Phylogeny of whey acidic protein (WAP) four-disulfide core proteins and their role in lower vertebrates and invertebrates

Valerie J. Smith1
Scottish Oceans Institute, Gatty Laboratory, University of St Andrews, St Andrews, Scotland KY16 8LB, U.K.

Abstract
Proteins containing WAP (whey acidic protein) domains with a characteristic WFDC (WAP four-disulfide core) occur not only in mammals (including marsupials and monotremes) but also in birds, reptiles, amphibians and fish. In addition, they are present in numerous invertebrates, from cnidarians to urochordates. Many of those from non-mammalian groups are poorly understood with respect to function or phylogeny. Those well characterized so far are waprins from snakes, perlwapins from bivalves and crustins from decapod crustaceans. Waprins are venom proteins with a single WAP domain at the C-terminus. They display antimicrobial, rather than proteinase inhibitory, activities. Perlwapins, in contrast, possess three WAP domains at the C-terminus and are expressed in the shell nacre of abalones. They participate in shell formation by inhibiting the growth of calcium crystals in the shell. The crustin group is the largest of all WFDC-containing proteins in invertebrates with the vast majority being highly expressed in the haemocytes. Most have a single WAP domain at the C-terminus. The presence and type of the domains between the signal sequence and the C-terminus WAP domain separate the different crustin types. Most of the Type I and II crustins are antimicrobial towards Gram-positive bacteria, whereas the Type III crustins tend to display protease inhibition. Expression studies show that at least some crustins have other important biological effects, as levels change with physiological stress, wound repair, tissue regeneration or ecdysis. Thus WAP domains are widely distributed and highly conserved, serving in diverse physiological processes (proteinase inhibition, bacterial killing or inhibition of calcium transport).

Introduction
WFDC [WAP (whey acidic protein) four-disulfide core]-containing proteins have been studied primarily in placental mammals but are not confined to these higher vertebrates. Sequences with a WAP domain motif have also been reported for other vertebrates: for example, marsupials, monotremes, birds, reptiles, amphibians and teleost fish (Figure 1a). In addition, WFDC proteins are present in a diversity of invertebrates (arthropods, cephalochordates, ascidians, echinoderms, molluscs, insects, crustaceans, nematodes and cnidarians; Figure 1b), plus some cereal plants and bacteria. There are some 1765 hits for ‘WAP domain’ in animals in the GenBank® Nucleotide Sequence Database at the time of writing. Of these, 85% are from vertebrates, mainly placental mammals, with a mere 15% from invertebrates, despite the fact that invertebrates constitute approximately 95% of all animal species. Figure 1 shows the breakdown of the percentage of sequences according to taxon. Among the mammals, most are for rodents and humans, whereas among the invertebrates, most are for insects and crustaceans (Figure 1); no doubt this reflects the intensity of sequencing efforts on species of medical or commercial importance. For many species, however, there have been no subsequent analyses of the natural or recombinant proteins. The few WFDC-containing proteins from lower vertebrates and invertebrates for which some functional or proteomic information is available are from reptiles, molluscs and crustaceans. The various WAP proteins from these taxa are described and compared in the present mini-review, considering in particular their structure and biological roles in order to identify any trends or themes.

Waprins
Waprins are a family of 66–84 amino acid proteins present in the venom of snakes [1], in particular the elapids and colubrids [2]. In most species they comprise a signal sequence at the N-terminus and a single WAP domain at the C-terminus. Although some snake venoms have been known for many years to contain four-disulfide core proteins [3], the first to be designated ‘waprin’ was from the spitting cobra, Naja nigricollis [1]. Subsequently, many waprins have been identified, with some 127 sequences reported in the GenBank® Nucleotide Sequence Database from 15 species. They tend to be named using key letters from the species giving, for example, nawaprin from N. nigricollis and omwaprin from Oxyuranus microlepidotus [1]. Among
Figure 1 | Relative proportions of sequences listed in the GenBank® Nucleotide Sequence Database for different animal taxa

Data are based on 1510 nucleotide or protein sequences derived from vertebrates and 255 from invertebrates.

In addition, a WAP-domain-containing protein, with an unusual arrangement of domains, occurs in venom of the rattlesnakes *Sistrurus catenatus edwardsii* and *Sistrurus catenatus ergeminus* from the family Viperidae. In these venoms, the waprin protein is replaced by a Ku-wap-fusin protein, in which the WAP domain is fused to a Kunitz-like domain [5,6]. This Kunitz-like proteinase inhibitor domain is similar to the one present in elapid toxins, namely, dendrotoxins of *Dendroaspis* spp., textilins of *Pseudonaja textilis* and the bungarotoxins of kraits belonging to the genus *Bungarus* [7]. None of these three toxin groups have a WAP domain adjacent to the Kunitz domain (Figure 2). Thus the Ku-wap-fusin venom proteins represent an intermediate type between the waprin toxins and the non-WAP Kunitz neurotoxins [5,6]. The three classes of venom proteins have sufficient conservation in signal sequences, other exons and some non-coding regions [4] to permit analyses of phylogenetic relationships. Construction and analysis of such a Bayesian phylogenetic tree have led to the proposal of a theoretical model for the likely evolution of these proteins [5]. This posits that both Kunitz-containing and WAP-type venom proteins arose from a common ancestral gene encoding a toxin with a Kunitz-like proteinase inhibitor domain. This is postulated to have given rise, through deletion and insertion of new domains, to two gene lineages: one encodes toxins with Kunitz-like serine protease inhibitors and the other a lineage that led to the Ku-wap-fusins through the acquisition of a WAP domain [5]. The gene encoding the Kunitz-inhibitor proteins in kraits is proposed to have then undergone further insertions to become the bugarotoxins [5]. Conversely, the deletion of the Kunitz domain from the Ku-wap-fusins is thought to have created the waprins with a single WAP domain, which then, by duplication, may have given rise to the DWD (double WAP domain) of the *P. olfersii* venom [5]. Importantly, these lines of evolution occurred separately from the evolution of WAP proteins in mammals and so are not ancestral to the WAP or Kunitz proteins of higher vertebrates [5].

Waprins are believed to have antimicrobial activities, largely inferred from the WAP signature (Table 1). Omwaprin has been shown by in vitro study to act only against some Gram-positive bacteria, killing target micro-organisms by disruption of the membrane [8]. It has neither proteinase inhibitory nor haemolytic activities and, unlike many antimicrobial peptides, is stable in high (∼2%, w/v) salt [8].

### Chelonianin

Chelonianin is the name given to a 110 amino acid cysteine-rich basic protein isolated from the egg white of the red sea turtle (probably loggerhead turtle, *Caretta caretta*) [9]. This protein is an extracellular serine-type trypsin/subtilisin endopeptidase inhibitor (UniProt P00993) with a single WFDC domain of 46 amino acids at the C-terminus and a 13 amino acid signal sequence at the N-terminus [10]. These flank a central 51 amino acid Kunitz-type domain, similar to the Ku-wap-fusin proteins in the venom of vipers and colubrid snakes (Figure 2). Despite its early discovery, little is known about the expression patterns, biological roles or occurrence of chelonianin in either *C. caretta* or other cheloniens, but the name has been used to describe, perhaps inappropriately, a number of non-mammalian WAP-containing proteins in fish, shrimp and whipworm. However, the fish sequence (GenBank® accession number BAC00855) has three WAP domains and is similar to the perlwapins (described below); the shrimp ‘chelonianin’ [11] lacks a typical 6-cysteine Kunitz domain; and the whipworm protein (GenBank® accession number EFV61339) is lacking a complete WFDC domain.

### Perlwapins

Perlwapins are 134 amino acid proteins secreted in the shell matrix of the abalones [12,13]. They were first identified in the green lip abalone *Haliotis laevigata* [12], with a similar protein...
present in the shell nacre of the donkey ear abalone *Haliotis arsinima* [13]. Both perlwapins contain three tandem WAP domains of 42–44 amino acids, each of which have 40–53% identity in amino acids [12] (Figure 2). The third WAP domain in *H. laevigata* is unusual in possessing an extra (ninth) cysteine residue at the C-terminus [12]. In abalones, the perlwapins appear to play a major role in shell formation in *vivo* (Table 1). They bind mineral (calcite) crystals in the shell matrix and are thought to regulate the growth of the calcite layers within the nacre by inhibiting calcium deposition [12].

A perlwpain-like sequence (GenBank® accession number P86855) also occurs in the shell secretome of the Mediterranean mussel (a bivalve) *Mytilus galloprovincialis*, although this appears to have only one complete WAP domain. Its role in calcium deposition and shell growth is unknown. Similarly, triple WAP-domain-containing sequences, resembling perlwapins, have, curiously, been found in a number of other taxa, including fish (GenBank® accession numbers ACI67873 and XP_684531.2), but as these have been found in EST (expressed sequence tag) libraries or from genome analyses, virtually nothing is known about their function. Intriguingly, the WAP-domain-containing sequence (GenBank® accession number ACI67873) from Atlantic salmon (*Salmo salar*) occurs in the head kidney [14], the organ that produces the macrophages; a finding indicating that it might participate in inflammation.

Lustrin A is a large multidomain extracellular nacre protein of 1428 amino acids from *Haliotis rufescens* [15]. It contains one WAP domain at the C-terminus that shares high (~61%) similarity to the third WAP domain in *H. laevigata*, and has not only the extra cysteine residue at the C-terminus but also a tenth cysteine residue between Cys3 and Cys4 in the WFDC [12,15]. Lustrin A is a complex protein that, as well as the WAP domain, has a long (954 amino acids) region of nine cysteine-rich motifs interspersed with eight proline-rich motifs, followed by two glycine-serine loops, another cysteine-rich motif and then a short 31 amino acid basic region [15] (Figure 2). The glycine region is thought to confer some elasticity and flexibility on the protein, and the WAP domain to be primarily involved in proteinase inhibition, protecting the shell matrix from degradation [15]. However, this does not rule out other biological effects and lustrin A could well be multifunctional (Table 1) on account of its complex domain composition (Figure 2).

**Crustins**

Invertebrate WAP-domain-containing proteins are particularly well described in the crustaceans. The first report for this taxon (and arguably the first WAP domain proteins discovered for invertebrates) was an 11.5 kDa cationic protein, later termed ‘carcinin’, isolated from the circulating haemocytes of the crab, *Carcinus maenas*, on the basis of antibacterial activity [16]. Many others (now over 64 proteins or genes and isoforms) from 37 decapod species have subsequently been found, primarily from shrimp, lobster, crayfish and other brachyuran crabs, and they are now referred to as ‘crustins’ rather than as ‘carcinins’ [17]. Intriguingly EST sequences with typical WAP domains have also been found in non-decapod crustaceans, including the copepod, *Calanus finmarchicus* (GenBank® accession number EL273178), the brine shrimp *Artemia franciscana*.
Table 1 | Functional characteristics of WFDC proteins from reptiles, crustaceans and molluscs

<table>
<thead>
<tr>
<th>Taxon</th>
<th>WAP protein type</th>
<th>Number of species*</th>
<th>Antibacterial activity†</th>
<th>Proteinase inhibition</th>
<th>Other functions</th>
<th>Key reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reptiles</td>
<td>Waprins</td>
<td>15</td>
<td>+ (Omwaprin)</td>
<td></td>
<td>Envenomation</td>
<td>[8]</td>
</tr>
<tr>
<td></td>
<td>Ku-wap-fusins</td>
<td>2</td>
<td>?</td>
<td>+/? (Inferred)</td>
<td>Envenomation</td>
<td>[5,6]</td>
</tr>
<tr>
<td></td>
<td>Chelonanin</td>
<td>1</td>
<td>?</td>
<td>+ (Assumed)</td>
<td>?</td>
<td>[9]</td>
</tr>
<tr>
<td>Crustaceans</td>
<td>Crustins Type I</td>
<td>15</td>
<td>+ (Mainly Gram-positive bacteria)</td>
<td>- (Crayfish only)</td>
<td>?</td>
<td>[17,25-27]</td>
</tr>
<tr>
<td></td>
<td>Crustins Type II</td>
<td>15</td>
<td>+ (Mainly Gram-positive bacteria)</td>
<td></td>
<td>?</td>
<td>[17,18,20]</td>
</tr>
<tr>
<td></td>
<td>Crustins DWD (Type IV?)</td>
<td>2</td>
<td>-/?</td>
<td>+ (Some spp.)</td>
<td>+ (Subtilisin)</td>
<td>bind to bacteria</td>
</tr>
<tr>
<td>Molluscs</td>
<td>Perlwapins</td>
<td>4</td>
<td>?</td>
<td></td>
<td></td>
<td>[12]</td>
</tr>
<tr>
<td></td>
<td>Lustrin A</td>
<td>2</td>
<td>?</td>
<td></td>
<td>Unknown role in shell nacre, Multifunctional (?)</td>
<td>[15]</td>
</tr>
</tbody>
</table>

*The number of species for which full protein amino acid or full coding cDNA nucleotide sequence data are available for the taxon specified.
†Derived from studies with natural (purified) or recombinant proteins in vitro.

(∗GenBank® accession number Q1W1H4), and the amphipod *Gammarus pulex* (GenBank® accession numbers EH273178 and EH268909), but so far the full proteins have not yet been characterized and no further information is available.

All decapod crustins have a 45–50 amino acid WAP domain at the C-terminus and a 16–24 amino acid signal sequence at the N-terminus, with usually one or two additional domains between the signal sequence and the WFDC. In some species (mainly crabs, lobsters and crayfish), the extra domain has 48–53 residues, including six cysteines (Figure 2). These crustins are classed as Type I [17]. Type II crustins have two flanked domains: one is cysteine-rich and similar to that of the Type I crustins and the other is a 50–60 residue glycine-rich domain with a high representation of repeat VGGGLF motifs (Figure 2) [17]. They are present in shrimp but are not confined to these decapods as similar proteins occur also in crayfish *Pascifastacus leniusculus* [18] and spiny lobster *Panulirus japonicus* [19]. A third group of crustins are the Type IIIs. These are known only in a few species of shrimp, crab and spiny lobster and all comprise only a single WAP domain with a signal sequence and a short region containing proline and arginine residues (Figure 2) [[17,20-22] and an unpublished sequence (GenBank® accession number GU137452) from porcelain crab]. These were once described as chelonianin-type proteins [11], but this is misleading as they lack the Kunitz domain found in the turtle protein (above). Recently, two studies have described crustin-like proteins in shrimp that have two WAP domains, a signal sequence and no other motifs [22,23] (Figure 2). These are named DWD proteins, but can be tentatively considered to be Type IV crustins, at least until additional information is forthcoming. The different crustin types are not mutually exclusive, as more than one crustin type may be co-expressed in a single species [18,19,26].

Regarding bioactivity, many of the Type I and II crustins have been shown to have antimicrobial activities, primarily against Gram-positive bacteria ([16,17,25,26]; also reviewed in [17]), with relatively few found to kill Gram-negatives [20] (Table 1). The Type III crustins, however, may exhibit either antibacterial activity [21,22,27] or proteinease inhibition, usually towards subtilisin [11] (Table 1). Of the two DWD (putative Type IV) crustins, both appear to have proteinase inhibitory activity [23,24] and the recombinant DWD from *Fenneropenaeus chinensis* binds bacteria [24] although its ability to inhibit the growth of bacteria is uncertain (Table 1). It is also unclear the extent to which Type I and Type II crustins have proteinase inhibitory properties, as this has rarely been investigated. Notwithstanding, one study [26] has reported that two Type I crustins from crayfish do not inhibit proteinases.

The role of crustins in host defence is usually assumed not only because of their effects on bacterial growth (although little is known about the killing mechanism) but also because in the vast majority of species, crustins are expressed by the circulating haemocytes, in particular the granular cells [17]. Expression levels may be very high in these cells. For example, in the spider crab *Hyas araneus*, expression of crustins can be up to 200 times as high as that of other antimicrobial proteins.
produced within the same cells [28]. In addition, some authors have also detected mRNA transcripts in other tissues ([29–31], also reviewed in [17]), but as crustaceans have an open-type circulation and all tissues are bathed in haemolymph, it is not clear whether the transcripts are derived mainly from the haemocytes. What is more interesting is the finding of high expression of a Type I crustin in developing limb buds of the fiddler crab *Cceluca pagulator* (GenBank® accession number DW176897), and another (GenBank® accession number AY340636) in regenerating olfactory bulb of the spiny lobster *Panulirus argus* with experimental ablation [32]. While one might reasonably expect haemocytes carrying antimicrobial proteins to accumulate at wound sites or in exposed tissues unprotected by a chitin exoskeleton to prevent colonization by bacteria from sea water, elevated expression of crustins does not seem to correlate with the moult cycle. Rather, in the swimming crab *Portunus pelagicus*, crustin expression is approximately seven to eight times lower during ecdysis than that at intermoult when the crab is fully armoured [33,34]. This is not due to a change in the size of the population of the crustin-bearing granular haemocytes, as this is higher at ecdysis than at intermoult [35]. It is possible that at least some crustins may contribute to the regulation of protease cascades such as those involved in clotting and phenoxidoxidase activation [36] or in hardening of the carapace post-moult, perhaps through calcium regulation (see above).

Certainly expression levels of crustins may also vary with environmental experiences such as temperature [27,37], changed water quality [31] or exposure to non-self agents ([29,37–39]; also reviewed in [17]). However, expression patterns are either inconsistent or contrary to what might be expected. Non-self challenge experiments, in particular, yield conflicting results although the only *in vitro* study published so far has shown that there is little change in expression after exposure to bacteria by isolated granular cells from the spider crab *H. araneus* [28]. Thus differences between the findings of authors may reflect experimental design, where fluxes in haemocyte number are not taken into account, or that the response is more strongly connected with stress or trauma. Certainly some studies indicate that the latter is a factor [37,40].

**Conclusions**

This brief survey of WFDC-containing proteins in lower vertebrates and invertebrates shows that the WAP domain occurs in many separate phyla and has therefore been conserved. Within individual taxa, protein families with this domain have diversified and continued to evolve but there is no clear phylogeny overall across groups. Clearly, the domain has important properties, the best known of which are antisepsis and proteinase inhibition (Table 1), both of which are shown by mammalian WDFC proteins. It should also be borne in mind that demonstration of antibacterial activity *in vitro* does not prove that the protein functions primarily as an anti-infective *in vivo*, especially if the *in vitro* assay conditions are not physiologically relevant. Some of the WFDC proteins in invertebrates appear to also play roles in calcium regulation, transport or deposition, which is not surprising given the presence of WFDC proteins in milk of monotremes and marsupials plus some placental mammals. Unfortunately, the extent to which the invertebrate proteins assist in other processes is unclear as functionality is often assessed on the basis of extrapolation from mammalian studies or based on *in vitro* analyses that may not mimic the *in vivo* state.

Certainly, the tightly coiled conformation of WAP structures must provide stability and strength to their proteins, perhaps accounting for their presence in the shell nacre of mollusces, the extracellular collagen matrix of the cnidarian *Hydra vulgaris* [41], and in regenerating tissues in decapod crustaceans. Certainly the tightness of coiling must govern functionality, as it will determine how the protein can interact with its target; hence, relating the subtle differences in the shape of the coil to functionality might be an interesting way forward with these proteins. Invertebrates are useful, and ethically acceptable, models for such studies. More importantly, studies on invertebrate and lower vertebrate taxa might offer insights into hitherto unknown roles that WFDC proteins might play in higher mammals.

**References**


©The Authors Journal compilation ©2011 Biochemical Society


Received 24 May 2011
doi:10.1042/BSI0391403