

Review

The lipocalin protein family: structural and sequence overview

Darren R. Flower^{a,*}, Anthony C.T. North^b, Clare E. Sansom^c^a The Edward Jenner Institute for Vaccine Research, Compton, Newbury, Berkshire RG20 7NN, UK^b School of Biochemistry and Molecular Biology, University of Leeds, Leeds, West Yorkshire LS2 9JT, UK^c Department of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK

Abstract

Lipocalins are remarkably diverse at the sequence level yet have highly conserved structures. Most lipocalins share three characteristic conserved sequence motifs – the kernel lipocalins – while others are more divergent family members – the outlier lipocalins – typically sharing only one or two. This classification is a useful tool for analysing the family, and within these large sets are smaller groups sharing much higher levels of sequence similarity. The lipocalins are also part of a larger protein superfamily: the calycins, which includes the fatty acid binding proteins, avidins, a group of metalloproteinase inhibitors, and triabin. The superfamily is characterised by a similar structure (a repeated +1 topology β -barrel) and by the conservation of a remarkable structural signature. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Lipocalin; Fatty acid binding protein; Avidin; Structural comparison; Characteristic motif; Protein superfamily

1. Introduction

The special edition of *Biochimica Biophysica Acta*, of which this paper forms a small part, describes, in detail, the biological function of many members of the lipocalin protein family. The lipocalin family is a large, and ever expanding, group of proteins exhibiting great structural and functional variation, both within and between species [1]. Lipocalins are typically small (160–180 residues in length), extracellular

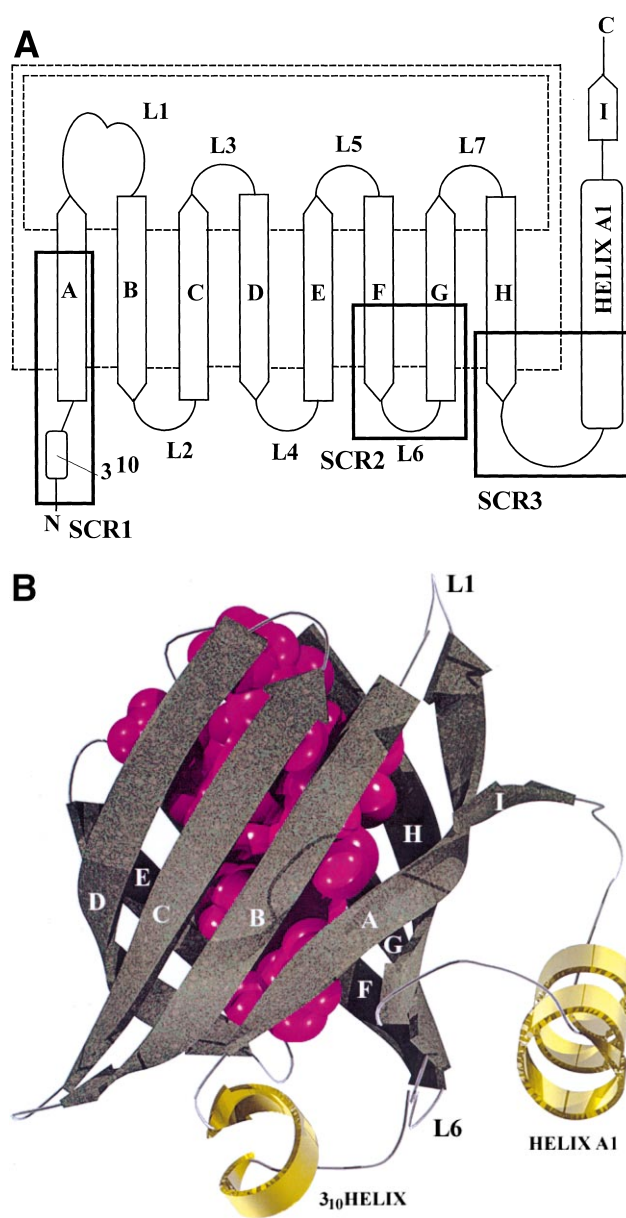
proteins sharing several common molecular recognition properties: the binding of small, principally hydrophobic molecules (such as retinol); binding to specific cell-surface receptors; and the formation of covalent and non-covalent complexes with other soluble macromolecules. Although they have been classified mainly as transport proteins, it is now clear that members of the lipocalin family fulfil a wide variety of different functions.

Despite many common characteristics and common functions, membership of the lipocalin family has been defined largely on the basis of sequence, or structural, similarity and it has grown to encompass a large corpus of proteins. Within this the lipocalins display unusually low levels of overall sequence conservation, with pairwise comparisons often falling below 20%, the nominal threshold for a reliable alignment. However, all lipocalins share sufficient similarity, in the form of short characteristic conserved sequence motifs, to form the basis of a useful definition of family membership [2,3]. Most

Abbreviations: FABP, fatty acid binding proteins; Blg, β -lactoglobulin; MPI, metalloproteinase inhibitors; OBP, odorant binding protein; Apod, apolipoprotein D; SCR, structurally conserved region; PYP, photoactive yellow protein; MUP, mouse major urinary protein; BBP, bilin binding protein; AGP, α -1-acid glycoprotein; MIR, multiple isomorphous replacement; VEGP, von Ebner's gland protein; LESP, lizard epididymal secretory protein

* Corresponding author. Fax: +44-1635-577901;
E-mail: darren.flower@jenner.ac.uk

Fig. 1. Structure of the lipocalin fold. (A) An unwound view of the lipocalin fold orthogonal to the axis of the barrel. The nine β -strands of the antiparallel β -sheet are shown as arrows and labelled A–I. The C-terminal α -helix A1 and N-terminal 3_{10} like helix are also marked. Connecting loops are shown as solid lines and labelled L1–L7. A pair of dotted lines indicates the hydrogen-bonded connection of two strands. One end of the lipocalin β -barrel has four loops (L1, L3, L5, and L7); the opening of the internal ligand-binding site is here and so is called the Open end of the molecule. The other end has three β -hairpin loops (L2, L4 and L6); the N-terminal polypeptide chain crosses this end of the barrel to enter strand A via a conserved 3_{10} helix closing this end of the barrel: the Closed end of the molecule. Those parts that form the three main structurally, and sequence, conserved regions (SCRs) of the fold (SCR1, SCR2, and SCR3) are marked as heavy boxes. SCR3 corresponds closely with the sequence conserved region rather than the structurally conserved region. (B) A schematic or ribbon-drawing of the lipocalin fold. The structure shown is a prototypical, rather than an actual, structure. The nine conserved β -strands, the closed β -barrel, and the N-terminal and C-terminal helices are shown (see A above). Loops L1 and L6 are also marked. The size of the ligand binding ultrasite is represented by a collection of atoms, in a purple space-filling depiction, representing the structural overlay of a variety of lipocalin ligands which include retinol, retinoic acid, haem, and bilin. β -Strands are shown as smoothly curving arrows textured as grey granite, α -helices as golden spiral ribbons, and loops as silver cords. Figure generated using ALTER and POVray [57].



lipocalins share three characteristic conserved sequence motifs – the kernel lipocalins – while other more divergent family members – the outlier lipocalins – typically share only one or two.

Lipocalins of known three-dimensional structure include a group of kernel lipocalins: retinol-binding protein (RBP) [4], β -lactoglobulin (Blg) [5], insecticyanin [6], bilin-binding protein (BBP) [7], major urinary protein [8], α_{2u} -globulin [8,9], epididymal retinoic acid-binding protein [10], and neutrophil lipocalin [11]. There are also a number of outlier lipocalin groups including odorant-binding protein [12,13], Bos d2 allergen [14], nitrophorin [15], and a histamine binding protein from the tick *Rhipicephalus appendiculatus* [16].

The common structure of the lipocalin protein fold is now well-described [1,3,17]. The lipocalin fold is a highly symmetrical all- β structure dominated by a single eight-stranded antiparallel β -sheet closed back on itself to form a continuously hydrogen-bonded β -barrel. In cross-section, this has a flattened

or elliptical shape. The β -barrel encloses a ligand-binding site composed of both an internal cavity and an external loop scaffold. It is this diversity of cavity and scaffold that gives rise to a variety of different binding modes each capable of accommodating ligands of different size, shape, and chemical character. The eight β -strands of the barrel, labelled A–H (see Fig. 1), are linked by a succession of +1 connections, giving it the simplest possible β -sheet topology. The seven loops, labelled L1–L7, are all typical of short β -hairpins, except loop L1: this is a large Ω loop. Loop L1 forms a lid folded back to

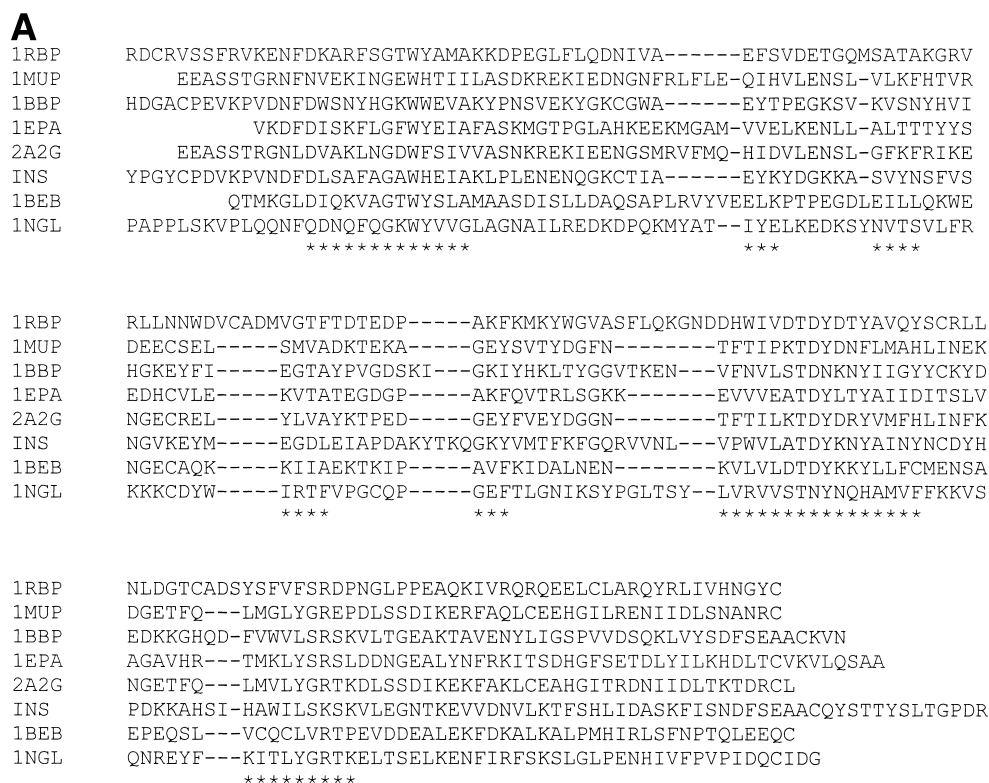


Fig. 2. Structural alignment of the lipocalin protein family. (A) Structure-based sequence alignments of those kernel lipocalin structures which have been determined experimentally by crystallography and multidimensional NMR. Equivalent positions within the lipocalin common core are indicated by an asterisk. Labels are of the protein databank files from which the structures were taken. 1MUP, mouse major urinary protein; 1NGL, neutrophil lipocalin; INS, insecticyanin; 1BBP, bilin binding protein; 1EPA, epididymal retinoic acid binding protein; 1RBP, retinol binding protein; 1BEB, β -lactoglobulin; 2A2G, α_{2u} -globulin. (B) The set of superimposed α traces corresponding to the alignment shown in A. 1MUP, yellow; 1NGL, light blue; INS, cyan; 1BBP, green; 1EPA, magenta; 1RBP, red; 1BEB, blue; 2A2G, gold. (C) Structure-based sequence alignments of those outlier lipocalin structures which have been determined experimentally by crystallography. Equivalent positions within the lipocalin common core are indicated by an asterisk. Labels are of the protein databank files from which the structures were taken. 1NP1, nitrophorin; 1BJ7, bos d 2; 1A3Y, odorant binding protein. (D) The set of superimposed α traces corresponding to the alignment shown in C. Colours are given by: 1NP1, blue; 1BJ7, green; 1A3Y, red.

close partially the internal ligand-binding site found at this end of the barrel. Between strand H and the short terminal strand I is an α -helix; this is an ever-present feature of the lipocalin fold but is not conserved in its position relative to the axis of the β -barrel nor in its length (see Fig. 1). Previous work has analysed the conservation of sequence and structure in the lipocalin protein family [1,8]. These accounts show how the common core characteristic of the lipocalin fold is dominated by three large structurally conserved regions (SCRs): SCR1 (strand A and the 3_{10} -like helix preceding it), SCR2 (strands F and G, and loop L6 linking them), and SCR3 (strand H and adjoining residues). Other SCRs of the common core are small and can be neglected

(see Fig. 2). The three principal SCRs each contain a sequence motif that is wholly, or partly, invariant.

Together with three other distinct protein families: the fatty-acid-binding proteins (FABPs), avidins, metalloproteinase inhibitors (and the presently enigmatic triabin), the lipocalin family forms part of a larger structural superfamily: the calycins [3,18,19]. This is an example of a 'structural superfamily': a set of proteins with closely related three-dimensional structures that show no significant overall similarity at the sequence level. In this paper, we shall review and update structure and sequence relations within the lipocalins, and also those between the lipocalins and other members of the calycin protein superfamily.



Fig. 2 (continued).

2. Structural relationships to the calycin protein superfamily

The lipocalins, the fatty acid-binding proteins (FABPs), and the avidins – three families of ligand-binding proteins – together form the calycin protein superfamily. The FABPs are a family of predominantly intracellular proteins involved in lipid metab-

olism [20]. The avidins are proteins with a remarkable affinity for the vitamin biotin, which have found important applications in biotechnology [21]. Recently, the size of the calycin protein superfamily has been enlarged to incorporate a group of metalloprotease inhibitors (MPIs) [22] and triabin [19] (see Fig. 3 and Table 1). Both MPIs and triabin function as proteinase inhibitors, as do some lipocalins [23].

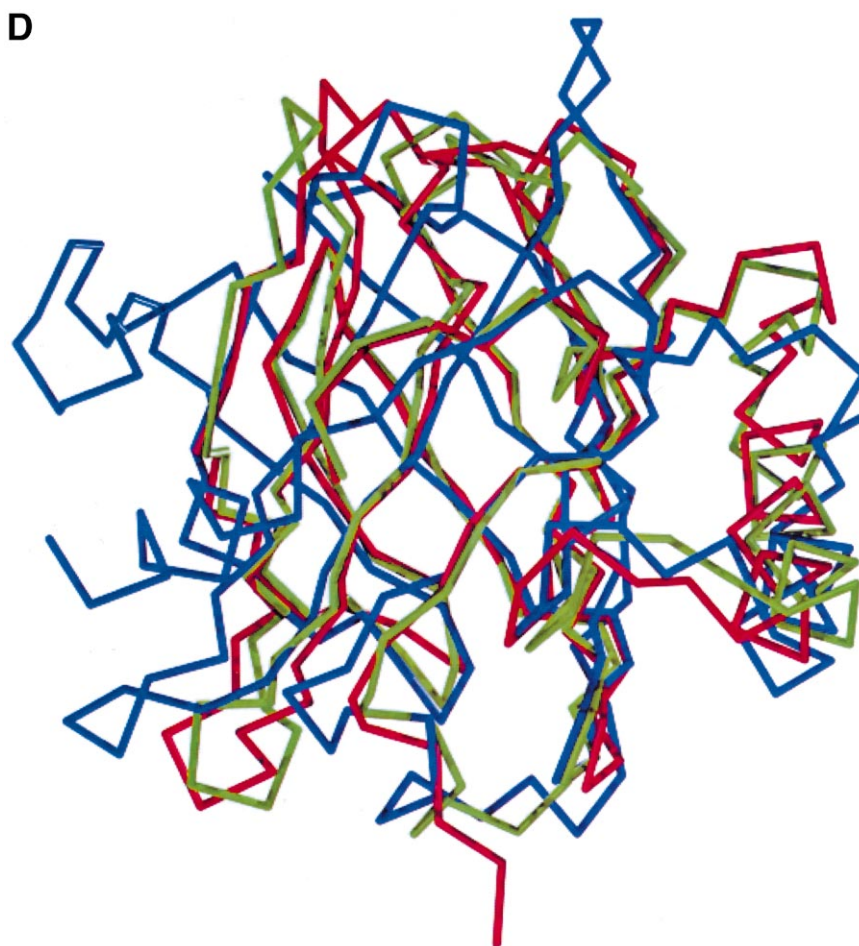


Fig. 2 (continued).

The structure of triabin is presently something of an enigma. Strands B and C are interchanged in position, perturbing the repeated +1 topology and antiparallel arrangement of adjacent strands. Dis-

Table 1

Number of aligned positions and root mean squared deviations (RMSD) from pairwise comparisons of representative members of the calycin protein superfamily

	1SMP	1ACD	1AVG	1RAV	1MUP
1SMP	—	1.70	1.70	1.84	1.64
1ACD	42	—	1.63	1.38	1.63
1AVG	39	46	—	1.50	1.76
1RAV	55	54	52	—	1.73
1MUP	54	62	46	66	—

1MUP, mouse major urinary protein; 1RAV, chicken avidin; 1SMP, metalloproteinase inhibitor; 1ACD, mouse adipocyte lipid binding protein; 1AVG, triabin. Upper triangle is RMSD and the lower triangle is number of aligned positions.

counting the possibility of a chain mistracing, then this may represent an interesting example of an uncommon topological transformation between related folds. Comparing parts A and B of Fig. 4 shows that, by allowing reverse matching of structural segments rather than exclusively ordered matching, large proportions of the triabin and lipocalin structures can be equivalenced. These topological differences between triabin and other members of the calycin protein family have generated some interest [24,25]. There are examples from eight-membered parallel α/β -barrels, where strands interconvert to yield a barrel of mixed topology [26], although the evolution of these barrels is deemed to be primarily convergent. Similar topological transformations are much more common in non-barrel structures. For example, the comparison of serine/threonine protein kinase (PK) and type II β phosphatidylinositol phosphate kinase [27] shows

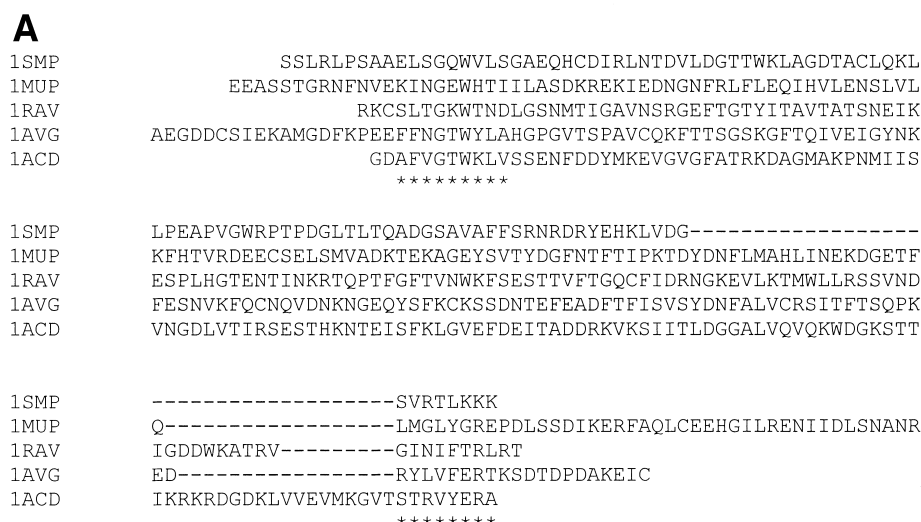


Fig. 3. Structural alignment of the calycin protein superfamily. (A) Structure-based sequence alignments of a representative set of calycin protein structures which have been determined experimentally by crystallography. Equivalent positions within the calycin common core are indicated by an asterisk. The overall alignment is concentrated in the N and C termini. Very short structurally conserved regions (<3) residues have been excluded. Sequence and/or structure similarity indicated by individual alignments was considerably greater between pairs of structures. Labels are of the protein databank files from which the structures were taken. 1MUP, mouse major urinary protein; 1RAV, chicken avidin; 1SMP, metalloproteinase inhibitor; 1ACD, mouse adipocyte lipid binding protein; 1AVG, triabin. (B) The set of superimposed C α traces corresponding to the alignment shown in A. 1MUP, green; 1RAV, blue; 1SMP, red; 1ACD, magenta; 1AVG, yellow.

topological variation between closely related folds. The possible implications of such changes for protein evolution are not clear, but they may require some relaxation in our present perceptions of the relatedness of protein structures in the face of topological change. At the sequence level, global similarity searches suggest there is some similarity of triabin to nitrophorin (an outlier lipocalin), and rather more to a *Rhodnius prolixus* salivary platelet aggregation inhibitor. Greater elucidation will come with the discovery of more sequences clearly related to triabin.

The calycins share a related β -barrel structure. The FABP family has a ten-stranded discontinuous β -barrel structure; the avidin and MPI barrels, although eight-stranded, are more circular in cross-section than that of the lipocalins and do not have a C-terminal helix or strand I; while triabin has a similar barrel geometry yet has a variant topology. The N- and C-terminal strands of the FABPs and lipocalins can be closely superimposed, with the loss (FABP to lipocalin) or gain (lipocalin to FABP) of two central strands necessary to effect the transformation of one to another [3]. Moreover, beyond

some functional similarity (hydrophobic ligand binding and/or macromolecular interaction) these families are characterised by a similar folding pattern (an antiparallel β -barrel dominated by a largely +1 topology), within which, large parts of their structures can be structurally equivalenced, although the families share no global sequence similarity (see Fig. 3). Fig. 5 shows a tree relating, in a thematic way, the main members of the calycin superfamily.

Previous work [28] has shown that members of the calycin superfamily share a distinct structural pattern. An arginine or lysine residue (from the last strand of the β -barrel) which forms hydrogen bonds to the main-chain carbonyl groups of the N-terminal 3_{10} -like helix and packs across a conserved tryptophan (from the first strand of the β -barrel). This pattern can be seen both in the structures of kernel lipocalins, which also share a conserved interaction from loop L6, and in the more structurally diverse outlier lipocalins. It is also apparent in the other four families comprising the calycins. Examination of the available structures of streptavidin and chicken avidin, the metalloproteinase inhibitor from *Erwinia chrysanthemi*, and the structure of triabin, all reveal

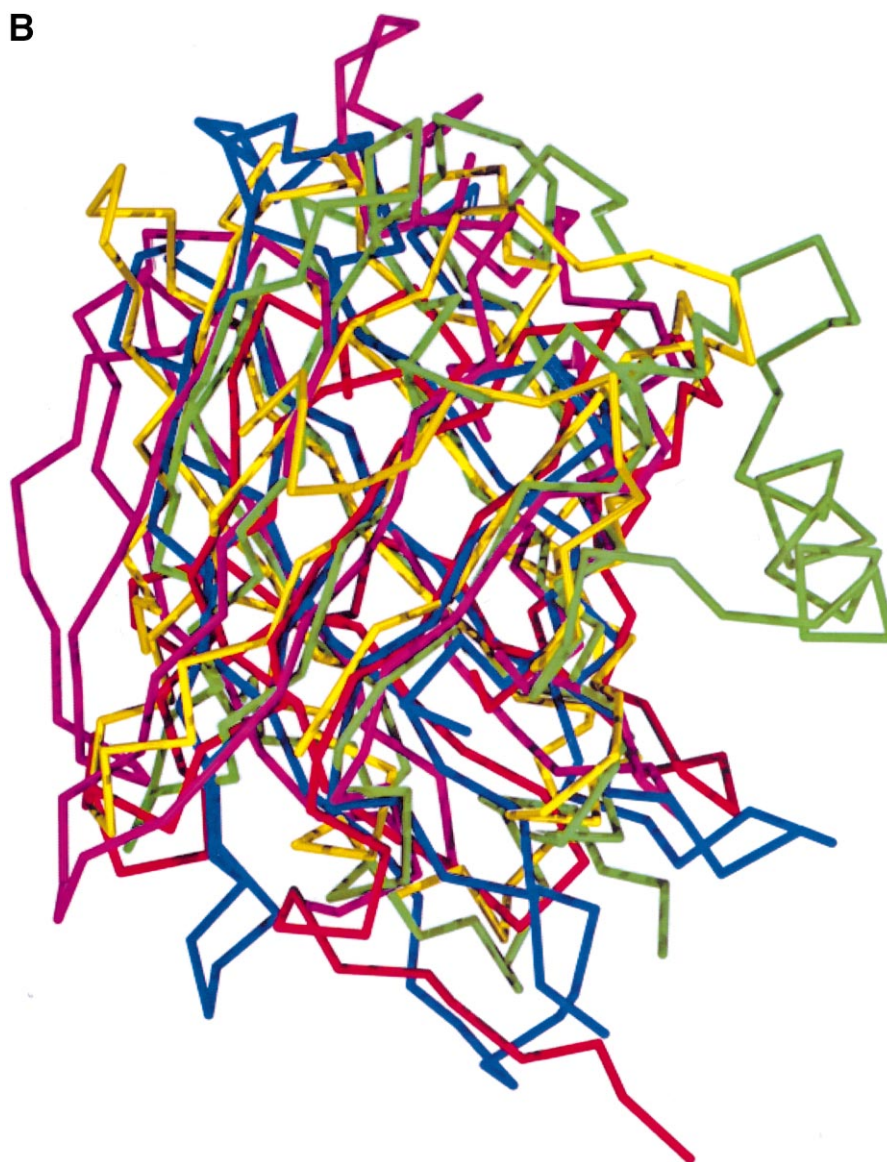


Fig. 3 (continued).

a very similar arrangement of interacting residues. Most of the known FABPs have an arrangement of side chain interactions similar to those described above, in which a tryptophan, from the first strand of the FABP barrel, packs against an arginine from near the end of the last. This feature is, however, lacking from a group of more highly diverged FABPs, typified by insect muscle FABPs. See Fig. 6.

This distinct structural signature (in which an arginine or lysine, able to form a number of potential hydrogen bonds with the main chain carbonyls of a short 3_{10} helix, packs across a conserved tryptophan

in a structurally superimposable, non-random manner) corresponds to sequence determinants common to the calycin member families: a characteristic N-terminal sequence pattern, displaying the complete preservation of key residues, and a weaker C-terminal motif. Fig. 7 shows the sequence patterns centred on these two interactive residues. Although conservation is apparent from these patterns, it has not yet proved possible to design sequence discriminators able to reliably identify all calycins, as has been done for the lipocalins [2,3]. Nonetheless, the preservation of particular residues and interactions, across

A

```

1AVG  AEGDDCSIEKAMG-DFKPEEFFNGTWTYLAHGPGVTSPAVCQKFTTSGSKGFTQIVEIGYN--KFE-----
1MUP  EEASST-----GRNFNVEKI-NGEWHTIILASDKREKI-----EDNGNFRFLFLE
Code  .....=====,=====.....

```

```

1AVG  -----S-----NVKFQCNQVDNKN-GEQYSFKCKSSDNTEFEA--DFTFISVSYDNFALVC
1MUP  QIHVLENSLVLFHTVRDEECSELSMVADKT-EKAGEYSVT-----YDGFNTFTIPKTDYDNFLMAH
Code  .....=====,.....=====

```

```

1AVG  RSITFTSQP---KEDRYLVFERTKSDTDPDAKEIC-----
1MUP  LINE----KDGETFQLMGLYGREPDLSSDIKERFAQLCEEHGILRENIIDLSNANRC
Code  ==.....=====,.....

```

B

Triabin GenFitted to MUP

14 ASP to	21 PHE aligned with	13 ASN to	20 ASN
23 GLY to	28 ALA aligned with	21 GLY to	26 ILE
40 GLN to	55 GLU aligned with	60 PHE to	45 PHE *
64 ASN to	72 VAL aligned with	68 CYS to	76 ASP
79 GLN to	82 PHE aligned with	83 GLU to	86 VAL
97 THR to	11 ARG aligned with	95 THR to	109 LEU
122 ASP to	132 SER aligned with	119 GLN to	128 PRO

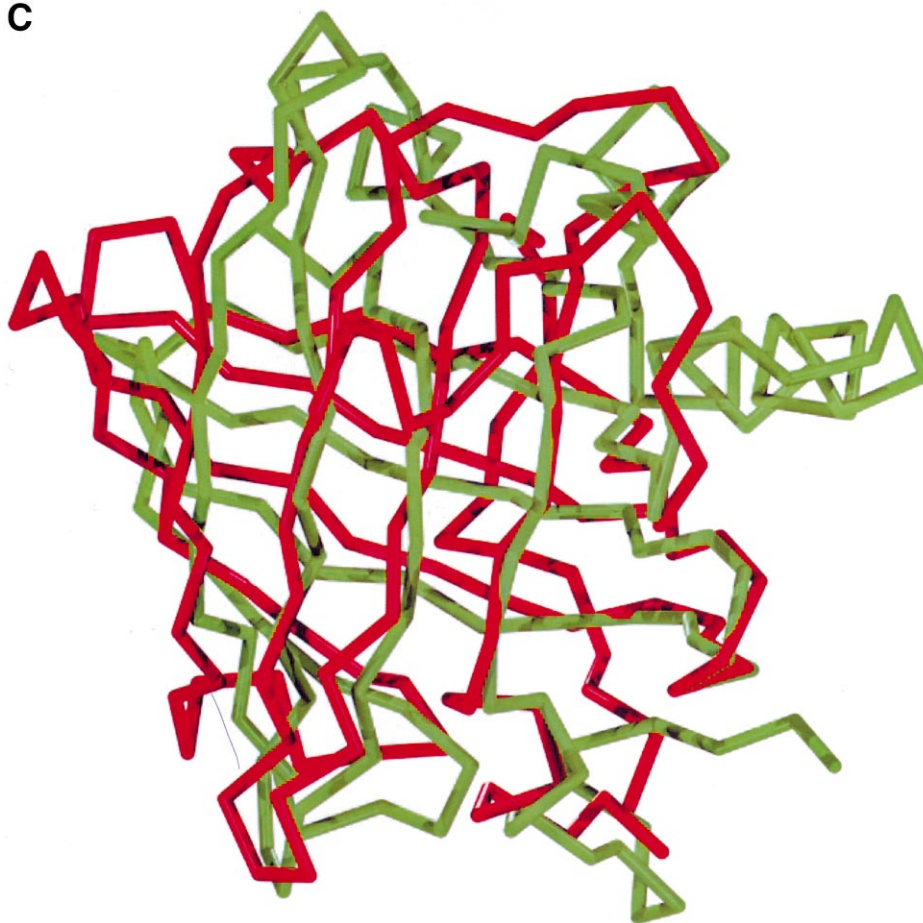
C

Fig. 4. Structural alignment of triabin and a representative lipocalin structure. (A) Straightforward ordered alignment of triabin and Mouse major urinary protein. (B) Equivalent alignment of triabin and Mouse major urinary protein using the computer program GENFIT [58]. This program allows continuous stretches of amino acids to match in any order and also allows reverse matching. (C) The pair of superimposed C α traces corresponding to the alignment shown above. MUP is shown in green; triabin is shown in red. Clearly the topological differences between MUP and triabin mask the high degree of structural similarity they share. In addition to these topological differences, triabin also lacks the large loop L1, a C-terminal α -helix, and short strand I. The sequence similarities in structurally conserved regions of the fold pose a number of intriguing questions: is triabin an extreme lipocalin outlier, a distinct fold, or the product of a chain mistracing?

the member families lends some support to the view that there was a common, if very distant, evolutionary origin for the calycin superfamily.

2.1. Structures without relatedness

Over the years, many proteins have, on the basis of real or perceived structural similarity, been suggested as having a relationship with the lipocalins. For some – the fatty acid binding proteins, avidins, metalloproteinase inhibitors, and now triabin [2,3,19] – this comparison is valid and is borne out by de-

tailed structural analysis. For other proteins, however, this appears not to be the case.

Photoactive yellow protein (PYP) was the first such suggested relationship [29]. The initial description of the structure suggested the protein to be composed of an open antiparallel β -barrel, with a repeated +1 topology, similar to FABP. However, the subsequent higher resolution MIR structure proved this initial chain tracing to be wrong [29]; PYP is actually closer, structurally, to the SH2 domain and profilin. Since that time, there have been a number of other suggestions ranging from new struc-

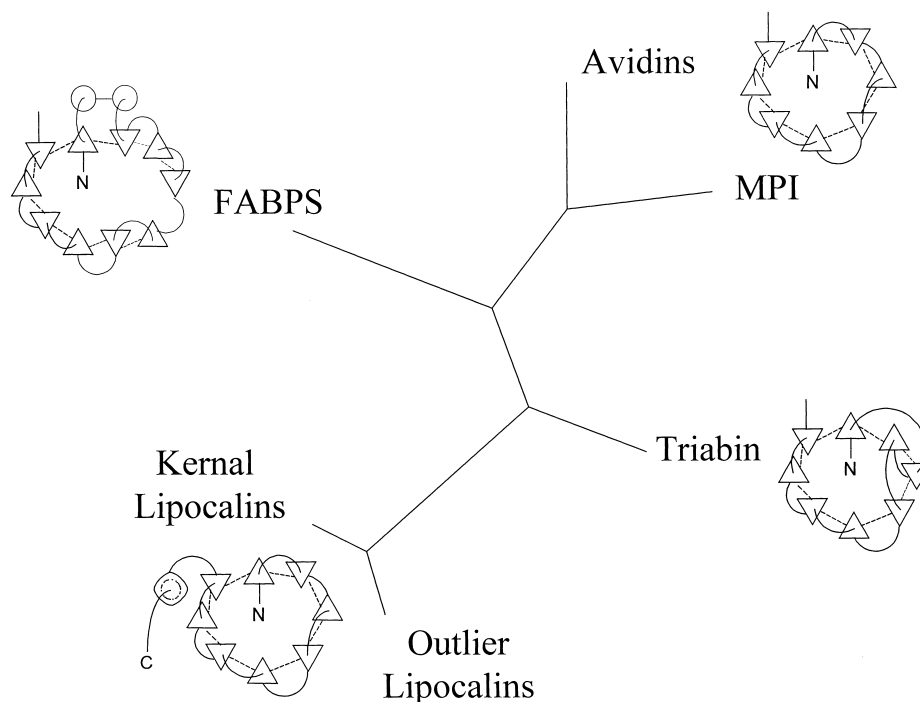


Fig. 5. Structural tree. A conceptual phylogenetic tree of the calycin protein superfamily, representing the structural and sequence similarities in its members. β -Strands are shown as triangles. Triangles pointing downwards indicate a strand direction into the plane of the paper and those pointing upwards indicate a strand direction out of the plane of the paper. The view shown, down the axis of the barrel, is orthogonal to that in Fig. 1. Connecting loops are shown as continuous lines. The topology of the tree succinctly summarizes the relative similarity within the superfamily. Likewise, while the branch lengths of the tree are arbitrary, they represent, in a highly approximate manner, the relative similarity of different folds.

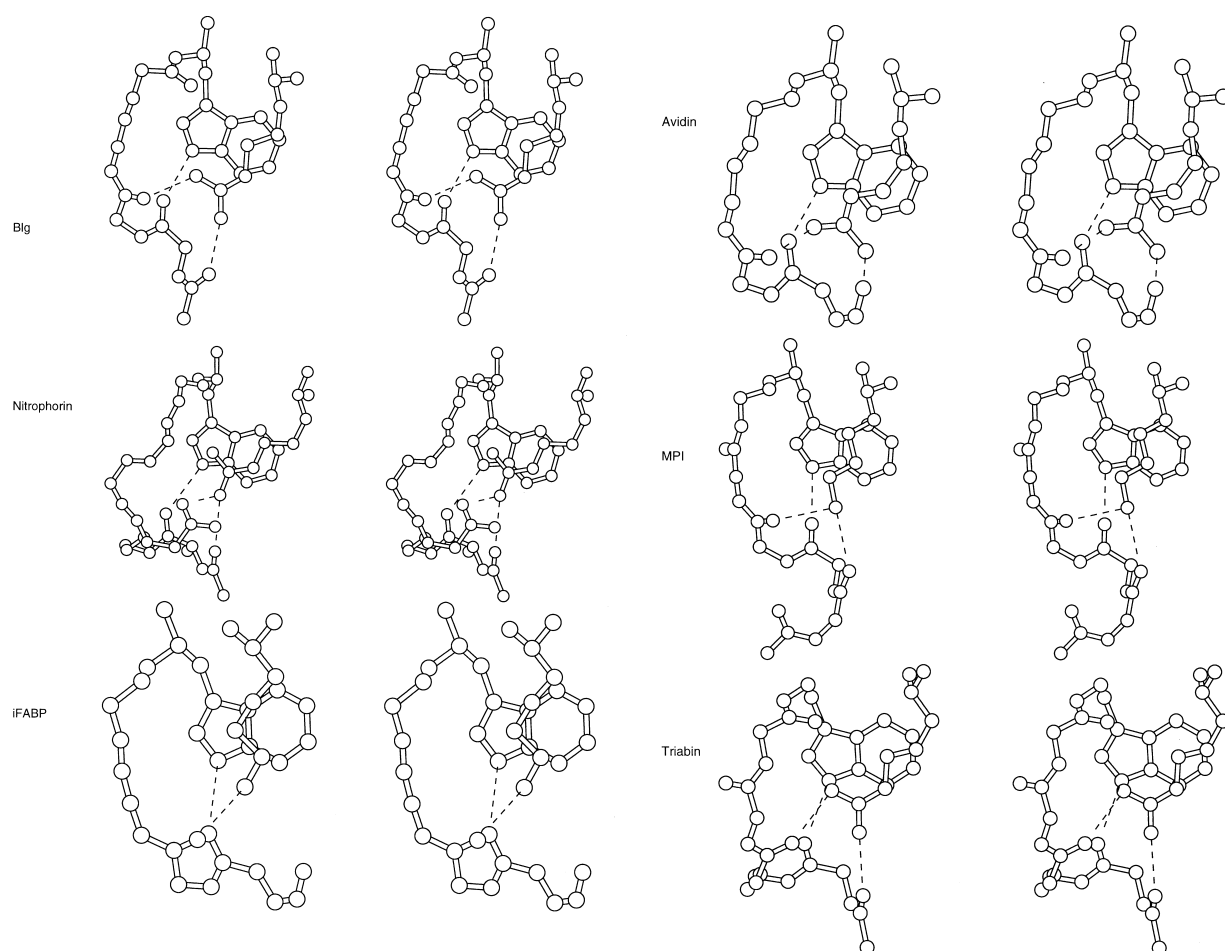


Fig. 6. Representative plots of the conserved structural signature characteristic of the calycin fold. (a) Kernel lipocalin: β -lactoglobulin. (b) Outlier lipocalin: nitrophorin. (c) Fatty acid-binding protein. (d) Avidin. (e) Metalloproteinase inhibitor from *Erwinia chrysanthemi*. (f) Triabin.

tures such as TCL-1 and MTCP-1 through to old favourites such as catalase [30,31]. Although the geometry and shear number of the catalase barrel is very similar to that of avidin and MPI, the barrels of calycin family members dominate their structures, whereas the barrel in catalase is embedded in a much larger structure involving a folding pattern not consistent with the modularity seen in certain avidin analogues [1]. Moreover, the exterior of the catalase barrel forms part of a haem binding site; a feature not present in any calycin structure. The sequence patterns discussed above are also absent in catalase. Although one cannot entirely refute a relationship between calycins and catalase, it seems, in our view, a remote possibility.

There are many more examples, where analogies,

in which one has varying degrees of confidence, have been drawn based on a greater or lesser degree of similarity in topology, size, and binding site. Cyclophilin is another interesting example of a structurally similar protein, which differs dramatically in its topology [32]. The most telling example is, perhaps, the pleckstrin homology domain. The similarity of this protein to a remarkable diversity of other proteins (including lipocalins, verotoxin, and FK-506 binding protein) has been noted [33].

At a superficial level, the calycins do show some overall similarity to certain other barrel-like all- β structures [34], and it is true that β -structures, particularly β -barrels, are easily confused. This results, in part, from the convergent similarities arising from the principles that seem to underlie barrel design:

	MOTIF1	MOTIF2
Triabin	fFNSTWVLa-----vFSTTs	
MPI	eLSGQWVLS-----rTLTa	
Chick Avidin	sLTSGWVNd-----iFTLt	
Human RetBP	nFACTWMr-----iVVV--	
Human MyeP2	kFLGTWLV-----iVSV--	
Bovine Blg	kVAGTWVSL-----cLVSTTe	
Mouse MUP	kINGEWFSi-----lVGTAd	
Human A1M	rIVGFWVNI-----lVGTApq	
Human NGAL	qFQGWVTV-----lVGTTe	
Human RBP	rFSGTWVAm-----vFSTDen	
Insect BBP	nyVGTWVsv-----vLSTSVv	
Human ApoD	kYLGTWYFi-----iLSTNEN	
Ecoli Blc	rVLTGWYFi-----iLSTTPT	
Bovine OBP	eLSGFWTV-----nVDTGq	
Human VEGP	dVSTGWVlk-----lVGTDPk	
Human A1GP	qITGFWFi-----yADDET	
nitrophorin	yFSGTWVt-----vLNTNd	

Fig. 7. Sequence motifs centred on the conserved residues of the calycin structural signature. Two motifs are shown centring on the conserved tryptophan and the arginine residues. Both are marked with an asterisk. The glycine position is completely conserved, while some others, such as G-2 (MOTIF1) and R-2 (MOTIF2), show a very strong conservation of hydrophobic and aromatic residues. Other positions, such as W-1 (MOTIF1), have a conserved polar character. Other positions are not conserved. The sequences chosen are typical of the calycin families. For simplicity, the relatively small proportion of lipocalin sequences that deviate from this pattern are not shown. MOTIF1 corresponds to SCR1 and MOTIF2 to SCR3 from the lipocalin multiple structural alignment.

only certain barrel geometries, shear numbers etc., are possible. Indeed, the authors of SCOP [35] have discounted the relationship of avidins and MPIs to the lipocalin and FABPs for this reason. However, similarities or differences in gross structural characteristics are insufficient to prove or disprove divergent or convergent evolution and one is obliged to look for synergy between clues from both sequence and structure in order to characterise fully such relationships.

3. Relations within the lipocalin family

Lipocalins have been found predominantly in eukaryotic organisms, mostly in vertebrates, although some have been identified in other phyla. Apart from vertebrates, which are over-represented for historical and pragmatic reasons, the best exemplified phylum is arthropoda, where proteins include butterfly insecticynin, grasshopper lazaro [36], cockroach Bla g 4

protein [37], lobster crustacyanin [38], a lipocalin from *Leucophaea maderae* [39], nitrophorin from *Rhodnius prolixus* [15], and histamine binding protein from *Rhipicephalus appendiculatus* [16]. Other calycons – FABPs, avidins, and MPIs – are present in vertebrates, invertebrates, including Arthropoda and Coelenterata, and prokaryotes. There have also been reports of lipocalins from plants [42]. Ganfornia et al. [43,44] have suggested that lipocalins are also present in the cellular slime mould *Dictyostelium discoideum* and *Drosophila*. There are now several long established prokaryotic lipocalins (*Scherichia coli*, *Citrobacter freundii* and *Vibrio cholerae*) [40,41]. A more recent analysis of microbial genomes by Bishop [45] indicates that lipocalins can be found in at least 14 species of Gram-negative bacteria.

The existence of prokaryotic lipocalins, has, potentially, profound implications for our understanding of lipocalin evolution. Although the possibility of horizontal transfer cannot be excluded, it suggests that the lipocalins are very much older and more widespread than has been supposed. To determine if the bacterial lipocalins result from horizontal transfer, Ganfornia et al. [43] estimated the G+C content in the first and third codon positions of bacterial gene samples. Their results showed no significant bias in the computed % G+C of the bacterial lipocalin genes, indicating no evidence for the recent acquisition for bacterial lipocalins by horizontal transfer. Bishop [45] suggests that horizontal transfer has occurred not to bacteria from multi-cellular organisms, as assumed above, but from bacteria to eukaryotes, perhaps at the time of the endosymbiotic uptake of mitochondria. Such transfers are seen as increasingly common in the history of life [46,47].

Lipocalins have yet to be found in the third main branch of the tree of life: the archaeobacteria. Possibly the genomes of archaeobacteria, and other microorganisms, are too small or have become too specialised to expect to find lipocalins easily. One explanation hinges on the observation that bacterial lipocalins are lipoproteins anchored to the outer membrane of Gram-negative bacteria. So, it may be expected that lipocalins, of this type, would be absent in Gram-positive bacteria and archaeobacteria, both lacking an outer membrane. However, absence of evidence is seldom evidence of absence, and it need not be stressed further that many lipocalins, and re-

lated calycins, are very hard to detect at the sequence level [15,16] and we must ultimately rely on crystallography for their initial identification. Thus there may be many more microbial lipocalins waiting to be found.

3.1. Kernel and outlier lipocalins

Harpaz and Chothia [48] have used the term ‘set’ to describe distant subdivisions within large protein families. We have previously identified the ‘kernel set’ of lipocalin sequences, which match the LIPOCALIN fingerprint in the PRINTS database more or less completely, and distinguished these from a number of other, even more divergent ‘outlier sets’. Kernel lipocalins, which form by far the largest self-consistent sub-group within the whole group of related sequences, share three conserved sequence motifs corresponding to the three main structurally conserved regions of the lipocalin fold.

There is considerable difficulty in producing a phylogeny of all lipocalin sequences. The sequences are not evenly spread through the many bifurcating branches of the tree of life, but are, for the most part, clustered into a few, recently diverged species. The optimal multiple sequence alignment of the lipocalin protein family also contains a large number of indels. Conservation across the family is so low that only within particular regions can the alignment, and thus derived phylogenetic distances, be established with certainty. Thus all attempts to derive a fine-detailed phylogeny of the family must be viewed with some scepticism [43,44]. Nonetheless, it is possible to produce a coarse-grained view based on sequence comparisons. For example, a pseudo-phylogenetic tree of the kernel lipocalins is shown in Fig. 8. It can be sub-divided into three lobes. Proteins in lobe 1 include prostaglandin D synthase, neutrophil lipocalin, and α 1-microglobulin. Lobe 2 has two branches: one branch is formed from retinol binding proteins and purpurin; the other from invertebrate colouration proteins; such as insecticyanin, crustacyanin, and gallerin; lazearillo; apolipoprotein D (apoD); and prokaryotic lipocalins. Retinol binding proteins have a disulphide-bonding pattern containing three bridges not found in other lipocalins, while bilin-binding proteins (BBP) and apoD have a characteristic two-bond pattern. Many other lipocalins

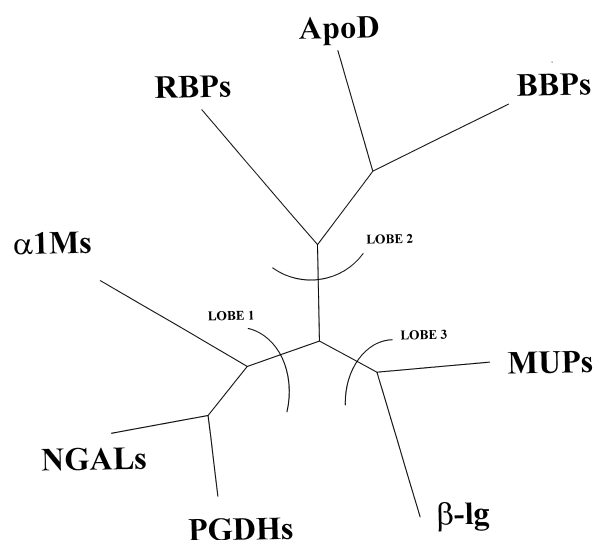


Fig. 8. Unrooted phylogenetic tree of the kernel lipocalins. A conceptual phylogenetic tree describing sequence relations within the kernel set of lipocalins. Major sequence groups are shown. The topology of the tree succinctly summarizes the relative similarity within the kernel lipocalin protein family. Likewise, while the branch lengths of the tree are arbitrary, they reflect, in a highly approximate way, the relative similarity of different sequences.

contain only one disulphide bridge. Lobe 3 is composed primarily of rodent urinary proteins (rat α _{2u}-globulin and mouse major urinary proteins (MUPs)) and β -lactoglobulins. These results are broadly in-line with the phylogenetic analysis presented by Ganfornia et al. [43,44].

The outlier lipocalins, which match no more than two of the three diagnostic motifs, are more diverse. They form distinct subsets at the sequence level: the odorant-binding proteins, Von Ebner's-gland proteins, α 1-acid glycoproteins (AGPs), nitrophorins, and tick histamine binding proteins. The first two outlier groups show considerably more similarity to the kernel lipocalins, in the sense of their overall sequence similarity (as assessed by, say, the program BLAST), than do the other three. The odorant binding proteins form a group of sequences comprising odour binding proteins (OBP), aphrodisin, probasin, and the major bovine dander allergen Bd a 20. The Von Ebner's gland proteins comprise a group containing vomeronasal proteins, OBP-II, the major dog allergen Can f 1, and the Von Ebner's gland proteins or tear lipocalins. The α 1 acid glycoprotein of hu-

man and other species forms a discrete subgroup with the outlier lipocalins. As well as a variety of immunological functions, human AGP is one of the most significant non-specific drug binders in human blood. The two remaining outlier groups are the Nitrophorins [15] and the Tick Histamine binding proteins [16]. Again, they form distinct, discrete groupings within the outliers, with only very distant relationships to other members of the family. Indeed their relationship to the lipocalins was first revealed by X-ray crystallography.

The bacterial lipocalins highlight an increasing problem in sequence analysis of the lipocalin family. Although they lack the second of the lipocalin motifs, their similarity to apolipoprotein D and related sequences is otherwise strong. Other sequences, for example horse allergen protein Equ c 1 and marsupial β -lactoglobulin and late lactation proteins, also clearly lack the three key sequence motifs but nonetheless show similarity, albeit somewhat distant, to the β -lactoglobulins and MUPs. Certain other lipocalins show even more remote similarities. The *Leuconophaea maderae* lipocalin and cockroach allergen protein both have some similarity to the BBP/ApoD group, but at great phylogenetic distances, even for the lipocalins. The plant sequences [42] are another phylogenetically isolated group. This may mean that they represent a new class of outlier lipocalin, or, it may transpire, they may not even be lipocalins at all.

4. Discussion

The lipocalin protein family has continued to grow, both in terms of sequences and structures determined, but also in their diversity and interest. In this brief review, we have focused our attention primarily on new relationships within the calycin superfamily and on a more careful appreciation of sequence relationships within the lipocalins.

The kernel-outlier division has proved a useful tool in the analysis of the lipocalins accounting for most sequences of the lipocalin family. The existence of outliers (sequences lacking the three key motifs) which are strongly similar to groups of kernel sequences is an increasing problem, but one that tends to affect limited numbers of sequences, and at the

present state of knowledge, particularly single sequences. This problem may increase, or decrease, as the phyletic distribution of lipocalins increases. One way around this problem, is to use the first of the three characteristic motifs, which is shared by almost all known lipocalins, as a single diagnostic of family membership. However, the diagnostic power of such approach varies considerably. PROSITE [49] – a deterministic database system – uses a regular expression, corresponding to this region, as a sequence signature characteristic of the family. Using this against SWISS-PROT produces 30 false negatives (including α 1 acid glycoproteins, late lactation proteins, VEGPs, and LESP IV), 70 true positives, and 65 false positives. More helpful results are found using PRINTS [50–53], where a probabilistic model in combination with the diagnostic power of multiple motifs considerably reduces noise.

Members of the calycin protein family share a β -barrel structure. The lipocalins, FABPS, and avidins also share the ability to bind ligands with a strongly hydrophobic character, although this does not extend to the apparent functions of the MPIs or triabin. Thus it is interesting to note that Russell et al., who looked for a close superimposition of bound ligands within a structural superfamily, did not identify the calycins as an example of a supersite [54]. This is a consequence of their structural diversity, the location of bound calycin ligands is more spatially diffuse [17]. Ligand binding in the lipocalins may be more like that observed in the G protein-coupled receptors, which again bind a remarkable range of molecules but in an extended ultrasite rather than in a confined supersite [55] (see Fig. 1b).

As well as an overall structural similarity, the calycin proteins show conserved main chain conformations, amino acid side chains, and the interactions they make, which together forms a structural signature characteristic of the superfamily. In particular, an arginine or lysine, able to form a number of potential hydrogen bonds with the main chain carbonyls of a short 3_{10} helix, and which packs across a conserved tryptophan. Seeking for the moment a figurative allusion, we can say that these conserved interactions act to pin together the two ends of the calycin β -barrel. But what role does this key structural motif play? Might it have some functional role: perhaps as a protein–protein recognition site [3]?

Might it stabilise the overall protein structure and help maintain the β -barrel fold? Or, as a key interaction in the folding pathway, could it guide the formation of the β -barrel? In the absence of further data this remains an open question. Several sequences of as yet undetermined structures deviate from this pattern. The late lactation proteins, for example, have a tyrosine exchanged for tryptophan and what appears to be a proline at the arginine position. A number of FABPs show a tryptophan to tyrosine, and even tryptophan to phenylalanine, substitution. Here an aromatic to arginine/lysine interaction is conserved, although structural data would suggest that it is not as strong as for tryptophan. There are several more such anomalies, but they are relatively small in number compared to the size of the whole calycin superfamily. It is not clear from exceptions such as these, whether these changes may make some form of compensating interactions within the fold or are examples of drift in protein evolution. If some proteins can tolerate these changes, then why, against a background of extreme sequence divergence across the family, is it needed by most?

As suggested by this non-universality, there may be an alternative explanation. In present day structures, the sequences that give rise to the individual folding patterns of each family may have evolved stability independent of this interaction of residues. A protein sequence will fold into a unique structure. Very different sequences, however, can fold into similar structures. As Rost reports [56], only 3–4% of amino acids appear to be crucial for protein structure and function. The mean identities for divergent and convergent evolution of proteins – evolution from the same or different ancestors – with similar structures are quite close and hence, it is difficult to distinguish between them. Low levels of overall sequence identity are thus not necessarily indicative of convergent evolution. All then that we see, in these structures, is, perhaps, an echo, a palimpsest, of a past interaction: still important but no longer essential: an evolutionary relic of the common calycin ancestor protein. Nonetheless, retention of common, characteristic sequence signatures corresponding to a highly conserved pattern of side chain packing is supportive of the conjecture that the calycons (lipocalins, FABPs, avidins, MPIs, and triabin) have a common evolutionary origin.

5. Conclusion

The classification of the lipocalins as kernel and outliers is a simple, yet powerful, definition. Nonetheless, despite its classificatory power there are examples, such as the late lactation proteins, that have proved exceptions, as they lack the Trp–Arg interaction characteristic of almost all calycons. As the number of calycons identified as members of the superfamily continues to increase, it seems likely that it will become even more diverse. Nonetheless, with this growth will come a new maturity in our knowledge of and insights into this family, combining these apparent inconsistencies into a new and deeper subtlety of understanding.

Acknowledgements

We thank the two anonymous referees for their helpful and constructive comments, which have doubtless strengthened this paper. We should also like to thank Dr T.K. Attwood, Professor J.-P. Sallier, Dr B. Akerstrom, and Dr R.E. Bishop for helpful discussions.

References

- [1] D.R. Flower, The lipocalin protein family: structure and function, *Biochem. J.* 318 (1996) 1–14.
- [2] D.R. Flower, A.C.T. North, T.K. Attwood, Mouse oncogene protein-24p3 is a member of the lipocalin protein family, *Biochem. Biophys. Res. Commun.* 180 (1991) 69–74.
- [3] D.R. Flower, A.C.T. North, T.K. Attwood, Structure and sequence relationships in the lipocalins and related proteins, *Protein Sci.* 2 (1993) 753–761.
- [4] S.W. Cowan, M.E. Newcomer, T.A. Jones, Crystallographic refinement of human serum retinol binding-protein at 2 angstroms resolution, *Proteins Struct. Funct. Genet.* 8 (1990) 44–61.
- [5] S. Brownlow, J.H.M. Cabral, R. Cooper, D.R. Flower, S.J. Yewdall, I. Polikarpov, A.C.T. North, L. Sawyer, Bovine beta-lactoglobulin at 1.8 angstrom resolution – still an enigmatic lipocalin, *Structure* 5 (1997) 481–495.
- [6] H.M. Holden, W.R. Rypniewski, J.H. Law, I. Rayment, the molecular structure of insecticyanin from the tobacco hornworm *Manduca sexta* L. at 2.6 Å resolution, *EMBO J.* 6 (1987) 1565–1570.
- [7] R. Huber, M. Schneider, I. Mayr, R. Muller, R. Deutzmann, F. Suter, H. Zuber, H. Falk, H. Kayser, Molecular structure of the bilin binding-protein (BBP) from *Pieris brassicae* after

- refinement at 2.0-Å resolution, *J. Mole. Biol.* 198 (1987) 499–513.
- [8] Z. Bocskei, C.R. Groom, D.R. Flower, C.E. Wright, S.E.V. Phillips, A. Cavaggioni, J.B.C. Findlay, A.C.T. North, Pheromone binding to 2 rodent urinary proteins revealed by X-ray crystallography, *Nature* 360 (1992) 186–188.
- [9] B.N. Chaudhuri, G.J. Kleywegt, J. Bjorkman, L.D. Lehmanmckeeman, J.D. Oliver, T.A. Jones, The structures of $\alpha_2\mu$ -globulin and its complex with a hyaline droplet inducer, *Acta Cryst. D55* (1999) 753–762.
- [10] M.E. Newcomer, Structure of the epididymal retinoic acid-binding protein at 2.1 angstrom resolution, *Structure* 1 (1993) 7–18.
- [11] M. Coles, T. Diercks, B. Muehlenweg, S. Bartsch, V. Zolzer, H. Tschesche, H. Kessler, The solution structure and dynamics of human neutrophil gelatinase-associated lipocalin, *J. Mol. Biol.* 289 (1999) 139–157.
- [12] M. Tegoni, R. Ramoni, E. Bignetti, S. Spinelli, C. Cambillau, Domain swapping creates a third putative combining site in bovine odorant binding protein dimer, *Nat. Struct. Biol.* 3 (1996) 863–867.
- [13] S. Spinelli, R. Ramoni, S. Crolli, J. Bonicel, C. Cambillau, M. Tegoni, The structure of the monomeric porcine odorant binding protein sheds light on the domain swapping mechanism, *Biochemistry* 37 (1998) 7913–7918.
- [14] J. Rouvinen, J. Rautiainen, T. Virtanen, T. Zeiler, J. Kauppinen, A. Taivainen, R. Mantyjarvi, Probing the molecular basis of allergy – three-dimensional structure of the bovine lipocalin allergen bos D 2, *J. Biol. Chem.* 274 (1999) 2337–2343.
- [15] J.F. Andersen, A. Weichsel, C.A. Balfour, D.E. Champagne, W.R. Montfort, The crystal structure of nitrophorin 4 at 1.5 angstrom resolution: transport of nitric oxide by a lipocalin-based heme protein, *Structure* 6 (1998) 1315–1327.
- [16] G.C. Paesen, P.I. Adams, K. Harlos, P.A. Nuttall, D.I. Stuart, Tick histamine-binding proteins: isolation, cloning, and three-dimensional structure, *Mol. Cell* 3 (1999) 661–671.
- [17] D.R. Flower, Multiple molecular recognition properties of the lipocalin protein family, *J. Mol. Recogn.* 8 (1995) 185–195.
- [18] D.R. Flower, Structural relationship of streptavidin to the calycin protein superfamily, *FEBS Lett.* 333 (1993) 99–102.
- [19] P. Fuentesprior, C. Noeskejungblut, P. Donner, W.D. Schleuning, R. Huber, W. Bode, Structure of the thrombin complex with triabin, a lipocalin-like exosite-binding inhibitor derived from a triatomine bug, *Proc. Natl. Acad. Sci. USA* 94 (1997) 11845–11850.
- [20] L. Banaszak, N. Winter, Z.H. Xu, D.A. Bernlohr, S. Cowan, T.A. Jones, Lipid-binding proteins – a family of fatty-acid and retinoid transport proteins, *Adv. Protein Chem.* 45 (1994) 89–151.
- [21] N.M. Green, Avidin and streptavidin, *Methods Enzymol.* 184 (1990) 51–67.
- [22] U. Baumann, M. Bauer, S. Letoffe, P. Delepelaire, C. Wandersman, Crystal-structure of a complex between *Serratia marcescens* metalloprotease and an inhibitor from *Erwinia chrysanthemi*, *J. Mol. Biol.* 248 (1995) 653–661.
- [23] W. Vanthof, M.F.J. Blankenvoorde, E.C.I. Veerman, A.V.N. Amerongen, The salivary lipocalin Von Ebner's gland protein is a cysteine proteinase inhibitor, *J. Biol. Chem.* 272 (1997) 1837–1841.
- [24] A.G. Murzin, How far divergent evolution goes in proteins, *Curr. Opin. Struct. Biol.* 8 (1998) 380–387.
- [25] D.R. Flower, A topological nomenclature for protein structure, *Protein Eng.* 11 (1998) 723–727.
- [26] T.M. Larsen, J.E. Wedekind, I. Rayment, G.H. Reed, A carboxylate oxygen of the substrate bridges the magnesium ions at the active site of enolase, *Biochemistry* 3 (1996) 4349–4358.
- [27] N.V. Grishin, Phosphatidylinositol phosphate kinase: a link between protein kinase and glutathione synthase folds, *J. Mol. Biol.* 291 (1999) 239–247.
- [28] D.R. Flower, A structural signature characteristic of the calycin protein superfamily, *Protein Pept. Lett.* 2 (1995) 341–346.
- [29] G.E.O. Borgstahl, D.R. Williams, E.D. Getzoff, 1.4 Angstrom structure of photoactive yellow protein, a cytosolic photoreceptor - unusual fold, active-site, and chromophore, *Biochemistry* 34 (1995) 6278–6287.
- [30] Z.Q. Fu, G.C. Dubois, S.P. Song, I. Kulikovskaya, L. Virgilio, J.L. Rothstein, C.M. Croce, I.T. Weber, R.W. Harrison, Crystal structure of MTCP-1: implications for role of TCL-1 and MTCP-1 in T cell malignancies, *Proc. Natl. Acad. Sci. USA* 95 (1998) 3413–3418.
- [31] R.B. Russell, M.J.E. Sternberg, A novel binding site in catalase is suggested by structural similarity to the calycin superfamily, *Protein Eng.* 9 (1997) 107–111.
- [32] J. Kallen, C. Spitzfaden, M.G.M. Zurini, G. Wider, H. Widmer, K. Wuthrich, M.D. Walkinshaw, Structure of human cyclophilin and its binding-site for cyclosporine-A determined by X-ray crystallography and NMR-spectroscopy, *Nature* 353 (1991) 276–279.
- [33] C.A. Orengo, M.B. Swindells, A.D. Michie, M.J. Zvelebil, P.C. Driscoll, M.D. Waterfield, J.M. Thornton, Structural similarity between the pleckstrin homology domain and verotoxin – the problem of measuring and evaluating structural similarity, *Protein Sci.* 4 (1995) 1977–1983.
- [34] A.V. Efimov, A structural tree for proteins containing 3 beta-corners, *FEBS Lett.* 407 (1997) 37–41.
- [35] A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia, Scop: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.* 247 (1995) 536–540.
- [36] D. Sanchez, M.D. Ganfornina, M.J. Bastiani, Developmental expression of the lipocalin lazarillo and its role in axonal pathfinding in the grasshopper embryo, *Development* 121 (1995) 135–147.
- [37] L.K. Arruda, L.D. Vailes, M.L. Hayden, D.C. Benjamin, M.D. Chapman, Cloning of cockroach allergen, Bla G 4, identifies ligand binding proteins (or calycins) as a cause of

- IgE antibody responses, *J. Biol. Chem.* 270 (1997) 31196–31201.
- [38] J.N. Keen, I. Caceres, E.E. Eliopoulos, P.F. Zagalsky, J.B.C. Findlay, Complete sequence and model for the C1 subunit of the carotenoprotein, crustacyanin, and model for the dimer, beta-crustacyanin, formed from the C1 and A2 subunits with astaxanthin, *Eur. J. Biochem.* 202 (1991) 31–40.
- [39] A. Korchi, R. Brossut, H. Bouhin, J. Delachambre, cDNA cloning of an adult male putative lipocalin specific to tergal gland aphrodisiac secretion in an insect (*Leucophaea maderae*), *FEBS Lett.* 449 (1999) 125–128.
- [40] D.R. Flower, C.E. Sansom, M.E. Beck, T.K. Attwood, The first prokaryotic lipocalins, *Trends Biochem. Sci.* 20 (1995) 498–499.
- [41] R.E. Bishop, S.S. Penfold, L.S. Frost, J.V. Holtje, J.H. Weiner, Stationary phase expression of a novel *Escherichia coli* outer membrane lipoprotein and its relationship with mammalian apolipoprotein D. Implications for the origin of lipocalins, *J. Biol. Chem.* 270 (1995) 23097–23103.
- [42] R.C. Bugos, A.D. Hieber, H.Y. Yamamoto, Xanthophyll cycle enzymes are members of the lipocalin family, the first identified from plants, *J. Biol. Chem.* 273 (1998) 15321–15324.
- [43] M.D. Ganfornina, G. Gutiérrez, M. Bastiani, D. Sánchez, A Phylogenetic analysis of the lipocalin protein family, *Mol. Biol. Evol.* 17 (2000) 114–126.
- [44] G. Gutiérrez, M.D. Ganfornina, D. Sánchez, Evolution of the lipocalin family as inferred from a protein sequence phylogeny, *Biochim. Biophys. Acta* 1482 (2000) 35–45.
- [45] R.E. Bishop, The bacterial lipocalins, *Biochim. Biophys. Acta* 1482 (2000) 73–83.
- [46] C.P. Ponting, L. Aravind, J. Schultz, P. Bork, E.V. Koonin, Eukaryotic signalling domain homologues in archaea and bacteria. Ancient ancestry and horizontal gene transfer, *J. Mol. Biol.* 289 (1999) 729–745.
- [47] W.F. Doolittle, Lateral genomics, *Trends Cell Biol.* 9 (1999) 5–8.
- [48] Y. Harpaz, C. Chothia, Many of the immunoglobulin superfamily domains in cell-adhesion molecules and surface-receptors belong to a new structural set which is close to that containing variable domains, *J. Mol. Biol.* 238 (1994) 528–539.
- [49] K. Hofmann, P. Bucher, L. Falquet, A. Bairoch, The PROSITE database, its status in 1999, *Nucleic Acids Res.* 27 (1999) 215–219.
- [50] T.K. Attwood, H. Avison, M.E. Beck, M. Bewley, A.J. Bleasby, F. Brewster, P. Cooper, K. Degtyarenko, A.J. Geddes, D.R. Flower, M.P. Kelly, S. Lott, K.M. Measures, D.J. Parry-Smith, D.N. Perkins, P. Scordis, D. Scott, C. Worledge, The PRINTS database of protein fingerprints: a novel information resource for computational molecular biology, *J. Chem. Inf. Comput. Sci.* 37 (1997) 417–424.
- [51] T.K. Attwood, M.E. Beck, D.R. Flower, P. Scordis, J.N. Selley, The PRINTS protein fingerprint database in its fifth year, *Nucleic Acids Res.* 26 (1998) 304–308.
- [52] T.K. Attwood, D.R. Flower, A.P. Lewis, J.E. Mabey, S.R. Morgan, P. Scordis, J.N. Selley, W. Wright, PRINTS prepares for the new millennium, *Nucleic Acids Res.* 27 (1999) 220–225.
- [53] T.K. Attwood, M.D. Croning, D.R. Flower, A.P. Lewis, J.E. Mabey, P. Scordis, J.N. Selley, W. Wright, PRINTS-S: the database formerly known as PRINTS, *Nucleic Acids Res.* 28 (2000) 225–227.
- [54] R.B. Russell, P.D. Sasieni, M.J.E. Sternberg, Supersites within superfolds. Binding site similarity in the absence of homology, *J. Mol. Biol.* 282 (1998) 903–918.
- [55] D.R. Flower, Modelling G protein coupled receptors for drug design, *Rev. Biomembr.* 1422 (1999) 207–234.
- [56] B. Rost, Protein structures sustain evolutionary drift, *Fold. Design* 2 (1997) S19–S24.
- [57] D.R. Flower, ALTER: eclectic management of molecular structure data, *J. Mol. Graph. Mod.* 15 (1997) 161–169.
- [58] J.V. Lehtonen, K. Denessiouk, A.C.W. May, M.S. Johnson, Finding local structural similarities among families of unrelated protein structures: A generic non-linear alignment algorithm, *Proteins* 34 (1999) 341–355.