Lipocalin Genes and Their Evolutionary History

Diego Sanchez,* María D. Ganfornina, Gabriel Gutierrez, Anne-Christine Gauthier-Jauneau, Jean-Loup Risler and Jean-Philippe Salier

Introduction

As extensively detailed elsewhere in this book, lipocalins exhibit three characteristic features, which include: (i) an unusually low amino acid sequence similarity (typically 15-25% between paralogs) (ii) a highly conserved protein tertiary structure, and (iii) a similar arrangement of exons and introns in the coding sequence of their genes. These shared protein and gene features are overwhelming arguments for the existence of a single lipocalin ancestral gene that once extended into a family.

The ancestral gene appears to have arisen in a group of bacteria, and possibly was inherited by eukaryotes as a result of genome fusion (see Chapter 4). Given this hypothetical beginning, lipocalins are expected to be found in all descendants of the eukaryotic common ancestor. Currently, and aside of prokaryotes, bona fide lipocalin have been recovered from a protoctist, a fungus, several plants, a nematode, several arthropods, a tunicate, a cephalochordate, and many examples of chordates.

This review will first focus on the structure of lipocalin genes in eukaryotes, and then on our current view of the evolutionary history of this family.

An Overview of Lipocalin Genes

Exon-Intron Organization of Eukaryote Lipocalins

Exon-intron arrangements in lipocalins were first deciphered mostly in human and rodents, but the virtually complete information on genome sequence in an ever growing number of organisms, as released over the last few years, has considerably extended our knowledge of gene organization in eukaryotes at large. Whether by direct sequencing of known genes or bioinformatics-aided identification of novel open reading frames (ORF), the lipocalin genes whose structure is now established has benefited from this flow of information. Figure 1, provides an overview of the exon-intron arrangement of lipocalin genes, from unicellular eukaryotes to human. Given the large number of lipocalin genes that arose from duplication and retained a similar organization (see below), not every lipocalin currently known is depicted in Figure 2. The latter intends to provide an overview of gene structure with major trends rather than a comprehensive list of genes. When the position of introns is marked in a protein sequence alignment of different lipocalins, the pattern that emerges immediately highlights a model gene arrangement with a maximum of five exons (e1-e5) and four introns (A-D) in

*Corresponding Author: Diego Sanchez—Departamento de Bioquímica y Biología Molecular y Fisiología-IBGM, Universidad de Valladolid-CSIC, Valladolid, Spain. Email: lazarill@ibgm.uva.es

Lipocalins, edited by Bo Åkerström, Niels Borregaard, Darren R. Flower
and Jean-Philippe Salier. ©2005 Eurekah.com.

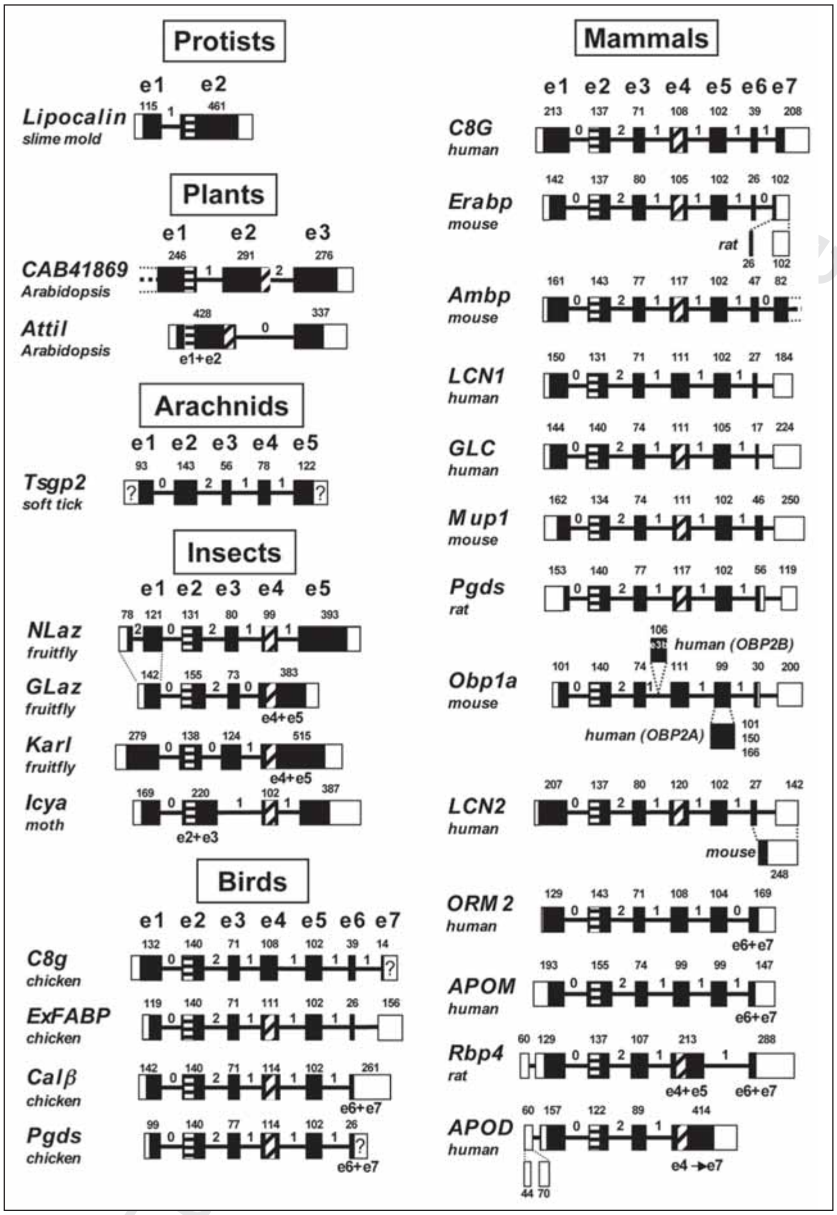

Figure 1. Intron-exon organization of lipocalin genes in eukaryotes. Only a limited series of informative genes is shown. Every gene name is identified on the left (capitals for human genes and lowercase letters for all others) and the species taken as an example is noted underneath. In a few instances of nonmammalian genes, the gene name of the mammalian ortholog is noted instead of the original name (e.g., *Pgds* instead of *Calγ* in chicken) for the sake of comparison. Introns are depicted as a solid line and they are not drawn to scale for the sake of exon comparison between genes. The intron phase is indicated between exons: 0 (intron inserted between codons) or 1 (intron after the first nucleotide in a codon) or 2 (intron in the second nucleotide in a codon). Every exon is depicted as a box with its size (bp) noted above and its coding part filled in. The figure legend is continued on the next page.

Figure 1, continued. When the 5' or 3' border of the utmost 5' or 3' exon is still undetermined, the noncoding part of the exon is given an arbitrary size and noted with a question mark, and only the size of the coding part is indicated above this exon. The archetypal lipocalin exons are noted e1-e7. Whenever possible, the exons of every gene are lined up with the archetypal exons. Note that in protoctists and plants the number of archetypal exons still remains highly speculative. A large exon noted with an arrow joining two exon numbers (e.g., ex→ez) suggest that this exon is the counterpart of the exons ex+ey+ez separately found in other genes. The genes are ordered by decreasing number of coding exons with respect to the maximal exon arrangement. When the exon arrangement of a given gene significantly varies among species, the arrangement closest to most lipocalin genes is shown in full, and some species differences are detailed underneath. The splice variants that occur within e5 of human *OBP2a* are shown, and the resulting exon sizes are noted in e5. When the 5' or 3' region of a fused gene does not code for a lipocalin-type polypeptide, this part is shown as a dotted line (e.g., *Attil, AMBP*). The canonical peptide motifs of lipocalins are shown as a stripped (GxW) or dotted (TDY) box within an exon. Absence of such a box in an exon means that the corresponding lipocalin lacks this motif. Modified from reference 1.

arthropods *vs* a maximum of seven exons (e1-e7) and six introns (A-F) in chordates, being the exon sizes and the pattern of intron phases of different lipocalins quite similar[1,2] (Figs. 1, 2). So far, too few lipocalin genes have been identified in other phyla to allow for rules of exon/intron arrangements to be inferred. Besides these features, a major tool for unambiguous exon identification is the presence in some of them of short peptide-encoding sequences. These are the well known common motifs of lipocalins (see Chapter 2), namely the GxW motif in e2, and the TDY motif in e4, that appear depicted in Figure 2, in the context of their preserved secondary structure. Because of the well aligned intron-exon boundaries in the ORF of all lipocalin family members, we can conclude that intron positions are homologous characters in this family. As an update of our previous work,[1,2] we will review structure properties of lipocalin genes in different organismal groups.

Unicellular Eukaryotes and Plants

As noted above very few lipocalin genes have been found in these groups, and therefore drawing any solid conclusion from the different exon/intron arrangement as currently found in two genes of a single plant species may be risky, nor is it possible to align the protist or plant genes with those found in higher eukaryotes. Again, the presence of the GxW and TDY motifs are the only reliable elements that allow one to propose that, for instance, the first exon in *Attil* may be a counterpart of the series of four e1 to e4 exons found in animals (see Fig. 1).

Arthropods

Exon alignment between lipocalin genes in arthropods highlights a gene arrangement with a maximum of five exons (e1-e5) / four introns (A-D), with the introns present in the 3'-end of chordate lipocalins being always absent (Fig. 1). Singular exons such as that appearing in the region coding for the signal peptide of *NLaz* are encountered as well. Also, *GLaz* and *Karl* lack the intron D that intervenes exons e4 and e5 of insects. In general, the lipocalin motifs GxW and TDY, encoded by e2 and e4 respectively, could help identify such exons. In the genes found so far in arachnids, the *TSGP1, -2*, and *-4* and the *HBP2* gene of the haematophagous ticks (*O. savignyi* and *R. appendiculatus*) both peptides motifs are absent.[3] However, the number, position and pattern of intron phases (0, 2, 1, 1) are shared with insect lipocalins (Fig. 1). This has permitted to align the tick proteins to other lipocalins[4] and include them in the phylogeny studies of the family (see below).

Vertebrates

Vertebrate lipocalins show a maximum of seven exons, usually including six coding exons (e1-e6), and the introns A-E show a fairly conserved pattern of intron phases (0, 2, 1, 1, 1) (Figs. 1, 2). Deviations from this scheme are seen in some genes which, for instance, lack intron F (*ORM2*, *APOM*), introns D and F (*RBP4*), or introns D to F (*APOD*). Yet, other more complex events have been noticed. For instance, extra noncoding 5' exons are found in *RBP4*

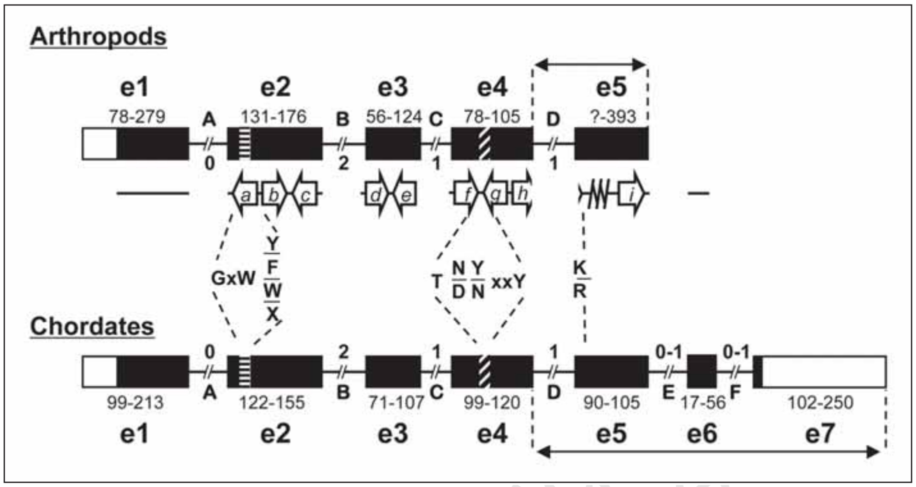

Figure 2. Lipocalin gene consensus and exon/peptide relationships in arthropods and chordates. Every exon e1 to e7 is depicted as a box whose coding part is filled in. The names (A-F) and phase (number) of introns intervening the ORF are indicated between the corresponding exons. The size range (bp) for e1-e7 is indicated (the range values are taken from Figure 1 and from other genes not shown, after excluding unusual cases of combined exons or alternative splicing). Areas where introns generate diversity in the 3' region of the lipocalin coding sequence are indicated with a double-headed arrow bordered by vertical dotted lines. An extra exon sometimes found at the 5' end of some genes as well as the variable location of the stop codon in e6 *vs* e7 in chordates (see text) are not depicted. The nine antiparallel β-strands (*a-i*) and the major C-terminal α-helix of lipocalins are indicated by head-to-tail arrows and by diagonal lines respectively. Other protein segments are depicted as horizontal lines. These elements are aligned with the corresponding gene exons. The three lipocalin motifs, namely two peptide (shown as a stripped or hatched bar within exons) and a basic amino acid residue are detailed below the appropriate β-strand or α-helix. Modified from reference 1.

and *APOD* in mammals. Also, in a limited number of cases (*ERABP* and *C8G*), the 3'-end of the ORF is intervened by intron F. Finally, a unique lipocalin gene (*AMBP*) exists (Fig. 1), which is part of a fusion of two unrelated genes whose 5' and 3' sides code for two in-frame proteins, namely the lipocalin A1m and the protease inhibitor bikunin, that is not a lipocalin. The gene region coding for A1m shows intron F at the 3'-end, and thus belongs to the group of lipocalins whose ORF is contained in 7 exons. Finally, other gene-specific events and arrangements have occurred as illustrated in Figure 1, (e.g., *Obp1a*), but they will not be commented any further herein.

Chromosomal Locations and Gene Clusters

The standard techniques of in situ hybridization and linkage studies, and the advent of information from several sequencing genome projects, have unveiled the physical mapping of lipocalins to particular chromosomes. In *Drosophila melanogaster*, three lipocalins are found in chromosome 2,[5] and one in chromosome X. The chromosomal arrangement in vertebrates displays a striking pattern, that is illustrated in Figure 2. With the exception of *APOD*, *APOM*, and *RBP4*, most human lipocalin genes are found on the long arm of chromosome 9 (HSA9q). Likewise, their orthologs in mouse and rat are clustered into two separate chromosomes that show syntenies with HSA9q.[1,6] Along these lines, a lipocalin gene cluster is also found in chicken.[7] This cluster locates onto chromosome 17 (Fig. 3), which is known to be a counterpart of HSA9.[8] Alike what is found in human, the *ApoD* and *Rbp* genes are isolated in different chromosomes in rodents and chicken. An intra-lineage duplication of *Rbp* that resulted in locations on two different chromosomes further appears in chicken.

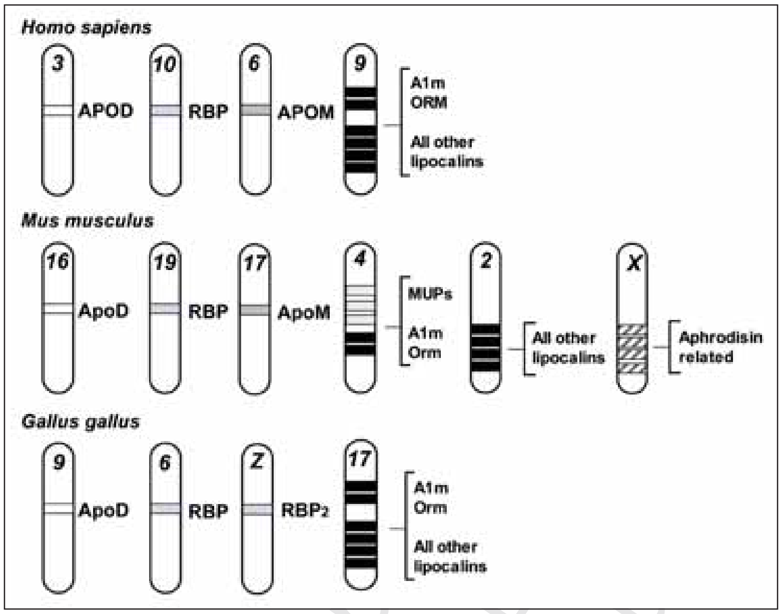

Figure 3. Chromosomal location of lipocalin genes in several chordate species. The areas of human chromosome 9, mouse chromosomes 2 and 4 and chicken chromosome 17 that harbor most lipocalin genes are known to be syntenic. The urinary proteins (*MUP* in mouse and *a2ug* in rats) of mouse chromosome 4, and the aphrodisin-related genes on mouse chromosome X are rodent-specific.

Not only are distantly related paralogs gathered in the above cluster(s) but, in many instances, more than one copy of a given lipocalin gene is further found within a cluster, which shows a trend for lipocalin genes to duplicate. This is illustrated, for example, by a duplication of the *LCN1* gene (human and rat), the *BLG-glycodelin* gene (in many mammals), a *PGDS*-like gene (chicken), the epididymal lipocalins (human and mouse), the *ORM* gene (human and mouse), and the cluster of over twenty genes in the *MUP* locus (mouse and rat). Remarkably, this trend to lipocalin gene duplication is also found in lower eukaryotes as illustrated with the *Icya* gene (in the tobacco hornworm) and the expansion of a series of lipocalins found in the saliva of hematophagous insects and chelicerates.[1,3,7,9,10] As stressed previously, these genes could not all be depicted in the limited setting of Figure 2.

Inferring the Evolution of the Lipocalin Family

Five properties can theoretically be used for inferring gene phylogenies: the gene/protein sequence, the protein three-dimensional structure, the exon-intron structure of the genes, their chromosomal position, and their organismal representation.

Only the protein tertiary structure remains to be used for reconstructing the evolutionary pathway followed by lipocalins, which is mainly due to the lack of a solid phylogenetic method that uses this information. We will now review the available phylogenetic information that can be extracted from all other properties,[2,11,12] and will update previous phylogenetic inferences with the addition of information from new lipocalins.

Lipocalin Evolution as Derived from Gene Structure Analysis

Since the gene architecture of lipocalins is well conserved (see above), the exon-intron boundaries can be assumed to be homologous, and therefore to contain traces of the evolutionary history of these genes. A new method to derive gene phylogenies from gene structure features has been developed and applied to the lipocalin family.[2] The method is based on a similarity measure of the intron-exon boundaries as mapped on a multiple protein sequence alignment of selected lipocalins with known gene structure. Three parameters are used to calculate a genetic distance: the number of introns intervening a lipocalin ORF, their phase, and the position of the exon-intron boundaries. A distance based method for phylogenetic reconstruction was then applied.[2]

Only the gene structure features present in the coding sequence are used for phylogenetic purposes. This restriction is based on the following grounds: (i) The difficulty to align the untranslated regions (UTR) of lipocalin genes, and therefore to assign homologous character to the exon-intron boundaries found in the UTR; (ii) the milder selective pressure on noncoding sequences that allows for more flexibility for independent intron gains or losses in these regions; and (iii) the presence of intron phase in the intron-exon boundaries located within the coding sequence, which is one of the phylogenetically informative characters used.

When this methodology was applied to the Lipocalin family,[2] we obtained a gene tree that is congruent with our previous protein sequence-based phylogenies of the family,[11,12] adding support and helping us refine the evolutionary history of lipocalins. This new method has also allowed Mans and Neitz[4] to include histamine-binding proteins and related sequences from several arachnids in the lipocalin phylogeny. Given the extreme divergence of these protein sequences from the lipocalin shared sequence motifs, the set of characters derived from gene structure are the most reliable approach to ascertain their relationship to lipocalins.

An updated version of a gene structure-based phylogenetic tree of lipocalins[2] is shown in Figure 4, where we have included *ApoM* as a novel member of the lipocalin family, as well as an

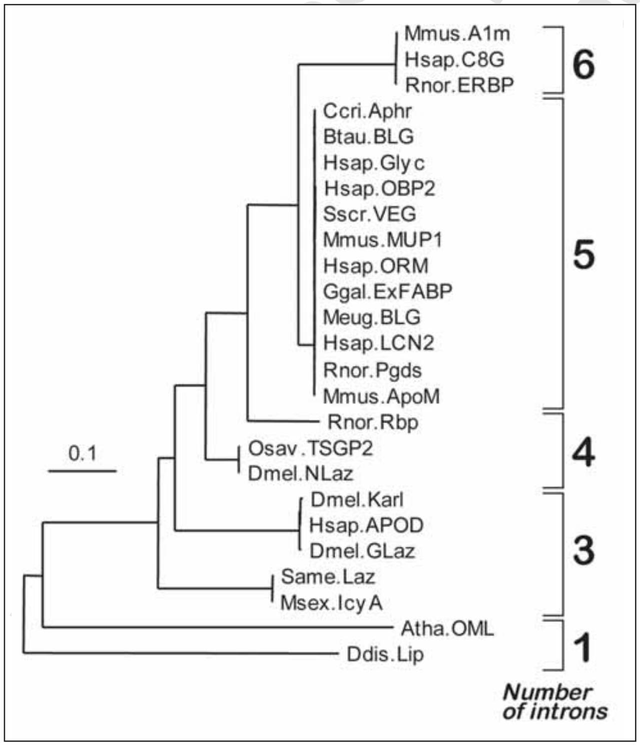

Figure 4. Lipocalin phylogeny (Neighbor-Joining) based on the exon-intron arrangement of selected genes, rooted with a protoctist lipocalin. Scale bar represents branch length (number of amino acid substitutions/site). The number of introns intervening the genes ORF are shown on the right.

example of the tick lipocalins. The tree, rooted with the lipocalin from the unicellular eukaryote Dictyostelium, shows agreement with the organismal phylogeny and sort arthropodan lipocalins in three groups, two of them with 3 introns and one group with 4 introns. The most basal chordate lipocalin in the gene tree is *ApoD*. The remaining chordate lipocalins assemble monophyletically and appear separated in three groups according to the 4-6 introns present in their ORF. The most important information that can be extracted from this tree is that lipocalins that have originated more recently contain more introns in their coding sequence. As mentioned above, introns E and F are missing in nonchordate lipocalins, while introns A-D display a broad phylogenetic distribution. In this sense, introns A and C are present in all metazoan lipocalins included in our study. Thus, the evolutionary history of metazoan lipocalins can be better traced from the distribution of introns B and D.

Lipocalin Evolution as Derived from Protein Sequence Analysis

The use of amino acid sequences and the multiple alignment of lipocalins allows us to explore their evolutionary history in more extent. Many more lipocalins can be studied in this way, given the larger dataset of mRNA and protein entries existing in molecular databases. This method also allows the addition of prokaryotic lipocalins, that could not be included in the phylogeny performed with the exon-intron arrangement.

We have aligned a set of 210 lipocalins using the same methods as in previous publications.[11,12] Namely, CLUSTAL X (1.8)[13] was used with a gap penalty mask to penalize the opening of gaps inside helices or strands, based on the secondary structure of lipocalins. Minor manual corrections were carried out based on the knowledge of lipocalin structure and function. With this alignment, we have performed a phylogenetic tree following a Bayesian approach,[14] as a newly developed method to be compared with our previous phylogenies in which a combination of neighbor-joining and maximum likelihood had been used (see refs. 11,12 for details). The use of different methodologies to reconstruct the phylogeny of lipocalins increases the confidence in the inferred relationships between extant lipocalins.

The parameters used in the Bayesian reconstruction are: 10^6 generations, a sample frequency of 10^2, a 'burn in' of 10^3, 4 chains, a mixed model, and a consensus tree representation following a 50% majority rule. The tree, rooted with the bacterial lipocalins is shown in Figure 2. Deep nodes supported by this and previous phylogeny reconstructions[11,12] are highlighted as open circles. As mentioned above, this tree includes new lipocalin members of paramount value to our phylogeny, such as those from the fungus *Debaromyces hansenii*, the nematode *Caenorhabditis elegans*, the tunicate *Ciona intestinalis*, and the cephalochordate *Branchiostoma belcheri*. Additionally, a group of Nitrophorins and of ApoM were included in our analysis.

Interestingly, the lipocalins of plants and fungi appear closely related to the bacterial lipocalins that root the tree. The protoctist dictyostelid lipocalin appears within a cluster of bacterial lipocalins. When comparing this phylogeny with our previous ones, the addition of lipocalins found in new organismal groups has resulted in the distribution of, for example, the bacterial and arthropodan lipocalins in several monophyletic groups, possibly reflecting functional relationships (see Chapter 6). Insect Nitrophorins appear grouped with a nematode lipocalin, but the tree position of this group is not well resolved possibly due to their strong sequence divergence, that could generate a long-branch attraction phylogenetic artifact.[15]

ApoD keeps being associated to a particular group of arthropodan lipocalins sharing gene expression in the nervous system, which suggest ApoD as the ancestral chordate lipocalin. This proposal gets further support from the finding that the tunicate lipocalin (Cint.Lip), which shows a basal tree position within the group of arthropodan lipocalins and ApoD, holds the highest pairwise sequence similarity with ApoD.

ApoM locates in our tree at a basal position of the chordate subtree with respect to all other chordate lipocalins. The rest of the tree does not significantly differ from our previous phylogenies, with Rbp, Blg and the Pgds-Ngal groups being most closely related to the ancestral ApoD.

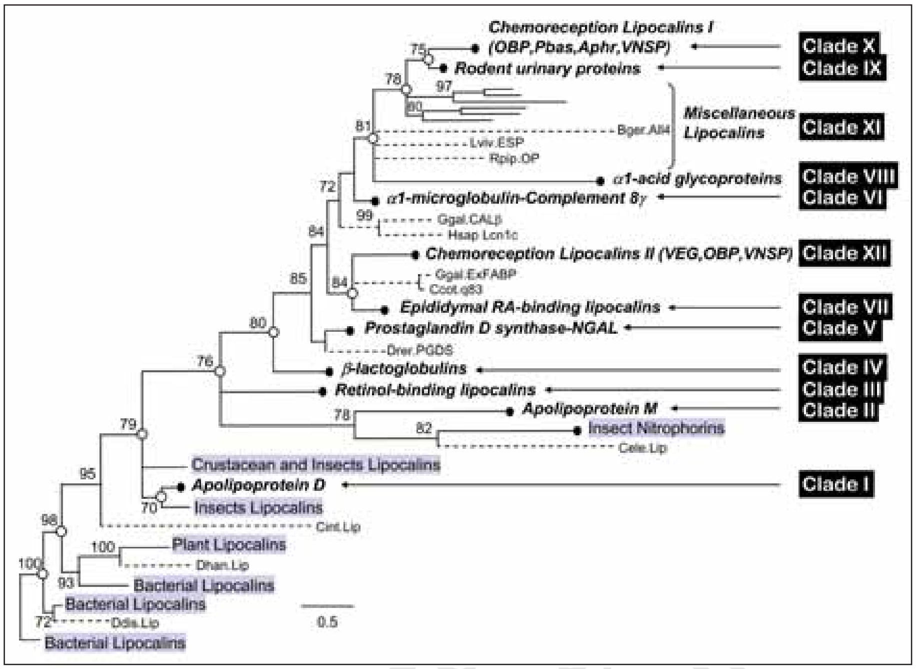

Figure 5. Phylogenetic tree of the Lipocalin family derived from a multiple alignment of 210 lipocalin protein sequences, reconstructed by a Bayesian method, and shown as a consensus tree (50% majority rule) rooted with a group of bacterial lipocalins. Posterior clade probability values (>70) are shown at each node. The scale bar represents the branch length (number of amino acid substitutions/site). Lipocalin clades from chordates are resumed to the main node, marked with a black dot, and numbered on the right. The remaining groups of lipocalins are clustered and boxed in gray. Individual lipocalins without supported grouping are shown with a dashed line, and named with an abbreviated species name and a functional label. The deep nodes marked with open circles are the ones supported by different tree building methods previously applied to the lipocalin family (see text).

As part of an ongoing project, these phylogenetic representations are helping us classify lipocalins in evolutionarily related clades. By ascribing a particular lipocalin to a clade, our goal is also to assist researchers in their experimental designs to test lipocalin functions based on the knowledge gathered from other clade members.

The lipocalin clades, numbered by roman numerals in our previous publications,[11,12] are therefore in constant refinement. Based on our updated tree, we propose a new clade arrangement (Fig. 5). This classification will hereafter only cover the well represented lipocalins of the chordate phylum, where there are several organisms with fully sequenced genomes, and that has been subjected to exhaustive lipocalin searches (ref. 16 and our continuing work). Lipocalins from other phyla, like the growing group from arthropods (see Chapter 6), need further sampling and refinement of phylogenetic associations before we can assign meaningful and reliable clade memberships. The new clade classification aims at maintaining most clade numbers as in our previous phylogenies, while accommodating the addition of the ApoM clade. Future understanding of the function of the assorted clades of so-called chemoreception and miscellaneous chordate lipocalins, as well as the ungrouped ones (labeled with dashed lines in the tree), will definitely resolve the evolutionary organization of the chordate lipocalin tree.

Finally, the phylogenetic classification also intends to help unify lipocalin nomenclature for the potential lipocalin sequences emerging from the various genome sequencing projects.

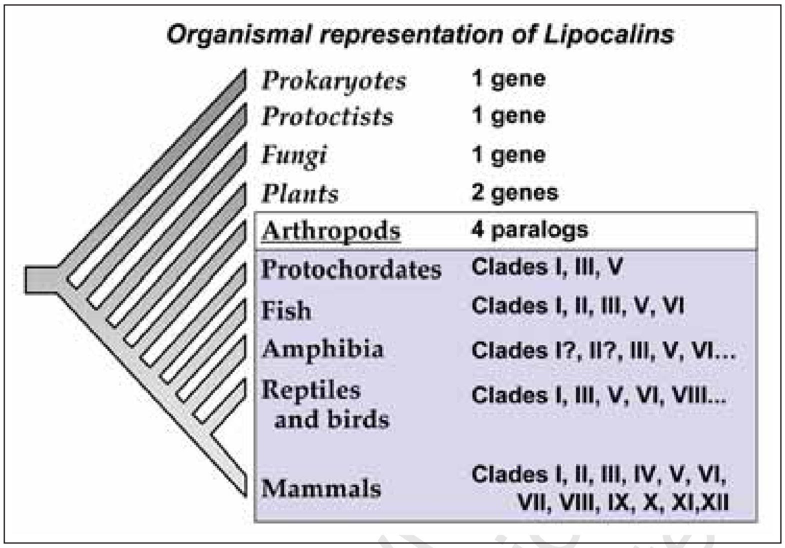
Organismal representation of Lipocalins
Prokaryotes 1 gene
Protoctists 1 gene
Fungi 1 gene
Plants 2 genes
Arthropods 4 paralogs
Protochordates Clades I, III, V
Fish Clades I, II, III, V, VI
Amphibia Clades I?, II?, III, V, VI...
Reptiles and birds Clades I, III, V, VI, VIII...
Mammals Clades I, II, III, IV, V, VI, VII, VIII, IX, X, XI,XII

Figure 6. Representation of presumed lipocalin genes in the tree of life showing the five organismal kingdoms (metazoans are boxed). Lipocalin genes are grouped in clades in the chordate phylum, and we show the clades that have been found in each chordate group so far.

Lipocalin Evolution as Derived from Chromosomal Position

As described above, the chromosomal position of many lipocalin genes has been precisely defined in the insect Drosophila genome, and in a handful of vertebrate genomes. The study of chromosomal arrangement of lipocalins in selected vertebrates, such as chicken, mouse and human (see Fig. 3) has gathered information that let us to propose hypotheses about lipocalin evolution in this phylum. The fact that both the mammalian and chicken lipocalin genes of clades IV-XII are assembled in a single chromosome, suggests that the clustering of these clades was present in the common ancestor of reptiles and birds. Further tandem duplications of some of the genes located in that cluster have occurred, which is confirmed by the similarities they show both at the protein sequence and gene structure levels (see above), and in many cases by their similar expression and function. The same reasoning applies to the *APOD* and *RBP* genes, for which we suggest a location in separate chromosomes in the first terrestrial vertebrates.

Organismal Representation and Overview of Our Hypothesis of Lipocalin Evolution

In order to present a supported hypothesis on Lipocalin evolution, besides reconstructing gene phylogenies, we need to assess their representation in the tree of life. Thus, we started by calculating the number of legitimate lipocalins present in extant taxa, without taking into account intra-species duplications. We then mapped this information onto a simplified organismal tree, which is shown in Figure 2. Using this scheme, and adding the information reported above, we can formulate a putative history of descent for lipocalins (illustrated in Fig. 7) since their appearance in the prokaryotic world.

The first thing that comes out of the analysis is that lipocalins are present in species of all five organismal kingdoms. However, a conservative number of 1 gene/species seems to be the catalog for prokaryotes, protoctists, and fungi. At least 2 lipocalin genes have been recovered from plants, but their singular exon-intron structure suggests their independent evolution. The animalia kingdom inherited the ancestral single lipocalin gene, but it subsequently duplicated

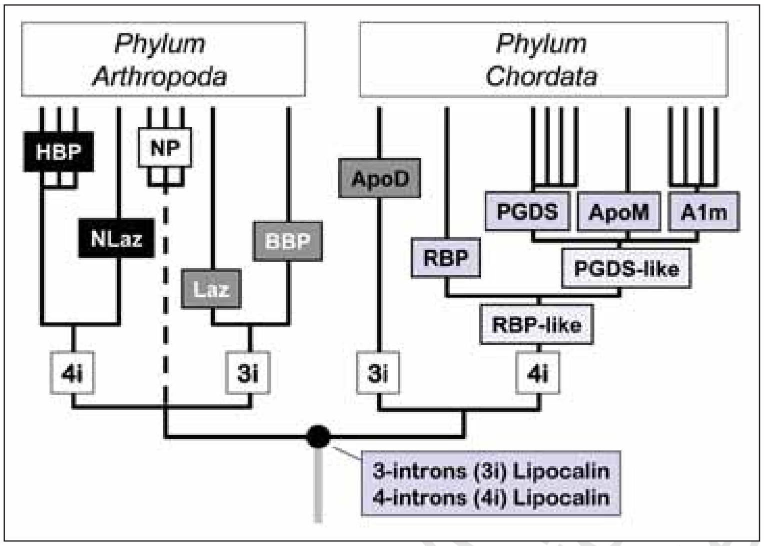

Figure 7. Hypothetical evolutionary pathway followed by arthropodan and chordate lipocalins after the split of these phyla from a common ancestor bearing two lipocalins with different gene structure.

giving rise to two genes, that evolved different gene structures with either 3 or 4 introns intervening their ORFs. Based on the commonalities observed in terms of exon-intron arrangement between the two animal phyla better sampled today, arthropods and chordates, we propose that those two lipocalin genes were present in their common ancestor (see Fig. 7).

The two ancestral lipocalins followed a different history of change in these separate phyla, that were exposed to different adaptive landscapes. In arthropods, at least four paralogs can be assumed to be common to this phylum. However, as will be reviewed in Chapter 6, the diverse arthropodan lifestyles and physiology can impose different selective constraints to the divergence of lipocalins, giving rise to particular expansions of the lipocalins repertoire, such as the tick and insect saliva proteins that are specific of blood-sucking arthropods, or the milk proteins of the viviparous cockroaches. However, the gene catalog of this phylum is increasing, and more functional and expression data are being published that will help refine the independent evolution followed by lipocalins in these organisms.

As was discussed above, all evidences coincide in ApoD being the successor of one of the ancestral lipocalins present in the first chordate-like organism. The ancestral 5-exon lipocalin of chordates was probably an RBP-like lipocalin, given the basal position of RBP in all our phylogenies. This is also supported by the presence of an RBP ortholog in the cephalochordate Branchiostoma (A. Xu, personal communication). Coincidental with the whole-scale genome duplications that occurred early during chordate evolution, the ancestral *RBP* underwent duplications, giving rise to two new lipocalins that located in separate chromosomes (see Fig. 3). These two lipocalins were possibly the ancestors of nowadays PGDS and ApoM, a proposal supported by sequence and gene-structure phylogenies, as well as the presence of these two lipocalins in fishes, and of PGDS in Branchiostoma (A. Xu, personal communication).

Several arguments point to a PGDS-like protein as the originator of a series of tandem gene duplications along the evolution of chordates, that resulted in a number of lipocalins clustered in a single chromosome: (1) a basal position for PGDS in the sequence-derived tree, (2) a gene structure similar to the duplicants, (3) its presence in every sampled chordate, and (4) a site of expression similar to ApoD.[17] In the process of *PGDS* duplications, we propose A1m as the first descendant of PGDS, as suggested by the presence of A1m in fishes.[18] Subsequent rounds

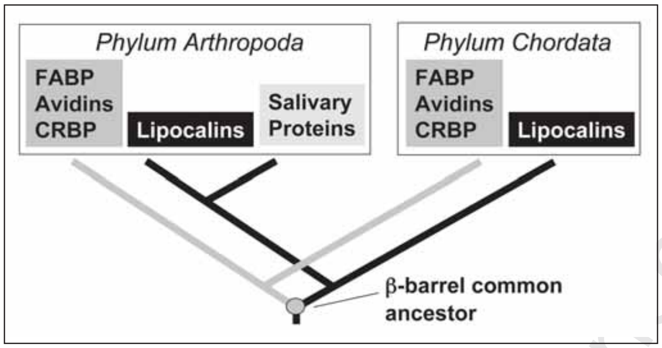

Figure 8. Evolutionary pathway of the members of Calycin protein superfamily after duplication from a hypothetical β-barrel common ancestor.

of tandem duplications of the genes coding for PGDS and A1m generated the remaining lipocalins of the cluster, following a pattern not clearly known yet. Our proposed model of lipocalin gene evolution is depicted in Figure 7, but this working hypothesis keeps being refined by our ongoing studies including information from new genomes, and about the expression and function of lipocalin genes.

As an example of the ‘evolving’ character of our own work, a lipocalin sequence has been recovered from *Hydra magnipapillata* (UG # Hma.2173) in the process of writing this review. Although not included in the tree, this sequence bears strong similarities with the tunicate lipocalin and the chordate *ApoD*. This finding is very important, as it adds a new and fairly ancient metazoan phylum, the cnidaria, where lipocalins are found and can be studied.

Lipocalins and Their Sister Families in the Calycin Superfamily

In this review we have not included in depth the phylogeny of other families composing the Calycin superfamily,[19] such as the FABP, the avidins, and the CRBP that show a compelling similarity to lipocalins in protein structure. Because of their marginal sequence (see Chapter 2) and gene structure[2] relationship, and because of their organismal representation, we propose that these families emerged from an eukaryotic common ancestral protein displaying a β-barrel structure. After that, they have diversified in different phyla (Fig. 8 illustrates this variety for the arthropodan and chordate lineages) by following divergent and independent evolutionary pathways, while folding constraints act as a selective pressure that maintains the protein structure.

Conclusions

Lipocalins are present in every organismal kingdom, and the current knowledge of their function (reviewed in other chapters of this book) suggests they have undergone striking functional diversification both after speciation and after gene duplication. Many cases are also known that point to lipocalins as moonlighter proteins,[20] able to perform more than one function at once, emphasizing the versatile nature of this protein folding.

The pathway proposed in this work for lipocalin evolution highlights the expansion of this gene family in metazoans, as well as the maintenance of the duplicated paralogous genes. However, the evolutionary mechanisms leading to the extant set of lipocalins in the two most sampled metazoan phyla (Arthropoda and Chordata) are quite distinct. A small number of lipocalins is present in most species of arthropods, while intra-lineage duplications multiply the number of lipocalins in some arthropod species adapting to new lifestyles (see Chapter 6).

Chordates also show intra-lineage gene duplications (e.g., the urinary proteins of rodents), but this phylum is characterized by a large number of paralogous lipocalins generated by large-scale duplications at the genomic level. In general these paralogs do not preserve the same protein function, as is the case for other families such as the globins or the Hox genes. The divergent protein sequences of the paralogs opened new avenues for molecular interactions (at the internal ligand-binding pocket and at the protein surface), while preserving the structural fold, and consequently increased the availability of functional pathways where to perform a novel task to be screened by natural selection.

References

1. Salier J-P. Chromosomal location, exon/intron organization and evolution of lipocalin genes. Biochim Biophys Acta 2000; 1482(1-2):25-34.
2. Sanchez D, Ganfornina MD, Gutierrez G et al. Exon-intron structure and evolution of the Lipocalin gene family. Mol Biol Evol 2003; 20(5):775-783.
3. Mans BJ, Louw AI, Neitz AWH. The major tick salivary gland proteins and toxins from the soft tick, Ornithodoros savignyi, are part of the tick Lipocalin family: Implications for the origins of tick toxicoses. Mol Biol Evol 2003; 20(7):1158-1167.
4. Mans BJ, Neitz AWH. Exon-intron structure of outlier tick lipocalins indicate a monophyletic origin within the larger lipocalin family. Insect Biochem Mol Biol 2004; 34(6):585-594.
5. Sanchez D, Ganfornina MD, Torres-Schumann S et al. Characterization of two novel lipocalins expressed in the Drosophila embryonic nervous system. Int J Dev Biol 2000; 44(4):349-359.
6. Chan P, Simon-Chazottes D, Mattei MG et al. Comparative mapping of lipocalin genes in human and mouse: The four genes for complement C8 gamma chain, prostaglandin-D-synthase, oncogene-24p3, and progestagen-associated endometrial protein map to HSA9 and MMU2. Genomics 1994; 23(1):145-150.
7. Pagano A, Giannoni P, Zambotti A et al. Phylogeny and regulation of four lipocalin genes clustered in the chicken genome: Evidence of a functional diversification after gene duplication. Gene 2004; 331:95-106.
8. Consortium ICGS. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 2004; 432:695-716.
9. Suzuki K, Lareyre J-J, Sanchez D et al. Molecular evolution of epididymal lipocalin genes localized on mouse chromosome 2. Gene 2004; 339:49-59.
10. Ribeiro JMC, Andersen J, Silva-Neto MAC et al. Exploring the sialome of the blood-sucking bug Rhodnius prolixus. Insect Biochem Mol Biol 2004; 34(1):61-79.
11. Ganfornina MD, Gutierrez G, Bastiani M et al. A phylogenetic analysis of the lipocalin protein family. Mol Biol Evol 2000; 17(1):114-126.
12. Gutierrez G, Ganfornina MD, Sanchez D. Evolution of the lipocalin family as inferred from a protein sequence phylogeny. Biochim Biophys Acta 2000; 1482(1-2):35-45.
13. Thompson JD, Gibson TJ, Plewniak F et al. The ClustalX windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tool. Nucleic Acids Res 1997; 24:4876-4882.
14. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 2003; 19(12):1572-1574.
15. Philippe H, Laurent J. How good are deep phylogenetic trees? Curr Opin Genet Dev 1998; 8:616-623.
16. Salier J-P, Akerstrom B, Borregaard N et al. Lipocalins in bioscience: The first family gathering. BioEssays 2004; 26(4):456-458.
17. Ganfornina MD, Sánchez D, Pagano A et al. Molecular characterization and developmental expression pattern of the chicken Apolipoprotein D gene. Implications for the evolution of vertebrate lipocalins. Dev Dyn 2005; 232:191–199.
18. Akerstrom B, Logdberg L, Berggard T et al. alpha(1)-Microglobulin: A yellow-brown lipocalin. Biochim Biophys Acta 2000; 1482(1-2 SU -):172-184.
19. Flower DR. Structural relationship of streptavidin to the calycin protein superfamily. FEBS Lett 1993; 333(1-2):99-102.
20. Jeffery CJ. Moonlighting proteins. Trends Biochem Sci 1999; 24:8-11.