# The Lipocalin Protein Family:
## Protein Sequence, Structure and Relationship to the Calycin Superfamily

**Lola Ganfornina, Diego Sanchez, Lesley H Greene and Darren R. Flower***

## Abstract

Lipocalins are remarkable in their diversity, as manifest at the levels of protein sequence and protein function. At the level of 3-dimensional structure, however, they are very similar. The lipocalins are also part of a larger protein superfamily: the calycins, which also 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. In this review, both of these aspects are explored.

## Introduction

The lipocalin protein family is one of the most interesting and perplexing groups of homologous proteins. Diversity is the lipocalins' watchword. Lipocalins demonstrate extreme divergence at the sequence level: often beyond the ability of sequence analysis to readily recognize their relatedness. However, despite this sequence variability, the 3-dimensional structures of distinct lipocalins show remarkable similarity. Their structure comprises a slightly distended β-barrel, composed of eight β-strands, which forms an internal cavity lined by apolar residues ideally suited for the carriage of small hydrophobic residues. The barrel winds in a right-handed and conical manner around a central axis such that the first strand is hydrogen-bonded via its backbone to the last strand. The diversity of lipocalin sequences is matched by the variety of their function, their mechanisms of action, and their phyletic spread through a wide variety of species, which range from bacteria, through plants, to animals from ticks and arthropods to man. Physically, lipocalins are small (typically 150-250 residue) extracellular proteins, which share several common molecular recognition properties: the binding of small, predominantly hydrophobic molecules (such as retinol, long-chain lipids, or steroids); binding to specific cell-surface receptors;[1] and the formation of covalent and noncovalent complexes with other soluble macromolecules, such as human IgA.[2] Initially lipocalins were classified as transport proteins, their roles including carriage of retinol, odorants, and pheromones.[3] It is now clear, however, that lipocalins fulfill a wide variety of different functions: some act as tick anticoagulents, some show enzymatic activity, while others guide growing insect nerves or limit invading bacterial growth by iron sequestration.[4] The biological roles of other lipocalins are less clear-cut, with their main biological interest being coincidental to their, as yet, unknown physiological

*Corresponding Author: Darren R. Flower—The Edward Jenner Institute for Vaccine Research Compton, Newbury, Berkshire RG20 7NN, U.K. Email: darren.flower@jenner.ac.uk

*Lipocalins*, edited by Bo Åkerström, Niels Borregaard, Darren R. Flower and Jean-Philippe Salier. 

functions. For example, α-1-acid glycoprotein (AGP) is implicated in drug binding in human blood and a set of lipocalins form a major group of mammalian aeroallergens. Lipocalins have also been shown experimentally to form partially unfolded states at low pH called molten globules.[5,6] The conversion to these nonnative conformers have been proposed to be an essential step in the mechanism of ligand release for vitamin A and lipids,[5,6] opening up another dimension to their sequence-structural relationships.

Despite many common themes and functions, membership of the lipocalin family has been defined primarily on the basis of sequence, or structural, similarity and now encompasses many different proteins with a wide phyletic spread. Within this the lipocalins display unusually low levels of overall sequence conservation, with pairwise comparisons routinely falling below 20%, the nominal cutoff for reliable alignment. However, almost all lipocalins share sufficient similarity, in the form of short characteristic conserved sequence motifs, for this to effectively define family membership.[7,8]

Known 3-dimensional Lipocalin structures include a wide of variety of family members.[9,10] Recent additions include horse allergen Equ c 1,[11] Aphrodisin,[12] several anticalins,[13,14] various forms of crustacyanin,[15] Tear lipocalin,[16] and bacterial lipocalin.[17] Moreover, the structure of prostaglandin D synthase (PGDS) is imminent,[18] as well as those of several other anticalins. The common structure of the lipocalin protein fold is now well-described.[2,8,9] 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. Together with three other distinct protein families: the fatty-acid-binding proteins (FABPs), avidins, metalloproteinase inhibitors (MPIs) (and the presently enigmatic triabin), the lipocalin family forms part of a larger structural superfamily: the calycins.[2,8,19-21] 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 these pages, we shall review and update structure and sequence relations within the lipocalin protein family and also within the larger calycin protein superfamily.

## Protein Sequence Relations within the Lipocalin Protein Family

Lipocalin genes are transcribed, with few exceptions, into mRNAs of 0.6-1 kb long. These, in turn, code for proteins of 160-230 amino acids. Most of these polypeptides have a signal peptide that would export the proteins into the extracellular environment. Exceptions to this include some bacterial lipocalins (see chapter by R.E. Bishop in this volume), a dictyostelid lipocalin, and the temperature-induced lipocalins of plants. Following the signal peptide cleavage, most eukaryotic lipocalins are secreted to the extracellular milieu, while those of prokaryotes are mainly attached by lipids to the bacterial membranes. However, some prokaryotic lipocalins appear as soluble proteins in the bacterial periplasmic space, and the grasshopper Lazarillo is GPI-linked to neuronal membranes (see chapter by M.D. Ganfornina et al). Moreover, the sub-cellular location of Rat probasin appears to be nuclear. Also, the chordate apolipoprotein M (ApoM) is unique in showing no cleavage site for the N-terminal signal peptide (see chapter by W. van Dijk et al).

The mature polypeptides of lipocalins have an average predicted molecular weight (not counting postranslational modifications) of 19.4 KDa, ranging from the 17.7 KDa of ERBP to the 21.7 KDa of AGP. This variation in molecular weight results from the variable span of the N-terminal and C-terminal sides of the proteins revealed by a multiple protein sequence alignment that has been used for phylogenetic inference (see chapter by D. Sanchez et al). Plant lipocalins, most arthropodan lipocalins, the group containing both α1-microglobulin (AMG) and C8γ, and also AGP show long C-terminal extensions. However, the C-terminal region of plant lipocalins and the grasshopper Lazarillo show cleavage sites for GPI-linkage. Similarly, the mature N-termini of lipocalins from Dictyostelium, the fungus Debaromyces, the ascidian ortholog of apolipoprotein D (ApoD), and the Drosophila Karl, are unusually long. Averaged values of the molecular weight of individual clades of lipocalins are shown in Table 1.

***Table 1. Biochemical properties for lipocalin groups belonging to well established phylogenetic clades, as predicted from their protein primary sequence***

| Name | Clade | Length | MW (Da) | pI | N-Glycosyl | O-Glycosyl | S-S |
|---|---|---|---|---|---|---|---|
| Bacterial Lipocalins | I | 157 | 17915 | 8.1 | 0 | 0 | 0-1? |
| Plant Lipocalins | I | 190 | 21957 | 5.1 | 1 | 0 | 0 |
| Arthropodan Lipocalins | II | 189 | 21391 | 6.6 | 0 / 1-5 | 0-1 | 2 |
| ApoD | II | 168 | 19275 | 5.9 | 2 | 0 | 2 |
| RBP | III | 179 | 20655 | 5.9 | 0 | 0 | 3 |
| Blg | IV | 162 | 18490 | 4.9 | 0-1 | 0 | 2 |
| PGDS | V | 169 | 19084 | 7.2 | 2 | 2-4 | 1 |
| NGAL | V | 177 | 20339 | 8.2 | 1-2 | 0-3 | 1 |
| A1mg | VI | 182 | 20735 | 6.6 | 1-2 | 0-1 | 1 |
| C8GC | VI | 181 | 20193 | 8.7 | 1 | 0 | 1 |
| ERBP | VII | 159 | 17792 | 7.6 | 0-2 | 0-1 | 1 |
| A1GP | VIII | 186 | 21723 | 5.5 | 4-5 | 0 | 2 |
| RUPs | IX | 164 | 18952 | 5.1 | 0-1 | 0 | 1 |
| Chemoreception I | X | 159 | 18388 | 5.9 | 0-1 | 0 | 0-1 |
| Chemoreception II | XI | 161 | 18262 | 6.5 | 0-1 | 0-2 | 0-1 |
| ApoM | XII | 167 | 18668 | 6.1 | 0-1 | 0-2 | 3 |
| Miscellaneous Lipocalins | - | 162 | 18795 | 6.5 | 1-2 | 0-1 | 0-1 |

The alignment of lipocalins has always been a complicated and perplexing task, given the overall low conservation of their protein sequence. Pairwise sequence identities of 20-30% are common between family paralogs. This falls within the "twilight zone" for sequence assignment based on the protein primary structure. Alignments here are much less dependable than for higher levels of sequence identities. Certain pairs of lipocalins show less than 12% identity. This is generally apparent between the prokaryotic lipocalins and the odorant-binding lipocalins of chordates, while more specific examples include retinol-binding protein and the major urinary protein, aphrodisin and α-crustacyanin, the bilin-binding protein and lipocalin allergen Bos D2. This value is deep within the "midnight zone", where sequence alignments lose almost all reliability and certainty. However, the presence of several conserved sequence motifs does allow the generation of accurate automated multiple alignments of most family members. Moreover, the structure of several lipocalins that has been solved experimentally and the conserved exon-intron arrangement of lipocalins (see chapter by D. Sanchez et al) both offer additional guidance for the alignment process.

Three sequence motifs revealed in lipocalin multiple alignments, called SCRs (structurally conserved regions,[8] described in detail below), were initially proposed to be present in genuine lipocalins, and served to classify them as kernel (bearing all three SCRs) or outlier (missing one or more motifs). The discovery of new lipocalins, based on finding such motifs or a similar protein tertiary structure, has revealed that bona-fide lipocalins show a looser requirement for preserving sequence signatures. In a multiple alignment of 209 lipocalins: >90% show the SCR1 motif GxW and the SCR3 motif R/K, while >60% of the proteins show conservation in the SCR2 motif TDYxxY. Entire clades of orthologous lipocalins lack individual motifs (e.g., ApoM, AGP and odorant binding proteins lack SCR2), and individual proteins such as the marsupial late-lactation proteins, some AGPs, and some nitrophorins show significant nonconservation in SCR1.

The gaps generated in the lipocalin multiple alignment highlight important sequence characters since they represent atypical regions shared by groups of lipocalins or present only in individual proteins. Because of the secondary structure mask used to guide alignment, these gaps are all located in expected loop regions of the tertiary structure. Most gaps appear in loops placed at the open end of the β-barrel (L1, L3, L5 and L7). An expanded L1 occurs in AMG, ApoM, AGP, and some chemoreception lipocalins. The loop L5 appears elongated in retinol binding protein (RBP) and plant lipocalins, whereas L7 is extended in RBP, most arthropodan lipocalins, and a number of PGDS enzymes. The unique lipocalins Glaz of Drosophila and the BL baboon show a lengthened L3. Finally, the only elongated gap located in the closed end of the barrel is L2 in the Drosophila Glaz and Karl.

Other properties derived from the protein primary sequence are also shown in Table 1, and help us to catalog lipocalins in different family clades. The isoelectric point, calculated for the mature protein sequence, is an important factor for polypeptide solubility and folding. Most prokaryotic lipocalins, and the vertebrate neutrophil gelatinase associated lipocalins (NGALs) and C8γ show a basic p*I*, although the highest individual values (p*I*>10) are those for probasin (Clade X) and the epididymal Lcn10 (Clade VII). Lipocalin clades with predicted acidic p*I* are the Blg, AGP, RUP, RBP, Plant lipocalins, and ApoD, although in the later there is a subgroup of fish ApoD with a basic p*I*.

We have also explored the glycosylation potential for lipocalins, both for N-linked and O-linked oligosaccharides (see Table 1). The clade with the highest number of predicted glycosylation sites is a1GP. These sites are exclusively of the N-linked type. Other clades with high predicted glycosylation are PGDS (with both N-linked and O-linked residues), and the ApoD clade (only with N-linked residues). At the other extreme, RBP shows no glycosylation sites, except for some fish RBP and the functionally divergent Purpurin of chicken. The clade IV (Blg) shows a few members with one potential N-linked site, but the lipocalin Glycodelin presents three potential N-linked and one O-linked residues, highlighting its divergent function in humans (see chapter by M. Seppälä et al). Also, the arthropodan members of the clade

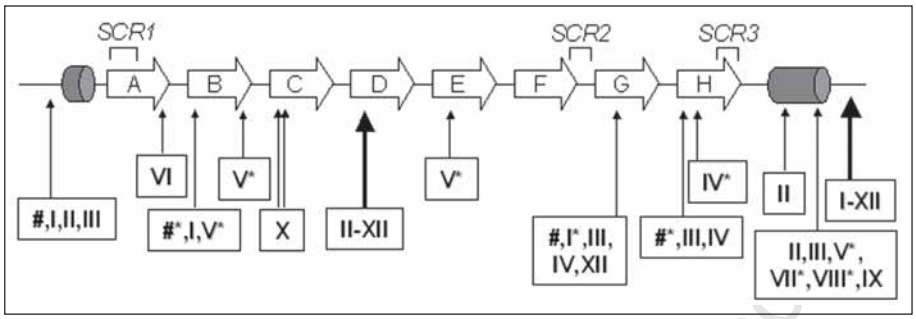


Figure 1. Cysteine locations in the lipocalin fold. Location of cysteine residues conserved in lipocalin clades in the context of the secondary structure of a prototypical lipocalin. β strands are represented by white arrows and α helices by gray cylinders. Clade numbers are defined in the chapter by D. Sanchez et al of this book. Cys residues marked with asterisks are present in subgroups of lipocalins belonging to a particular clade.

II, show a high heterogeneity in terms of glycosylation potential (see chapter by M.D. Ganfornina et al). The remaining lipocalin clades display a low number of glycosylation sites.

Finally, most lipocalins have cysteine residues involved in intramolecular disulfide bonds. Figure 1 shows a schematic representation of the secondary structure of a model lipocalin, in which the position of conserved cysteines are indicated by arrows. Cysteines conserved in lipocalin clades are depicted by clade number, while those conserved only in subgroups, within clades, are labeled by clade number with an asterisk. Table 1 shows the number of disulfide bonds present in lipocalin clades. Plant, dictyostelid and fungal lipocalins lack disulfide bonds. However, the cysteines located in β-strand B and the protein C-terminus are conserved in most chordate lipocalins and form a disulfide bond. Six cysteines found in RBP and ApoM form three disulfide bridges. Finally, clades II, IV and VIII show a conserved pattern of two disulphides, while of the rest of the lipocalins, which constitute a heterogeneous group, have a single disulfide bond.

## Structural Relationships in the Lipocalin Protein Family and Calycin Protein Superfamily

The folding pattern shared by members of the lipocalin protein family is that of a highly symmetrical all-β structure. Overall, it is dominated by a single antiparallel eight-stranded β-sheet. This is closed back on itself to form a continuously hydrogen-bonded β-barrel, which is slightly flattened in cross-section. The eight β-strands of the barrel, usually labeled A-H, are linked by a succession of + 1 connections, giving it the simplest possible β-sheet topology. The barrel winds in a conical and right-handed manner around the central axis so that strand A is hydrogen-bonded via its backbone to strand H. The seven loops, labeled L1 to L7, are all short β-hairpins, except loop L1: this is a large Ω loop. Loop L1 forms a lid folded back to close the internal ligand-binding site found at this end of the barrel. One end of the barrel (formed by loops L1, L3, and L5, and L7) is open, while the other end of the barrel is closed, with residues facing into the barrel form a tightly packed core. Beyond the eighth strand of the β-barrel is an α-helix. While this is a constant feature of all lipocalin structures, it is not conserved in its length, nor is it conserved in its position relative to the barrel. A simplified schematic is shown in Figures 2 and 3.

The β-barrel encloses a ligand-binding site composed of both an internal cavity and an external loop scaffold. It is the inherent structural and sequence diversity of this ensemble of cavity and scaffold that enables the lipocalin family to exhibit such a diversity of binding modes. Each of these modes is capable of accommodating ligands with different chemotypes, sizes, and

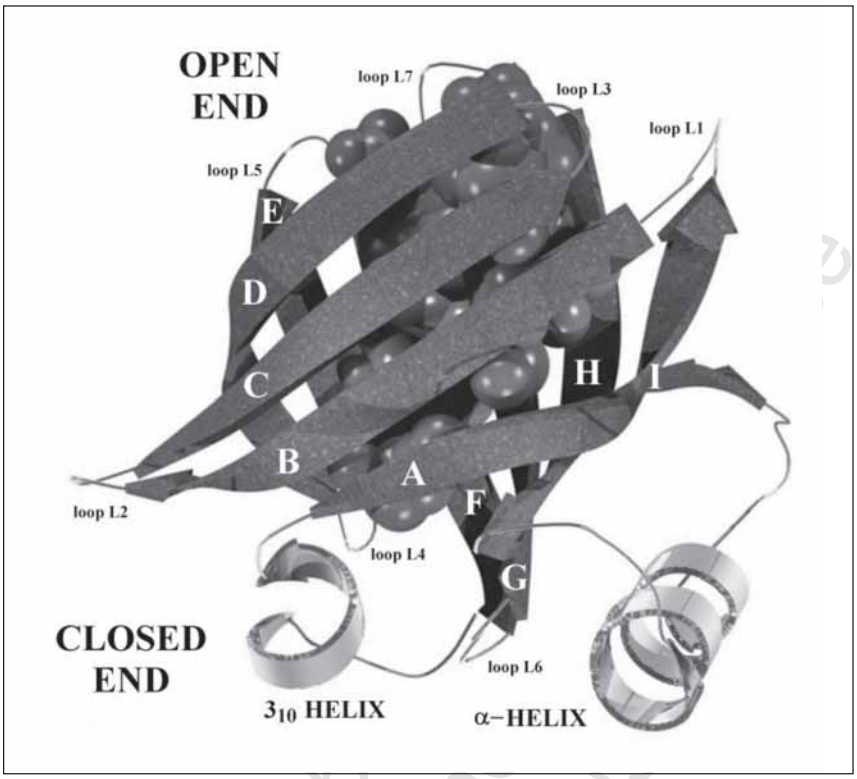


Figure 2. A schematic or ribbon-drawing of the lipocalin fold. The structure shown is a prototypical, not an actual, structure. The nine β-strands are shown labelled A-I. The N-terminal and C-terminal helices are shown and labelled. All loops (labelled L1-L7) are marked. The open end of the lipocalin β-barrel has four loops (loops L1, L3, L5, and L7). The closed 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 size of the ligand binding is shown by a collection of spheres. β-strands are depicted as curving arrows in grey, α-helices as spiral ribbons, and loops as thin cords. The figure was generated using ALTER[48] as an interface to POVRAY.

shapes. It is this diversity of structure in the ligand binding site which underlies much, but not all, of the functional diversity characteristic of the family. Contrasting with the overall highly conserved β-barrel topology, the loop region differs considerably between members of the family, both in amino acid composition, conformation, and length of the contributing polypeptide segments. This, in turn, gives rise to the particular ligand specificities displayed by individual lipocalins. Indeed, lipocalin binding sites can adopt very differing shapes. In the case of NGAL, it forms a wide, funnel-like opening to the solvent.[22] In mouse major urinary protein, the loops of the binding site close over the cavity fully encapsulating the ligand.[23] In RBP, the lumen of the binding site reaches down into the hydrophobic core of the barrel, deeply burying the α-ionone ring of retinal.[24] Finally, the binding site of human tear lipocalin (Tlc), forms an extended cavity with several lobes close to the base of the barrel.[16]

The common core characteristic of the lipocalin fold—which contains static features, like certain strands, rather than mobile features, such as the Helix—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

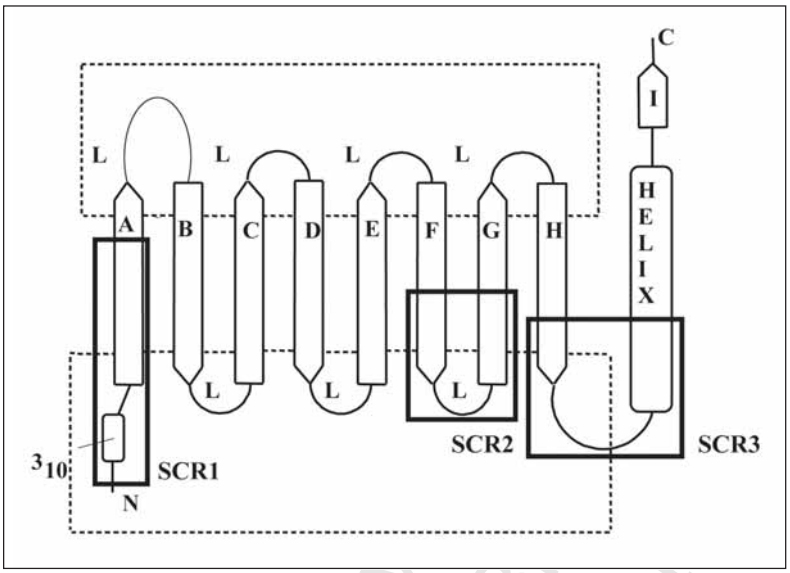


Figure 3. Schematic Structure of the lipocalin fold. An unwound view of the lipocalin fold orthogonal to the axis of the barrel. The nine β-strands of the antiparallel β-sheet are shown as arrows. The C-terminal α-helix A1 and N-terminal $3_{10}$ like helix are also marked. loops are labelled L1-L7. A pair of dotted lines indicates the hydrogen-bonded connection of two strands. Those parts that form the three main conserved regions (SCRs) of the fold (SCRl, SCR2, and SCR3) are marked as heavy boxes.

residues). Recently, two other sequence motifs have been identified and implicated in the lipocalin folding process.[25] One of these motifs is located at the closed end of the β-barrel. The other is found at the beginning of the main C-terminal α-helix. However, the new motifs cannot be found in most nonchordate lipocalins, and thus cannot help to unequivocally ascribe new members to the family outside the phylum Chordata. As mentioned above, the three principal SCRs contain a sequence motif that is wholly, or partly, invariant. Interestingly, a complete absence of the three SCRs occurs in histamine-binding lipocalins of haematophagous insects. In this case, their membership of the lipocalin family is supported by a conserved overall protein structure and conserved exon-intron pattern at the level of gene structure (see chapters by D. Sanchez et al and M.D. Ganfornina et al).

The Calycin protein superfamily, when originally identified,[8] was composed of a group of three families of ligand-binding proteins which includes the lipocalins, the fatty acid-binding proteins (FABPs), and the avidins. The avidins display an astonishing affinity for biotin, and have thus found key applications in biotechnology.[26] The FABPs are a family of predominantly intracellular proteins involved in lipid metabolism.[27] Calycins share related, if distinct, barrel structures. The structure of FABPs is built around a broken ten-stranded β-barrel structure. Avidins and related proteins, while eight-stranded, when compared to lipocalins lack a C-terminal helix or strand I and are more circular in cross-section. Others dispute the obvious and overwhelming structural similarity of the Avidins to other Calycins,[19] primarily on the basis of differences in the hydrogen bonding pattern, which manifests itself as disparities of topology and shape.[28,29] Despite such differences, and the absence of global sequence similarity, these families are characterized by a similar folding pattern—an antiparallel β-barrel with a largely +1 topology—within which significant regions can be structurally equivalenced. Moreover, beyond structural propinquity, the Calycins have a degree of functional similarity: many bind hydrophobic, or at least small, ligands and/or have key macromolecular interactions.

Over time, the size of the calycin superfamily has, on the basis of real or perceived structural similarity, grown to include a burgeoning group of other proteins: metalloprotease inhibitors,[30] triabin,[31] staphostatins,[32] domains from D-amino peptidase,[33] domains from Quinohemoprotein amine dehydrogenase,[34] the exclusion domain from Cathepsin C,[35] and various hypothetical proteins from bacteria. For some this is clearly correct and is borne out by detailed analysis. For other proteins, however, this is demonstrably not the case.[9] Photoactive Yellow Protein,[36] TCL-1,[37] MTCP-1,[37] Catalase,[38] and Cyclophilin,[39] have, for example, all been suggested as potential calycins. As, indeed, has the pleckstrin homology domain.[40] This haüãalso been suggested as a structural homologue of other many proteins, including verotoxin and FK-506.

Superficially, calycins do resemble certain other all-β proteins with barrel-like structures,[41] and it is correct to say that β-barrels, in particular, are easily confused. This is contingent upon coincidental similarities that may arise as a product of the principles underlying barrel structure: only certain barrel geometries and shear numbers are possible. However, the relative resemblance at the level of gross structure is often insufficient to verify or refute convergent or divergent evolution, obliging us to look for synergistic indications, such as a greater or lesser degree of similarity in topology, size, and binding site composition, from both structure and sequence, to further characterise such similarities.

Beyond the similarities described above, members of the calycin superfamily have in common a characteristic structural pattern:[20] a lysine or arginine (from the terminal strand of the β-barrel) forms several potential hydrogen bonds to carbonyl groups of the main-chain of the short N-terminal $3_{10}$-like helix while packing across a conserved tryptophan (from the initial strand of the barrel) in a structurally superimposible, nonrandom manner. Visual inspection of available lipocalin, Avidin, and FABP structures all reveal a very similar arrangement of interacting residues. This signature corresponds to sequence determinants common to the calycin member families: a characteristic N-terminal sequence pattern centred on Tryptophan, which displays preservation of key residues, and a weaker C-terminal motif centred on arginine or lysine. Although some conservation is apparent within these patterns, it is not of sufficient strength to allow the design of sequence discriminators able to identify all calycins with certainty.

For the newly identified Calycins, several appear, on the basis of retained structural signatures, to be bone fide members of the superfamily. These include the MPIs and the relevant domains from D-amino peptidase, Cathepsin C, and Quinohemoprotein amine dehydrogenase. For other proteins, such as YodA[42] and the *Thermus thermophilus* HB8 polyprenyl pyrophosphate binding protein,[43] both globally and in terms of retained structural signatures, apparent structural propinquity may be simply coincidental. Many members of the Calycin superfamily have variant signatures, which maintain some of the overall specificity of interaction but use different amino acids. Certain lipocalins—the late lactation proteins, for example—have a tyrosine instead of 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. Other Calycins substitute lysine for arginine. Examples of this come from a group of highly diverged FABPs, typified by insect muscle FABPs, and also Quinohemoprotein amine dehydrogenase.

Triabin and staphostatin remain enigmatic, in that they have similarity in terms of global conformation and the conservation of the family signature, yet have perturbed topologies. In triabin, Strands B and C are interchanged in position, altering the repeated +1 topology and antiparallel arrangement of adjacent strands. By allowing reverse matching of structural segments rather than exclusively ordered matching, large proportions of the triabin and other calycin structures can be equivalenced. At the level of protein sequence, Triabin has some global similarity to nitrophorin and *Rhodnius prolixus* salivary platelet aggregation inhibitors. Staphostatin has a structural signature similar to other Calycins but it is displaced with the sequence. The first three strands of the staphostatin barrel form a β-meander rather than two successive β-hairpins. As a consequence, the first strand interacts with the third and fourth

strands (sequential numbering) so that the signature tryptophan, rather than coming from strand 1, instead comes from strand two (sequential numbering), which itself interacts with the third and eighth strands (sequential numbering).

It remains unclear whether these changes—variations in topology and conservation—generate alternative compensating interactions within these structures or are examples of drift in protein evolution. If certain proteins are tolerant to such alterations, then why, against a backdrop of significant sequence divergence across the whole family, are they required by most? In current structures, sequences giving rise to particular folding patterns may have evolved stability independent of this interaction of residues. Ignoring disordered proteins,[44] protein sequences in aqueous solvents generally fold into, essentially, a unique structure. However, distinct sequences can fold into similar structures. As Rost reports,[45] only 3-4% of amino acids appear to be crucial for protein structure and function. Residue identities for proteins which have evolved from the same (divergent) or different (convergent) ancestors are similar and it is problematic to differentiate them. Low sequence identity does not necessarily indicate a convergent route. What we see in the Calycins is, perhaps, a distant evolutionary relic of the common calycin ancestor protein: still an important structural interaction but no longer essential. Nonetheless, conservation of a characteristic sequence signatures corresponding to an even more highly conserved structural signature supports the view that there is a common, if now very remote, origin from which the members of the calycin superfamily (lipocalins, FABPs, avidins, MPIs, triabin, etc.) have diverged.

## Folding and Stability

The lipocalin protein family has continued to grow, both in terms of sequences and structures determined, but also in their diversity and interest. Members of the calycin protein superfamily share a β-barrel structure and certain member families share the ability to bind hydrophobic ligands, although many do not. 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. Certainly, these conserved interactions act to "pin" together the two ends of the calycin β-barrel, but what other role does this structural signature play? Might its role be functional or might it be a protein-protein recognition site perhaps? Does it stabilize the structure maintaining the overall fold? Or is it involved in the folding pathway, perhaps guide the formation of the β-barrel?

A recent paper,[46] throws considerable light on this. Using RBP as their exemplar, the authors generated a series of mutant proteins. Changes to TRP24 or ARG139, both involved in the Calycin signature, lead to similar significant losses in stability and decreased yields of protein as generated by folding in vitro. As a control they also mutated several other, more accessible trytophans and found that they did not affect stability or expression. These results, in concert with the nature of natural amino acid mutations at these positions, support the notion that conserved residues in homologous proteins act to increase the proportion of folded to misfolded proteins, thus stabilizing the native structure.

In a separate, more recent paper, these authors show that the main sequence motifs in the lipocalins, which form a superset that includes the main calycin motifs, are involved critically in the folding process of RBP.[25] This result synergises with results from Goto and coworkers, who have analysed the β-lactoglobulin (Blg) folding process.[47] In common with many other proteins (α-lactalbumin, lysozyme, plasminogen activator inhibitor type-1, annexin 1, etc.), studies on Blg indicate that there is a transient intermediate with significantly more α-helical content and less β-sheet than the native protein. They used ultra-rapid mixing techniques to monitor folding over a 100 μs to 10 s timescale. The folding intermediate detected in their experiments contains a well-ordered region formed from strands F, G and H, as well as the *C*-terminal helix and a region of the *N*-terminus (see Fig. 1). This last region normally adopts

a β-strand conformation preceded by a rare, but conserved, $3_{10}$ helix. It is this region that adopts a nonnative α-helical conformation different to that in the fully folded protein. These results also help to explain the long-standing observation that polar solvents, such as water–alcohol mixtures, tend to increase the α-helical structure apparent in Blg, as observed by circular dichroism spectroscopy or NMR. Their observation fits well with an emerging consensus on the fundamental mechanisms underlying protein folding: a multi-dimensional energy landscape, sometimes described as a folding funnel, allows a large number of unequally populated alternative routes from the unfolded protein to the native state. For all but the simplest proteins, some of these routes will involve intermediates, or local minima, and may lead to kinetically trapped misfolded proteins.

Resolving the protein folding problem is one of the greatest challenges in science today. The lipocalins offer a promising system to investigate the determinants of topology, stability and the relationship between sequence conservation and folding. The great variation in their critical biological functions, their potential for specifically engineered drug transport, interactions with receptors, and their overwhelming pervasiveness throughout the eukaryotic kingdom have, biologically speaking, placed the lipocalins centre stage, with a thrilling new era of untold discoveries waiting to unfold.

## References


1. Flower DR. Beyond the superfamily: the lipocalin receptors. Biochim Biophys Acta 2000; 1482:327-336.
2. Flower DR. The lipocalin protein family: Structure and function. Biochem J 1996; 318:1-14.
3. Pervaiz S, Brew K. Homology of beta-lactoglobulin, serum retinol-binding protein and protein HC. Science 1985; 228:335-337.
4. Flo TH, Smith KD, Sato S et al. Lipocalin 2 mediates an innate immune response to bacterial infection by sequestering iron. Nature 2004; 432:917-921.
5. Bychkova VE, Dujsekina AE, Fantuzzi A et al. Fold Des 1998; 3:285-291.
6. Gasymov OK, Abduragimov AR, Gasimov EO et al. Tear lipocalin: Potential for selective delivery of rifampin. Biochim Biophys Acta 2004; 1688:102-111.
7. Flower DR, North ACT, Attwood TK. Mouse oncogene protein-24p3 is a member of the lipocalin protein family. Biochem Biophys Res Commun 1991; 180:69-74.
8. Flower DR, North ACT, Attwood TK. Structure and sequence relationships in the lipocalins and related proteins. Protein Sci 1993; 2:753-761.
9. Flower DR, North AC, Sansom CE. The lipocalin protein family: Structural and sequence overview. Biochim Biophys Acta 2000; 1482:9-24.
10. Flower DR. Experimentally determined lipocalin structures. Biochim Biophys Acta 2000; 1482:46-56.
11. Lascombe MB, Gregoire C, Poncet P et al. Crystal structure of the allergen Equ c 1. A dimeric lipocalin with restricted IgE-reactive epitopes. J Biol Chem 2000; 275:21572-21577.
12. Vincent F, Lobel D, Brown K et al. Crystal structure of aphrodisin, a sex pheromone from female hamster. J Mol Biol 2001; 305:459-469.
13. Korndorfer IP, Beste G, Skerra A. Crystallographic analysis of an "anticalin" with tailored specificity for fluorescein reveals high structural plasticity of the lipocalin loop region. Proteins 2003; 53:121-129.
14. Korndorfer IP, Schlehuber S, Skerra A. Structural mechanism of specific ligand recognition by a lipocalin tailored for the complexation of digoxigenin. J Mol Biol 2003; 330:385-396.
15. Habash J, Helliwell JR, Raftery J et al. The structure and refinement of apocrustacyanin C2 to 1.3 A resolution and the search for differences between this protein and the homologous apoproteins A1 and C1. Acta Crystallogr D Biol Crystallogr 2004; 60:493-498.
16. Breustedt DA, Korndorfer IP, Redl B et al. The 1.8-A crystal structure of human tear lipocalin reveals an extended branched cavity with capacity for multiple ligands. J Biol Chem 2005; 280:484-493.
17. Campanacci V, Nurizzo D, Spinelli S et al. The crystal structure of the Escherichia coli lipocalin Blc suggests a possible role in phospholipid binding. FEBS Lett 2004; 562:183-138.
18. Irikura D, Kumasaka T, Yamamoto M et al. Cloning, expression, crystallization, and preliminary X-ray analysis of recombinant mouse lipocalin-type prostaglandin D synthase, a somnogen-producing enzyme. J Biochem (Tokyo) 2003; 133:29-32.
19. Flower DR. Structural relationship of streptavidin to the calycin protein superfamily. FEBS Letters 1993; 333:99-102.

20. Flower DR. A structural signature characteristic of the calycin protein superfamily. Protein Pept Lett 1995; 2:341-346.
21. Flower DR. The up-and-down beta-barrel proteins: Three of a kind. FASEB J 1995; 9:566-567.
22. Goetz DH, Willie ST, Armen RS et al. Ligand preference inferred from the structure of neutrophil gelatinase associated lipocalin. Biochemistry 2000; 39:1935-1941.
23. Bocskei Z, Groom CR, Flower DR et al. Pheromone binding to 2 rodent urinary proteins revealed by X-Ray crystallography. Nature 1992; 360:186-188.
24. Cowan SW, Newcomer ME, Jones TA. Crystallographic refinement of human serum retinol binding- Protein at 2 angstroms resolution. Proteins 1990; 8:44-61.
25. Greene LH, Hamada D, Eyles SJ et al. Conserved signature proposed for folding in the lipocalin superfamily. FEBS Lett 2003; 553:39-44.
26. Green NM. Avidin and streptavidin. Meth Enzymol 1990; 184:51-67.
27. Banaszak L, Winter N, Xu ZH et al. Lipid-Binding Proteins - A family of fatty-acid and retinoid transport proteins. Adv Protein Chem 1994; 45:89-151.
28. Murzin AG, Lesk AM, Chothia C. Principles determining the structure of beta-sheet barrels in proteins. I. A theoretical analysis. J Mol Biol 1994; 236:1369-1381.
29. Murzin AG, Lesk AM, Chothia C. Principles determining the structure of beta-sheet barrels in proteins. II. The observed structures. J Mol Biol 1994; 236:1382-1400.
30. Baumann U, Bauer M, Letoffe S et al. Crystal-Structure of a complex between Serratia-marcescens metalloprotease and an inhibitor from Erwinia-chrysanthemi. J Mol Biol 1995; 248:653-661.
31. Fuentesprior P, Noeskejungblut C, Donner P et al. Structure of the thrombin complex with triabin, a Lipocalin-like Exosite-binding inhibitor derived from a triatomine bug. Proc Natl Acad Sci USA 1997; 94:11845-11850.
32. Rzychon M, Filipek R, Sabat A et al. Staphostatins resemble lipocalins, not cystatins in fold. Protein Sci 2003; 12:2252-2256.
33. Bompard-Gilles C, Remaut H, Villeret V et al. Crystal structure of a D-aminopeptidase from Ochrobactrum anthropi, a new member of the 'penicillin-recognizing enzyme' family. Structure Fold Des 2000; 8:971-980.
34. Satoh A, Kim JK, Miyahara I et al. Crystal structure of quinohemoprotein amine dehydrogenase from Pseudomonas putida. Identification of a novel quinone cofactor encaged by multiple thioether cross-bridges. J Biol Chem 2002; 277:2830-2834.
35. Turk D, Janjic V, Stern I et al. Structure of human dipeptidyl peptidase I (cathepsin C): Exclusion domain added to an endopeptidase framework creates the machine for activation of granular serine proteases. EMBO J 2001; 20:6570-6582.
36. Borgstahl GEO, Williams DR, Getzoff ED. 1.4 Angstrom structure of photoactive yellow protein, a cytosolic Photoreceptor - Unusual fold, Active-Site, and chromophore. Biochemistry 1995; 34:6278-6287.
37. Fu ZQ, Dubois GC, Song SP et al. Crystal structure of MTCP-1: Implications for role of TCL-1 and MTCP-1 in T cell malignancies. PNAS 1998; 95:3413-3418.
38. Russell RB, Sternberg MJE. A novel binding site in catalase is suggested by structural similarity to the calycin superfamily. Protein Eng 1997; 9:107-111.
39. Kallen J, Spitzfaden C, Zurini MGM et al. Structure of human cyclophilin and its Binding-Site for Cyclosporine-A determined by X-Ray crystallography and Nmr- Spectroscopy. Nature 1991; 353:276-279.
40. Orengo CA, Swindells MB, Michie AD et al. Structural similarity between the pleckstrin homology domain and Verotoxin - the problem of measuring and evaluating structural similarity. Protein Sci 1995; 4:1977-1983.
41. Efimov AV. A structural tree for proteins containing 3 Beta-Corners. FEBS Letters 1997; 407:37-41.
42. David G, Blondeau K, Schiltz M et al. YodA from Escherichia coli is a metal-binding, lipocalin-like protein. J Biol Chem 2003; 278:43728-43735.
43. Handa N, Terada T, Doi-Katayama Y et al. Crystal structure of a novel polyisoprenoid-binding protein from Thermus thermophilus HB8. Protein Sci 2005; 14:1004-1010.
44. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 2005; 6:197-208.
45. Rost B. Protein structures sustain evolutionary drift. Fold Des 1997; 2:S19-S24.
46. Greene LH, Chrysina ED, Irons LI et al. Role of conserved residues in structure and stability: Tryptophans of human serum retinol-binding protein, a model for the lipocalin superfamily. Protein Sci 2001; 10:2301-2316.
47. Kuwata K, Shastry R, Cheng H et al. Structural and kinetic characterization of early folding events in β-lactoglobulin. Nature Struct Biol 2001; 8:151-155.
48. Flower DR. ALTER: Eclectic management of molecular structure data. J Mol Graph Mod 1997; 15:161-169.