologous PGDSs, given their expression pattern and retinoic acid binding (Achen et al. 1992; Lepperdinger et al. 1996). The ancestral position of the chicken QSPs in the tree, along with the unreported presence of PGDS and NGAL in birds, suggests that QSPs are ancestral forms of PGDS and that they may bear other features common to PGDSs besides stabilizing growth in cell populations (reviewed by Flower 1996). The well-supported monophyletic relationship of the mammalian NGALs and PGDSs and the presence of both protein types in the same species suggest a case of gene cooption after duplication and divergence of an ancestral PGDS-like gene, specifically in the mammalian lineage. In clade II, the arthropodan lipocalins cluster together with the vertebrate ApoD's. The crustacean lipocalins are particularly closely related to ApoD, while the insect lipocalins stand apart as two polytomic clusters separately relating the fat body and epidermis lipocalins on one side and the nervous system lipocalins on the other. The duplications found in the crustacean *Homarus gammarus* and

in the lepidopteran *Manduca sexta* seem to be recent events in the evolution of these lineages. The unresolved node associating insect lipocalins and the lengthy branches of the arthropodan tree indicate a long and divergent history for these lipocalins. Further analysis of the expression pattern and functional role of the proteins associated in this clade is needed in order to formulate hypotheses about the ancestral role of lipocalins in the metazoan lineage.

Finally, the monophyletic superclade grouping clades VIII–XIII contains proteins found so far only in mammals. These lipocalins have been considered outliers in the family (Flower, North, and Atkwood 1993) because of their low amino acid sequence similarity and lack of some of the SCRs. Accordingly, these subtrees show long branch lengths and poorly supported nodes. These groups contain lipocalins that are particularly prone to gene duplications (e.g., MUPs, a2g's). However, attempts to derive further evolutionary implications within each clade are limited by the high sequence divergence and should take into account other characters for phylogenetic reconstruction.

Rooting the Lipocalin Tree

In the previous section, we analyzed the relationships among different lipocalins and lipocalin groups using an unrooted tree. Rooting our global lipocalin tree would help us to assign polarity to character changes and to suggest plausible scenarios for the evolution of specific lipocalin properties. The existence of a single cluster grouping the arthropodan lipocalins should allow an unambiguous rooting for the metazoan lipocalins. However, the discovery of bacterial lipocalins (Bishop et al. 1995; Bishop and Weiner 1996) opens up the possibility of rooting the tree using the clade of prokaryotic lipocalins.

The bacterial lipocalins have been found in a restricted number of species, raising the possibility that these lipocalins originated through horizontal transfer. In order to determine whether the bacterial lipocalins are the result of horizontal transfer, we estimated the G+C contents in the first and third codon positions of gene samples of the bacterial species under study. (The samples were retrieved from the GenBank database.) We then calculated the mean of these indices and the first and third quartiles. A biased G+C in the first and third codon positions would be suggestive of horizontal transfer (Lawrence and Ochman 1997). The results of these calculations (see table 2) show that none of the computed G+C contents of the bacterial lipocalin genes are outside of the expected limits (between the first and third quartiles). Given the predicted time sensitivity of this test (Lawrence and Ochman 1997), our data provide no support for a hypothesis whereby bacterial lipocalins were recently acquired through horizontal transfer.

Further evidence against the horizontal transfer hypothesis could come from finding more lipocalins in different bacteria, thus making the gene transfer hypothesis even more unlikely. We did a BLAST search at the NCBI for the bacterial genomes (both complete and incomplete) using the entire protein sequences of the

Table 2
Analysis of G+C Contents of Bacterial Lipocalins as Compared with Other Genes in Their Respective Organisms

SPECIES	SAMPLE	G+C %, FIRST CODON POSITION				G+C %, THIRD CODON POSITION			
		Mean	Q1	Q3	Lipocalin	Mean	Q1	Q3	Lipocalin
<i>Citrobacter freundii</i>	50	55.73	52.38	60.89	56.18	53.83	43.31	61.64	59.55
<i>Escherichia coli</i>	3,911	58.15	54.75	62.59	61.24	54.50	50.35	60.05	52.81
<i>Vibrio cholerae</i>	390	51.45	46.79	56.30	50.00	41.34	31.66	50.00	38.95

Escherichia and *Vibrio* lipocalins as independent queries. Four new putative lipocalins were discovered in three species of purple and green sulfur eubacteria (*Pseudomonas aeruginosa*, *Campylobacter jejuni*, and *Chlorobium tepidum*) that showed significant similarity to the existing lipocalins (BLAST E values ranging from 1×10^{-7} to 4×10^{-45}). These sequences were not annotated at the time of the search (September 1998), and therefore the results must be interpreted with caution. However, the conceptual sequences of these new bacterial proteins show clear lipocalin character (fig. 5).

The bacterial lipocalins included in our analysis are lipoproteins anchored to the outer membrane of the Gram-negative bacteria *Vibrio*, *Escherichia*, and *Citrobacter* (Barker and Manning 1997; Bishop et al. 1995). Accordingly, lipocalins are expected to be absent in Gram-positive eubacteria and archaeobacteria, both lacking an outer membrane. The absence of lipocalins from the completely sequenced genomes of the Gram-negative eubacteria *Haemophilus* and *Helicobacter* is puzzling, but gene loss is a plausible reason.

The results presented here suggest that the prokaryotic lipocalins have been vertically transmitted and that the common ancestor of these lipocalins with the metazoan lipocalins is very ancient indeed. Therefore, we rooted our global lipocalin tree using the clade containing all of the prokaryotic lipocalins as an outgroup.

A Rooted Tree for Lipocalins

The result of rooting the lipocalin tree with the bacterial lipocalins is shown in figure 6. It is puzzling that the *D. discoideum* lipocalin appears to be located within the bacterial clade (clade I). The dictyostelid lipocalin might be the result of a process of convergent evolution under bacterial-like selective forces. Alternatively, this could be a technical artifact of phylogenetic reconstruction which could be resolved by further sampling of protist lipocalins.

The arthropodan lipocalins (clade II) stem from the first branch of the metazoan lipocalin tree (fig. 6). The presence in this clade of mammalian ApoD's suggests that these lipocalins have an ancestral nature within the chordate lineage and that they should be present throughout the chordate phylum. Support for this hypothesis comes from the cloning and current characterization of an ApoD gene expressed in chickens (unpublished data). The next branch of the rooted tree clusters the RBPs (clade III), which have been widely found in vertebrates and seem to have shared a common ancestor with ApoD's after the split of arthropods and chordates.

Clades II and III, along with the outgroup lipocalins (clade I), are separated from the rest of the family members by a long branch that indicates a large sequence divergence. The rest of the tree seems to reflect

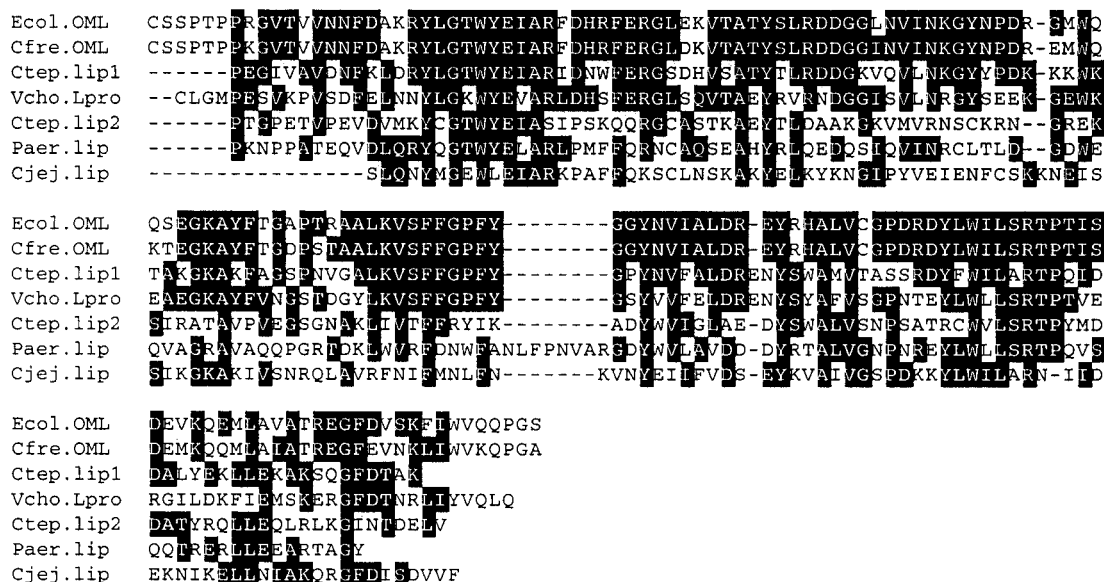


FIG. 5.—Protein sequence alignment of the bacterial lipocalins including novel conceptual protein sequences from the eubacteria *Pseudomonas aeruginosa* (Paer.lip) and *Campylobacter jejuni* (Cjej.lip) and two sequences from *Chlorobium tepidum* (Ctep.lip1 and Ctep.lip2). The sequences were aligned with CLUSTAL W. Gray boxes show residue identities present in two or more of the selected proteins.

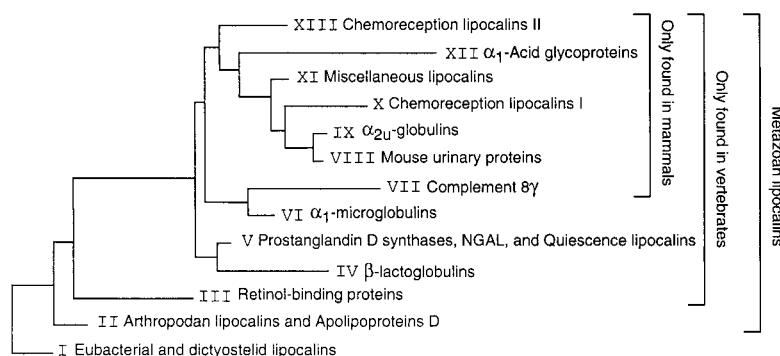


FIG. 6.—Phylogenetic tree of the supported lipocalin clades rooted with the eubacterial and dictyostelid lipocalins as the ancestral group. The scale bar represents branch length (number of amino acid substitutions/100 residues).

the entire genome duplications and/or individual gene duplication events that are thought to have occurred during vertebrate evolution (Holland et al. 1994). Clades IV and V represent the result of a duplication from an ancestral gene that was probably closer in sequence to the PGDSs. One of the duplicated genes gave rise to the exclusively mammalian BL clade, even more derived and rich in recent gene duplications (see above and fig. 4). A similar evolutionary pathway can be suggested for the related A1mg's and the C8GCs (clades VI and VII), given the existence of A1mg's in primitive chordates.

The chemoreceptor lipocalins, urinary proteins, and a1GP's (clades VIII–XIII) appear to be a monophyletic group. The origin of these outlier lipocalins poses a number of interesting evolutionary questions. This superclade is highly derived (based on the strong amino acid sequence divergence) and rich in individual gene duplications (e.g., MUPs). It is composed of lipocalins so far found only in marsupials and placental mammals, which suggests a shared common ancestor that was present before the split of these two mammalian groups. The marsupial late-lactation proteins (LLPs) appear to be related to a group of chemoreception lipocalins (in clade XIII), and the possum whey protein Trichosurin (Tvul.Lip) groups with an endometrial equine protein and a salivary dog lipocalin in clade XI. Similar phylogenetic relationships have been reported in a parsimony analysis (Piotte et al. 1998). The sequence relationships revealed by our trees suggest that the outlier lipocalins evolved from the ancient A1mg's (clade VI). This is consistent with the close relationship between a putative odorant-binding protein found in amphibians (Rpip.OP) and the A1mg's (see fig. 3). It is worth noting that there are also common characteristics between the superclade of outlier lipocalins and the BLs (clade IV). They all have similar sequence divergence values (see below), and both the BLs and the marsupial outlier lipocalins are expressed in the mammary glands. These characteristics, however, could be the result of convergence and of independent cooption events for a role in the mammary glands if it holds true that the A1mg's are the ancestors of the outlier lipocalins. An alternative evolutionary hypothesis, although not supported by our protein-sequence-derived trees, is that the outlier lipocalins originated by divergence from an ancestrally du-

plicated BL. Some of these proteins could have maintained the ancestral role in the mammary glands of marsupials, while the eutherian counterparts were coopted for other functions as ligand-binding proteins secreted to various body fluids. More sampling of lipocalins in these groups is necessary to resolve this issue.

Phylogenetic Partitioning of Lipocalin Properties

An important advantage of rooting the lipocalin family tree is that we can assign polarities to the evolution of particular lipocalin features. We studied the tissue pattern of protein/mRNA expression and the proposed function for each lipocalin clade. No straight correspondence is obvious between the tree topology and the physiological roles of different lipocalins. Although there are some clades with distinctive functions (e.g., retinol metabolism in RBPs or cryptic coloration in lepidopteran and crustacean lipocalins), some other functions are carried out by members of different clades. Binding of odorants, for example, has been reported for chemoreception lipocalins (clades VIII–XIII) and for human ApoD from clade II (Zeng et al. 1996), and a broad spectrum of lipocalins are thought to play roles in immune modulation and cell regulation (Flower 1994, 1996). Similarly, the extent of gene duplications within each clade does not show a trend according to the tree relationships. We also studied some biochemical properties of lipocalins, such as oligomerization and glycosylation. None of these characteristics showed a clear phylogenetic pattern.

Anchoring to the membrane surface has been proposed to be an ancestral character for lipocalins (Bishop and Weiner 1996) based on the membrane localization of bacterial lipocalins, the grasshopper protein Lazarillo, and ApoD. These authors suggest a plausible evolutionary scenario that would account for the different modes of membrane association observed in lipocalins. According to this hypothesis, the ancestral N-terminal lipid binding of bacterial lipocalins was followed by the appearance of a hydrophobic loop in eukaryotic lipocalins (L7, between β -strands 7 and 8) that associates them with membrane proteins. This was followed by the acquisition of the C-terminal GPI anchor, a more specialized type of membrane association. The last two types of membrane binding were subsequently lost in the rest

of the family. Our phylogenetic study nevertheless suggests a radically different hypothesis. Primitive eukaryotic lipocalins, such as Ddis.Lip, apparently lack hydrophobic loops that could be used for membrane anchoring. The same applies to the lipocalins found in crustaceans (Hgam.CRC1 and CRC2) and in several orders of insects, both primitive (*Schistocerca americana*) and highly derived (*D. melanogaster*). Moreover, the generality of the function for the hydrophobic surface loop of ApoD still remains to be proved, since the only demonstrated interaction of loop L7 (Yang et al. 1994) requires a cysteine residue found so far only in human ApoD. This points to a protein-protein interaction of recent invention in the evolution of the ApoD clade and therefore categorizes the ApoD membrane localization as an autapomorphic character for phylogeny reconstruction. Thus, we find no reasons to propose the hydrophobicity of the ApoD loop L7 to be a character present in the ancestor of both chordate and arthropod lineages. In addition, we so far have no evidence for the ancestrality of the GPI membrane anchor within the arthropod lipocalins, since it has only been found in the grasshopper Lazarillo protein. Currently, the most plausible hypothesis is that the GPI anchor is also an autapomorphic character.

We also explored the sequence divergence in the lipocalin family by calculating the maximum distance (amino acid substitutions/100 residues) observed in each individual clade. We only compared the clades within the subtrees that relate mammalian representatives because of possible biases in the sampling of sequences in different organisms. The values obtained (amino acid substitutions/100 residues) were as follows: clade II, 30; clade III, 17; clade IV, 103; clade V, 43; clade VI, 37; clade X, 189; clade XI, 113; clade XII, 104; clade XIII, 145. These results suggest an increasing sequence divergence in the more derived lipocalins. We did not include clades VII, VIII, and IX because of their scarce organismal representation, but the trend is still clear when the 50 amino acid substitutions calculated for the MUPs (clade VIII) are considered.

Likewise, there seems to be a phylogenetic trend in the degree of contact between the ligands and the internal cavities of lipocalins, which is evidenced when mapping onto the lipocalin tree several parameters that govern the ligand-binding interactions. The binding areas of the internal cavities of Msex.Icy and Pbra.Bbp (clade II) are approximately 1,255 Å², while those of Hsap.Rbp (clade III) and Mmus.MUP6 (clade VIII) decrease to 860 and 500 Å², respectively (Flower 1995). The surface area of the cavity of Btau.Obp (clade X) is estimated to be 390 Å² (Tegoni et al. 1996). These data are consistent with the fact that relatively large ligands bind to the more ancestral clades, while derived lipocalins bind to smaller ligands. The lipocalins of the ancestral clade II bind large ligands, such as the bilins that bind to the lepidopteran lipocalins and ApoD (biliverdin IX γ shows 920.7 Å² of solvent-accessible surface [SAS]) and the hemin that binds to a bacterial lipocalin (Barker and Manning 1997). The lipocalins of clade III bind ligands of intermediate size, since retinol has an

SAS of 588.9 Å². Finally, small ligands are bound by the odorant-binding proteins and MUPs of clades X and VIII (with endogenous ligands showing 407.1 Å² and 326.5 Å² of SAS, respectively). In addition, we studied the ligand-protein contacts using the complementary factors (Sobolev et al. 1996) in five representative lipocalins with known tertiary structure: Pbra.Bbp (code 1BBP in the Brookhaven data bank), Hsap.Rbp (1RBP), Btau.BLG (1BSO), Btau.Obp (1PBO), and Mmus.MUP6 (1MUP). The complementarity factors calculated for each ligand-protein pair are 0.4 (1BBP), 0.54 (1RBP), 0.34 (1BSO), 0.82 (1PBO), and 0.86 (1MUP). This parameter reflects that the contacts between, for instance, Pbra.Bbp and its bilin ligand are looser than those established between Mmus.MUP6 and the ligand 2-(s-butyl) thiazoline. When all of these results are mapped on the global rooted tree of lipocalins, they suggest a decrease in binding surface area during the evolution of lipocalins. Also, in close relation to this pattern, the number of intramolecular disulfide bridges ranges from the two to three bridges present in the ancestral lipocalins (clades II–IV) to the single bridges, and even absence of disulfide bonds (Tegoni et al. 1996), of more derived lipocalins, which will tend to confer more flexible protein structures.

In summary, a view of the lipocalin family is emerging from our phylogenetic analyses: The more derived lipocalins seem to have evolved a more flexible protein structure and a ligand-binding pocket that binds smaller hydrophobic ligands with more efficiency than the ancestral lipocalins. In addition, this trend is accompanied by a greater rate of sequence divergence within the more derived clades.

Phylogenetic Relationships Within the Calycin Superfamily

Flower, North, and Atkwood (1993) and Flower (1993) proposed a protein superfamily called calycins that includes the lipocalins, the avidins, and a complex group of mostly intracellular lipid transporters globally called the fatty-acid-binding proteins (FABPs). The hypothetical phylogenetic relationship of these three protein families was based on their remarkable structural resemblance: each had an antiparallel β -barrel with an internal ligand-binding site. Nonetheless, their low and local sequence similarity makes an overall alignment of these protein families very unreliable. However, lipocalins and FABPs have been aligned including their structural similarity as an important criterion (e.g., Flower, North, and Atkwood 1993). We used this alignment to add several members of the FABP family to our data set in order to construct trees that would test whether the FABPs constitute a sister group of the lipocalins. The tree obtained from a maximum-likelihood analysis is shown in figure 7. The FABP representatives form a monophyletic group with a high LBP value. However, the FABP clade does not group with the prokaryotic lipocalins, which would be expected if a common ancestor of the two protein families ever existed. Instead, the FABPs group with the more derived lipocalins (clades IV–XIII). These results confer no good evidence

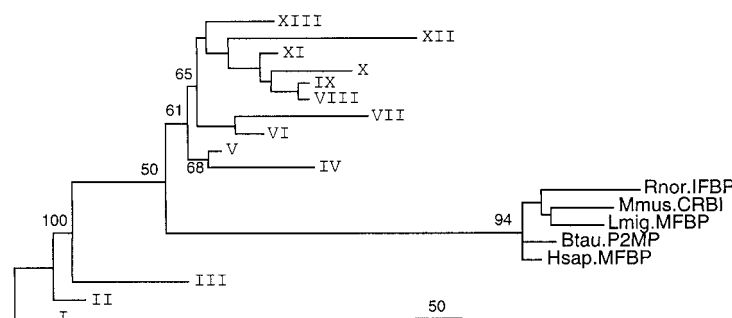


FIG. 7.—Maximum-likelihood tree showing the relationship between the lipocalins (clades labeled with Roman numerals) and representatives of the fatty-acid-binding protein (FABP) family. Abbreviations and accession numbers of the FABPs were as follows: Rnor.IFBP, rat intestinal fatty- acid-binding protein (P02693); Mmus.CRBP, mouse cellular retinoic-acid-binding protein (Q00915); Lmig.MFBP, locust muscular fatty-acid-binding protein (P41509); Btau.P2MP, cow myelin protein P2 (P02690); Hsap.MFBP, human muscular fatty- acid-binding protein (X56549). LBP values are indicated in each node. The scale bar represents branch length (number of amino acid substitutions/100 residues).

for a sister group relationship and suggest two mutually exclusive hypotheses: (1) that the FABPs evolved from already-existing lipocalins or (2) that the structural resemblance of FABPs and lipocalins reflects a process of evolutionary convergence to optimize a similar ligand-binding function.

Conclusions

The lipocalins are a complex family of proteins. The results presented in this work, along with the accumulating mass of data and ideas from other research fields, are bringing together an enticing view of lipocalin evolutionary history. Our phylogenetic analyses suggest that the lipocalins appeared early in the history of life and are thus expected to be present in all metazoan phyla, although gene loss or extreme amino acid divergence certainly occurred in particular lineages, such as the yeast *Saccharomyces cerevisiae* and the nematode *Caenorhabditis elegans*. A small number of genes are present in nonvertebrate metazoans, while an elaborate radiation of lipocalins occurred in vertebrates, as has been demonstrated for other genes (Holland and García-Fernández 1996). As a consequence of this evolutionary history, the function of the extant lipocalins reflects a panoply of gene cooptions arising with or without previous gene duplications (Ganformina and Sánchez 1999). The lipocalin fold seems to have the potential for many different uses, depending on where and when the protein is expressed in the organism and which ligand or protein interactions can take place. Hypotheses of lipocalin function can now be placed in a phylogenetic context, which in turn should help guide future sampling of new lipocalins, as well as deeper analyses of the expression and function of known lipocalins. We anticipate that the phylogenetic framework presented here will provide important insights into the history of changing functions of these apparently simple but versatile proteins.

Acknowledgments

We are grateful to W. J. Dickinson, S. Emerson, and J. Seger for their helpful discussion and comments on the manuscript.

LITERATURE CITED

- ACHEN, M. G., P. J. HARMS, T. THOMAS, S. J. RICHARDSON, R. E. WETTENHALL, and G. SCHREIBER. 1992. Protein synthesis at the blood-brain barrier. The major protein secreted by amphibian choroid plexus is a lipocalin. *J. Biol. Chem.* **267**: 23170–23174.
- ADACHI, J., and M. HASEGAWA. 1996. MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood. *Comput. Sci. Monogr.* **28**:1–150.
- AKERSTROM, B., and L. LOGDBERG. 1990. An intriguing member of the lipocalin protein family: alpha-1-microglobulin. *Trends Biochem. Sci.* **15**:240–243.
- 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.
- BARKER, A., and P. A. MANNING. 1997. VlpA of *Vibrio cholerae* 01: the first bacterial member of the α_2 -microglobulin lipocalin superfamily. *Microbiology* **143**:1805–1813.
- BISHOP, R. E., S. S. PENFOLD, L. S. FROST, J. V. HÖLTJE, and J. H. WEINER. 1995. Stationary phase expression of a novel *Escherichia coli* outer membrane lipoprotein and its relationship with mammalian apolipoprotein D. *J. Biol. Chem.* **270**:23097–23103.
- BISHOP, R. E., and J. H. WEINER. 1996. 'Outlier' lipocalins more than peripheral. *Trends Biochem. Sci.* **21**:127.
- BUGOS, R. C., A. D. HIEBER, and H. Y. YAMAMOTO. 1998. Xanthophyll cycle enzymes are members of the lipocalin family, the first identified from plants. *J. Biol. Chem.* **273**: 15321–15324.
- CALERO, M., J. ESCRIBANO, A. GRUBB, and E. MÉNDEZ. 1994. Location of a novel type of interpeptide chain linkage in the human protein HC-IgA complex (HC-IgA) and identification of heterogeneous chromophore associated with the complex. *J. Biol. Chem.* **269**:384–389.
- CAVAGGIONI, A., J. B. C. FINDLAY, and R. TIRINDELLI. 1990. Ligand binding characteristics of homologous rat and mouse urinary proteins and pyrazine-binding protein of calf. *Comp. Biochem. Physiol.* **96B**:513–520.
- COWAN, S. W., M. E. NEWCOMER, and T. A. JONES. 1990. Crystallographic refinement of human serum retinol binding protein at 2 Å resolution. *Proteins Struct. Funct. Genet.* **8**: 44–61.
- FELSENSTEIN, J. 1993. PHYLIP (phylogeny inference package). Distributed by the author, Department of Genetics, University of Washington, Seattle.
- FLOWER, D. R. 1993. Structural relationship of streptavidin to the calycin protein superfamily. *FEBS Lett.* **333**:99–102.

- . 1994. The lipocalin protein family: a role in cell regulation. *FEBS Lett.* **354**:7–11.
- . 1995. Multiple molecular recognition properties of the lipocalin protein family. *J. Mol. Recogn.* **8**:185–195.
- . 1996. The lipocalin protein family: structure and function. *Biochem. J.* **318**:1–14.
- FLOWER, D. R., A. C. T. NORTH, and T. K. ATKWOOD. 1993. Structure and sequence relationships in the lipocalins and related proteins. *Protein Sci.* **2**:753–761.
- GANFORNINA, M. D., and D. SÁNCHEZ. 1999. Generation of evolutionary novelties by functional shift. *Bioessays* **21**: 432–439.
- 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.
- HAEFLIGER, J. A., M. C. PEITSCH, D. E. JENNE, and J. TSCHOPP. 1988. Structural and functional characterisation of C₈, a member of the lipocalin protein family. *Mol. Immunol.* **28**: 123–131.
- 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.
- HOLDEN, H. M., W. R. RYPNIEWSKI, J. H. LAW, and I. RAYMENT. 1987. The molecular structure of insecticyanin from the tobacco hornworm *Manduca sexta* L. at 2.6 Å resolution. *EMBO J.* **6**:1565–1570.
- HOLLAND, P. W. H., and J. GARCÍA-FERNÁNDEZ. 1996. *Hox* genes and chordate evolution. *Dev. Biol.* **173**:382–395.
- HOLLAND, P. W. H., J. GARCÍA-FERNÁNDEZ, N. A. WILLIAMS, and A. SIDOW. 1994. Gene duplications and the origins of vertebrate development. *Dev. Suppl.* pp.125–133.
- 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.
- JONES, D. T., W. R. TAYLOR, and J. M. THORNTON. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8**:275–282.
- KAUMEYER, J. F., J. O. POLAZZI, and M. P. KOTICK. 1986. The mRNA for a proteinase inhibitor related to the HI-30 domain of inter-alpha-trypsin inhibitor also encodes alpha-1-microglobulin (protein HC). *Nucleic Acids Res.* **14**:7839–7850.
- KEEN, J. N., I. CACERES, E. E. ELIOPOULOS, P. F. ZAGALSKY, and J. B. C. FINDLAY. 1991. Complete sequence and model for the A₂ subunit of the carotenoid pigment complex, crustacyanin. *Eur. J. Biochem.* **197**:407–417.
- KJELDSSEN, L., A. H. JOHNSEN, H. SENGELOV, and N. BORREGAARD. 1993. Isolation and primary structure of NGAL, a novel protein associated with human neutrophil gelatinase. *J. Biol. Chem.* **268**:10425–10432.
- KREMER, J. M. H., J. WILTING, and L. H. M. JANSSEN. 1988. Drug binding to human alpha-1-acid glycoprotein in health and disease. *Pharmacol. Rev.* **40**:1–47.
- LAWRENCE, J. G., and H. OCHMAN. 1997. Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.* **44**:383–397.
- LEPPERDINGER, G., B. STROBL, A. JILEK, A. WEBER, J. THALHAMER, H. FLÖCKNER, and C. MOLLAY. 1996. The lipocalin Xlcp11 expressed in the neural plate of *Xenopus laevis* embryos is a secreted retinaldehyde binding protein. *Protein Sci.* **5**:1250–1260.
- NAGATA, A., Y. SUZUKI, M. IGARASHI, N. EGUCHI, H. TOH, Y. URADE, and O. HAYAISHI. 1991. Human brain prostaglandin D synthase has been evolutionarily differentiated from lipophilic-ligand carrier proteins. *Proc. Natl. Acad. Sci. USA* **88**:4020–4024.
- NEWCOMER, M. E., and D. E. ONG. 1990. Purification and crystallization of a retinoic acid-binding protein from rat epididymis. *J. Biol. Chem.* **265**:12876–12879.
- PASSEY, R. J., and A. G. MACKINLAY. 1995. Characterisation of a second, apparently inactive, copy of the bovine beta-lactoglobulin gene. *Eur. J. Biochem.* **233**:736–743.
- PIOTTE, C. P., A. K. HUNTER, C. J. MARSHALL, and M. R. GRIGOR. 1998. Phylogenetic analysis of three lipocalin-like proteins present in the milk of *Trichosurus vulpecula* (Phalangeridae, Marsupialia). *J. Mol. Evol.* **46**:361–369.
- ROSS, A. C. 1993. Cellular metabolism and activation of retinoids: roles of cellular retinoid-binding proteins. *FASEB J.* **7**:317–327.
- SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**:406–425.
- 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.
- SCHUBERT, D., M. LACORBIERE, and F. ESCH. 1986. A chick neural retina adhesion and survival molecule is a retinol-binding protein. *J. Cell Biol.* **102**:2295–2301.
- SIVAPRASADARAO, A., M. BOUDJELAL, and J. B. C. FINDLAY. 1993. Lipocalin structure and function. *Biochem. Soc. Trans.* **21**:619–622.
- SOBOLEV, V., R. C. WADE, G. VRIEND, and M. EDELMAN. 1996. Molecular docking using surface complementarity. *Proteins* **25**:120–129.
- TANAKA, T., Y. URADE, H. KIMURA, N. EGUCHI, A. NISHIKAWA, and O. HAYAISHI. 1997. Lipocalin-type prostaglandin D synthase (beta-trace) is a newly recognized type of retinoid transporter. *J. Biol. Chem.* **272**:15789–15795.
- TEGONI, M., R. RAMONI, E. BIGNETTI, S. SPINELLI, and C. CAMBILLAU. 1996. Domain swapping creates a third putative combining site in bovine odorant binding protein dimer. *Nat. Struct. Biol.* **3**:863–867.
- THOMPSON, J. D., D. G. HIGGINS, and T. J. GIBSON. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**:4673–4680.
- VON HEIJNE, G. 1990. The signal peptide. *J. Membr. Biol.* **115**: 195–201.
- YANG, C. Y., Z. W. GU, F. BLANCO-VACA, S. J. GASKELL, M. YANG, J. B. MASSEY, A. M. GOTTO, and H. J. POWNALL. 1994. Structure of human apolipoprotein D: locations of the intermolecular and intramolecular disulfide links. *Biochemistry* **33**:12451–12455.
- ZENG, C., A. I. SPIELMAN, B. R. VOWELS, J. J. LEYDEN, K. BIEMANN, and G. PRETI. 1996. A human axillary odorant is carried by apolipoprotein D. *Proc. Natl. Acad. Sci. USA* **93**:6626–6630.

MANOLO GOUY, reviewing editor

Accepted August 30, 1999