Published in final edited form as:
Insect Biochem Mol Biol. 2010 March ; 40(3): 189–204. doi:10.1016/j.ibmb.2010.02.001.

Structural cuticular proteins from arthropods: annotation, nomenclature, and sequence characteristics in the genomics era

Judith H. Willis
Department of Cellular Biology, University of Georgia, Athens, GA 30602, USA

Abstract

The availability of whole genome sequences of several arthropods has provided new insights into structural cuticular proteins (CPs), in particular the distribution of different families, the recognition that these proteins may comprise almost 2% of the protein coding genes of some species, and the identification of features that should aid in the annotation of new genomes and EST libraries as they become available. Twelve CP families are described: CPR (named after the Rebers and Riddiford Consensus); CPF (named because it has a highly conserved region consisting of about forty-four amino acids); CPFL (like the CPFs in a conserved C-terminal region); the TWDL family, named after a picturesque phenotype of one mutant member; four families in addition to TWDL with a preponderance of low complexity sequence that are not member of the families listed above. These were named after particular diagnostic features as CPLCA, CPLCG, CPLCW, CPLCP. There are also CPG, a lepidopteran family with an abundance of glycines, the apidermin family, named after three proteins in *Apis mellifera*, and CPAP1 and CPAP3, named because they have features analogous to peritrophins, namely one or three chitin-binding domains.

Also described are common motifs and features. Four unusual CPs are discussed in detail. Data that facilitated the analysis of sequence variation of single CP genes in natural populations are analyzed.

Keywords

R&R Consensus; whole genome sequences; *Anopheles gambiae*; *Bombyx mori*; cuticle

1. Introduction

1.1. Background

The most recent review of structural cuticular proteins (CPs) described the sequences of 139 CPs (Willis et al., 2005). This represents a considerable increase from the 38 complete sequences in the first major review (Andersen et al., 1995). In this group of 139 were 74 authentic CPs, defined as the sequence either coming from a protein extracted from cuticle, or corresponding to an N-terminal sequence of a protein extracted from cuticle. The remaining sequences came from isolation and sequencing of cDNAs, ESTs (expressed sequence tags) or short stretches of genomic DNA. Their assignment as CPs was based on sequence similarity to the verified CPs. The set of authentic CP sequences, most produced by Svend Andersen and

Appendix. Supplementary information: Eight files of supplementary information associated with this article can be found in the online version at

his collaborators, provides a solid foundation for all subsequent work, for the papers describing them identified or confirmed most of the motifs and other sequence features that are still used to classify a sequence as coding for a CP.

Since these reviews, several whole genome sequences have been made available. Detailed manual annotation has been carried out for the CPs of *Drosophila melanogaster*, *Anopheles gambiae*, *Apis mellifera*, *Bombyx mori* and *Nasonia vitripennis; Tribolium castaneum* is underway. One paper compares CPs of 7 *Drosophila* species (Cornman, 2009). Data from computer generated annotation are available for the pea aphid, *Acyrthosiphon pisum*, for the louse *Pediculus humanus corporis*, and for two non-insect arthropods, the crustacean, *Daphnia pulex*, and the tick, *Ixodes scapularis*. In addition, extensive collections of ESTs are coming on line for a broad array of arthropods. All of this has produced hundreds of sequences of putative CPs, recognized because of their similarity to the small number of authentic CP sequences. Only for *An. gambiae* has there been a concerted effort to verify that the annotated proteins are actually in the cuticle using LC/MS/MS to identify peptides isolated from cuticle that correspond to the translation products of the annotated genes (He et al., 2007). In addition to supporting over 90% of genes annotated on the basis of sequence similarity that study led to the recognition of new CP families. In total, genes for 240 cuticular proteins have been identified in *An. gambiae*, about 2% of its total protein coding genes (Table 1). The paper on the annotation of the *B. mori* CPs presents data for gene expression that, in addition to sequence similarity, were used to justify that these proteins are CPs (Futahashi et al., 2008). Thus the appearance of a transcript in epidermis during periods of cuticle secretion coupled with sequence similarity to known CPs is certainly adequate to assume that the particular transcript is coding for a putative CP. In both *Bombyx* and *Anopheles*, some proteins were identified that by sequence appear to be CPs, but have no additional supporting data. Such proteins were called CPH (for CP hypothetical) in *Bombyx* (Futahashi et al., 2008) and were the majority of the CPLCP family in *An. gambiae* (Cornman and Willis, 2009).

Help with protein identification is aided by Web sites that identify known consensus regions (described below) and the gene ontology category: GO:0042302. Of course, proteins identified in this manner are at best putative CPs. Furthermore, that GO term encompasses the collagens that make up the cuticle of nematodes as well as certain families of arthropod CPs. Information on spatial expression is available for many *D. melanogaster* transcripts at FlyBase (http://flybase.org/) when one searches under the “linkouts” for each gene. Especially useful are the microarray data for post-embryonic tissues at FlyAtlas and the *in situ* hybridization results on well-staged embryos at FlyExpress. Two other annotation studies have been accompanied by extensive expression data. Data for *Bombyx* are in (Futahashi et al., 2008; Okamoto et al., 2008), and at the Web site SilkBase (http://morus.ab.a.u-tokyo.ac.jp/cgi-bin/index.cgi). Temporal expression data across 19 developmental stages from hatching to adult eclosion for the *An. gambiae* CPs are available (Togawa et al., 2008; Cornman and Willis, 2009).

Rapid and inexpensive sequencing technology indicates that the number of sequences that resemble CPs will be expanding rapidly. Hence, before such data overwhelm us, it seems appropriate to summarize and categorize what we have learned from whole genome sequences and to summarize the defining characteristics and phylogenetic distribution of known CP families. One important advantage of whole genome data, when they have been properly mapped to chromosomes or at least scaffolds, is that it minimizes problems in accurate assessment of gene number that arise when one finds sequences that are almost identical. Such sequences could be due to different alleles of the same gene or to distinct but similar genes. Thus, this review was designed to summarize what has been learned about CPs in diverse arthropods, focusing primarily on data obtained from whole genome sequences.

After an introduction on cuticle protein nomenclature, the review will be organized by the protein families identified to date. It will point out, insofar as possible, the defining characteristics of each family and their taxonomic distribution. Then motifs shared among CP families will be described, a few specific CPs that illustrate interesting issues will be presented, and finally variations in four CP genes in natural populations will be described.

1.2. Cuticle protein nomenclature

This review has divided cuticular proteins into 12 different families. But such classification is artificial and subject to change. In most cases, it was based on a defining motif. But, some easily recognizable short motifs may be present in some members of different families (Section 3). While they provide support for calling a protein a cuticular protein, they do not define a family. Some families were identified based on chromosomal linkage of similar genes. There is a hierarchy to family nomenclature. A feature such as the R&R Consensus (named after the Rebers and Riddiford Consensus discussed in Section 2.1) takes precedence over shorter features. The 12 families of CPs (Table 1) fit the criterion of being a group of genes within a species that share common features. A collection of orthologs among species does not constitute a family. Hence orthologs of BcNCP1 (Section 4.4) are not a family. Indeed, in many cases orthologous genes are clearly members of well characterized families of paralogous genes. All of the CP families discussed in this paper have members in more than one species and in a limited number of cases, phylogenetic relationships among family members have been analyzed in some detail. Sequence motifs characteristic for several families are given in Supplementary Information File 1 in a FASTA format that can be used for BLAST searching.

Another goal of this review is to suggest guidelines for CP nomenclature, so that names alone will provide some clues as to the nature of the protein. This goal is complicated because different genome project leaders have established firm rules for naming genes of a particular species. The need for consistent nomenclature is especially critical with whole genome sequences so that distinct genes and differentially spliced transcripts, in the rare cases where they exist, are easily identified. Furthermore, authors are urged to indicate when sequences from whole genome analyses correspond to names previously given individual proteins. The database cuticleDB (http://bioinformatics2.biol.uoa.gr/cuticleDB/index.jsp) is attempting to serve as a repository for all structural CPs, but its success and utility will depend on investigators taking the time to properly submit their sequences.

An effective method for naming CP genes is to preface each name with a genus/species abbreviation of three or four letters followed by the protein family name and then the number of the gene in that family. Ideally, the genes should be numbered in their order on chromosomes, but annotation generally precedes complete assembly of a genome, and additional genes are frequently discovered when different search strategies are employed. And, of course, this method is not applicable to sequences obtained from ESTs and cDNAs. Thus although this naming strategy was planned for *An. gambiae*, problems quickly arose so that the stretches of genes in numerical order are frequently interrupted. Nonetheless, it is instantly obvious to those who work with cuticle proteins that a gene called *AgamCPR125* will code for a protein with the Rebers and Riddiford (R&R) Consensus and was identified in *An. gambiae*. Given the vast number of CPR genes, and complex patterns of amplification of paralogs, it is probably not wise to use the same number to name a similar CP in another species, although that was done in some pre-genomics work. Orthologs can best be described by presenting data in tables, and indeed, the meaning and identification of an ortholog among CPs is not straightforward (Section 4.3).

Such a logical scheme of nomenclature had to be abandoned for *D. melanogaster* where many CPR genes had prior names and where nomenclature rules forbade the use of three capital letters. Hence in one chromosomal region, 65A on chromosome 3L, one finds 18 CPR genes

with three distinct sets of names, *Lcp65A(a-g); Acp65A* and *Cpr65A*(*u-z*). Three of the genes have a number after the last letter because there are two or three genes coding for almost identical proteins. Use of ACP and LCP for adult and larval CPs was common in the early period of cuticle protein identification, but should be avoided as few CPs are stage specific. And, of course, now that 12 different species of *Drosophila* have been sequenced and sequencing has begun for at least an equal number of *Anopheles* species, it is essential that each of the names be preceded by an abbreviation designating the particular species. Three letter abbreviations for species and other guidelines for nomenclature are available at http://flybase.org/static_pages/docs/nomenclature/nomenclature3.html#2.5.1. The convention, followed in this review, is that both gene names and the three letter genus/species abbreviation are italicized. Protein names are not italicized.

2. Cuticular protein families

Several distinct families of CPs have been recognized. Their presence, based on current data, in three subphyla of Arthropoda including 8 orders of Hexapoda, is summarized in Table 2. It reveals that several families are found throughout the arthropods, but others are restricted to a particular order, or even lower taxonomic groups. All of these protein families appear to be restricted to arthropods except for the claim that the CPAP (obstructor) family with three chitin binding domains had been identified in nematodes (Behr and Hoch, 2005). But that appears not to be the case (Section 2.10). For the other CP families, it is not a lack of sequence data that explains the absence of these families from nematodes. It would be interesting to learn whether any of the CPs appear in the Onychophora, now accepted by most scientists as the sister group to arthropods (Edgecombe, 2009). Although 1904 EST sequences from onychophorans are available on PubMed (Roeding et al., 2007;Roeding et al., 2009) they came from adults so the absence of any recognizable CP sequences is not significant. Furthermore, given the similar cuticle construction throughout the Ecdysozoa (Schmidt-Rhaesa et al., 1998), it is intriguing that the arthropods seem to have adopted so many unique configurations for their CPs.

One group of sequences was originally lumped together based on the presence of low complexity sequence and named AgamCPLC#. Low complexity regions (Wootton and Federhen, 1993) are also commonly found in the CPR proteins, but there the R&R Consensus provides a defining characteristic. Careful examination of the large group of low complexity proteins revealed diagnostic domains that allowed them to be placed in distinct families (Cornman and Willis, 2009). As discussed below, a small group of *Bombyx* CPs appears to belong to the CPLCP family, first recognized in mosquitoes, rather than the CPGs where they were originally placed. As more sequences become available, our recognition of distinct families is certain to change. What must remain constant is the need to look for defining characteristics of an assembly of proteins within a species before they can be called a family. And, it must be remembered that in the absence of confirming evidence, calling a protein a CP based solely on sequence similarity at best makes the protein a putative CP.

2.1. CPR Family

The CPR family, named after the presence of the R&R Consensus, is by far the largest CP family in every species of arthropod examined and had the most representatives in the pre-genomic era. In 1988, Rebers and Riddiford recognized a common motif of 35 amino acids in 6 CP genes. The motif was G-x(8)-G-x(6)-Y-x(2)-A-x-E-x-G-F-x(7)-P-x-P. The number of sequences and species with this motif rose quickly, appeared in arachnids and crustaceans, and the consensus came to be referred to as the R&R Consensus. It appears to be restricted to arthropods. The one reported CPR sequence in a vertebrate, *Xenopus* (Klein et al., 2002), NP_001090156.1) turns out to be identical (in both protein and nucleic acid sequences) to a

CP of *Drosophila erecta* (XP_001972081.1). So, at present, the CPR family is arthropod-specific.

Now that several hundred CPR sequences are known, it is appropriate to see how well the initial Consensus has fared. The Pfam database includes an extended version of the R&R Consensus as pf00379, chitin_bind_4. Hence, any protein suspected of having this Consensus can be submitted to Pfam at http://pfam.sanger.ac.uk/search. The 2004 pf00379 consensus version based on 82 sequences was: PDGDYNY+YETSNGIADQETGD+KSQGETRDG+ +AVDVV+//**G**SYSYVDPD**G**TTRTVT**Y**T**A**DDEN**G**FQ PVGAHLP. A + indicates that multiple amino acids may occupy this position, just like the x in the original Consensus above. A double slash indicates the start of the original Consensus. Bolded and underlined are what Pfam found as invariant residues. So, by the time this version was constructed, two lessons had been learned. Most importantly, the conserved region extends N-terminally by 40 amino acids, doubling its length. Second, this sequence shows that near the N-terminus is an "aromatic triad", here shown as YNY+Y (underlined). Examination of over 500 CPR sequences revealed some more variation but the common structure remains. The original start of G-x(8)-G-x(6)-Y remains almost invariant. The amino acids in the aromatic triad (positions 5,7,9) can be Y or F or W. This triad is present in the vast majority of CPR proteins, occasionally there is only a diad, and even more rarely only a single aromatic residue is present in that region. The Y at position 57 is very rarely F, but the residue at position 65 is commonly Y. Interestingly, 11 *Bombyx* proteins lack the terminal aromatic residue (position 65), some of these terminate at G (position 64), and others go on for a few more amino acids (Futahashi et al., 2008). Nonetheless, the R&R Consensus remains easy to identify and is by far the most abundant motif among cuticular proteins.

Two major groups of CPR proteins are recognized, RR-1 and RR-2. The pfam00379 sequence is really a composite of both RR-1 and RR-2. A third, very minor form called RR-3 has been recognized (Andersen, 2000), but distinctive features have not been defined. The Web site cuticleDB (http://bioinformatics2.biol.uoa.gr/cuticleDB/index.jsp) has a tool that uses Hidden Markov Modeling to judge whether a sequence can be called RR-1 or RR-2. The conservation of these two groups is supported because while the model was developed on sequences from *D. melanogaster* it works well with proteins from diverse species including several chelicerates. The sequence logo for the chelicerates (Fig. 1) reveals that the RR-2 Consensus is consistently 2 amino acids shorter than for the other arthropods. Many RR-2 proteins have GFNAVV near the end of the Consensus, but this is not found in the available crustacean and chelicerate sequences. Some proteins that clearly have the R&R Consensus cannot be classified with this tool for reasons that are not clear. Furthermore, the tool uses amino acids N-terminal to the aromatic triad in its classification of RR-2 sequences; additional data indicate that these are not informative. The RR-1 form is variable in length from position 1 to the start of the original Consensus (here position 41) but almost constant in length thereafter. There are almost always 5 amino acids between the aromatic residue at position 57 and the G before the (FY) near the carboxyl terminus. The RR-2 form of the Consensus is constant in length and has 6 residues in that region. These differences were first described in 2005 based on 95 proteins and are valid today when hundreds more CPR sequences are known. Sequence logos for RR-2 from a collection of chelicerates and crustaceans can be compared to those from *Bombyx* and *Anopheles* (Fig. 1). They illustrate how conserved this form of the Consensus has been over more than 600 myr. The greater variability in the RR-1 Consensus is illustrated in Fig. 2. Interestingly, the cuticleDB classification tool does not recognize RR-1 sequences in chelicerates, although examples of YTADENGF, a common region in insect RR-1 proteins is found.

Early examples of proteins with the Consensus showed a correlation of RR-1 proteins coming from soft (flexible) cuticle and those with the RR-2 form from hard (rigid) cuticles.

Subsequently, Andersen (2000) suggested that the difference might be that RR-2 proteins were predominantly from exocuticle and RR-1 from endocuticle. It was anticipated that analysis of the expression patterns of most of the 156 CPR genes in *An. gambiae* would resolve the issue. Quantitative real-time RT-PCR was used to measure transcript levels for individual genes at 19 time points from hatching to just after adult eclosion (Togawa et al., 2008). If a transcript does not appear until after ecdysis, it can be assumed that following translation the protein will appear in endocuticle, as by definition, endocuticle deposition is post-ecdysial. Similarly, if a transcript is present only in a pharate animal, the corresponding protein probably will be used in forming exocuticle. But if a transcript appears first in a pharate stage and persists into the next stage, is it being used to contribute to post-ecdysial cuticle? One immediate complication is that the mRNA for this study was isolated from entire animals, and cuticle formation proceeds at different rates in different regions. Thus scales are fully formed at the time of eclosion while abdominal and thoracic cuticle continues to grow after eclosion and some CPs are used in tracheae and gut linings. The vast majority of CPR genes, both RR-1 and RR-2, had transcripts present in both pharate and post-eclosion stages which does nothing to resolve the issue. But out of 99 genes with transcripts exclusively in pharate stages, only 11 (10%) were RR-1, while RR-1 proteins are 34% of the total. Definitive data await an examination of the precise location of individual proteins in the cuticle, something that will require EM immunolocalization. (See Willis et al., 2005 for review of such studies, most of which were done with antibodies raised against proteins whose sequences were not known.)

Most proteins have only one occurrence of the R&R Consensus. Thus it was a surprise when Ikeya et al. (2001) published the sequence of a protein from the tailfin of a prawn (*Marsupenaeus* [*Penaeus*] *japonicus)* that had 14 consecutive Consensus regions, of 6 different varieties. A more recent paper describing chitin binding proteins from the horseshoe crab, *Tachypleus tridentatus*, (Iijima et al., 2005) described a protein (BAE44187) with 5 Consensus regions, and the *Ixodes* genome project has many predicted proteins with multiple Consensus regions, although manual annotation has been yet to be done on these. The computer-generated annotations of several insect species had defined several proteins with multiple Consensus regions that subsequently turned out to be multiple genes, easily recognized by the presence of TATA boxes, transcription initiator elements, Kozak consensus translation start sites, polyA addition sites, etc. At present there are very few insect CPR proteins that have more than a single Consensus region. The only one known with more than two Consensus regions has orthologs in several species (Section 4.2).

That the Consensus must be serving some fundamental function was surmised by Rebers and Riddiford (1988). The suggestion that the Consensus might confer chitin binding properties to a protein was first mentioned by Bouhin et al. (1992) and Charles et al. (1992), and was frequently repeated. Then Rebers and Willis (2001) published a paper that established that the *An. gambiae* protein, Agcp2b (now known as AgamCPR97) would bind to chitin beads. More importantly, 65 amino acids corresponding to the extended version of its Consensus was sufficient to convert glutathione-S-transferase into a chitin-binding protein. The region studied began 4 amino acids upstream from the aromatic triad and ended 8 after the final GF, i.e. positions comparable to those on the pfam00379 sequence given above. It is becoming common to use the term R&R Consensus for this expanded version and that practice is encouraged. There are three more published accounts that verify that this extended Consensus can confer chitin binding properties on a protein (Qin et al., 2009; Togawa et al., 2004; Togawa et al., 2007). Provocatively, Togawa et al. (2004) also obtained chitin binding activity from shorter regions of the extended Consensus and other regions of the protein, but they used a low salt concentration in their binding buffer and had background binding from glutathione-S-transferase alone.

Homology structural models of the R&R Consensus region from a lepidopteran RR-1 protein (HCCP12) and from a composite of numerous RR-2 proteins were constructed using a lipocalin (retinol binding protein) as a template, even though CPRs are not lipocalins (Hamodrakas et al., 2002; Iconomidou et al., 2005). The models presented a half barrel structure with an opening nicely sized to hold a chitin chain and with aromatic residues, shown by Rebers and Willis (2001) to be essential for binding, positioned inside the opening (Iconomidou et al., 2005). Unfortunately, we still do not know how many chitin chains interact with a single protein, and now there is evidence that deacetylated chitin may also be an essential component of cuticle (Arakane et al., 2009). Furthermore, such models, especially when based on such limited similarity, are tentative and experimental data are essential to provide definitive structures.

From 32-156 CPR proteins have been found in the species whose genomes have been subjected to manual annotation (Table 1). It appears that *Aedes aegypti* may have even more (Cornman and Willis, 2008). Sufficient data exist to make it possible to speculate on the basis for the differences in numbers. *Apis mellifera* with 28 CPRs (plus 4 more not yet submitted) at present is the lowest followed by *Nasonia vitripennis* with 62 (Honeybee Genome Sequencing Consortium, 2006;*Nasonia* Genome Working Group, 2010). Neither species has to deal with the outside world until they reach the adult stage. *Apis* is provisioned by worker bees and *Nasonia* in its larval and pupal stages is a parasite residing within the protected environment of the puparium of its host fly. Of course, it is possible that the small number of CPR proteins is a hymenopteran condition; something that should be resolved as soon as the genome of the tsetse fly, *Glossina morsitans*, becomes available for this dipteran also has a protected larva that pupariates immediately after extrusion from its mother.

The CPR genes in *Apis* and *Nasonia* are a subset of those found in other species. Three quarters of the *Apis* CPR proteins had convincing orthologs in *Nasonia*. All others were represented by two or more closely related *Nasonia* sequences (*Nasonia* Genome Working Group, 2010). All *Nasonia* CPR sequences had clear orthologs or paralogs in *Drosophila*. It is also intriguing that while the R&R Consensus region is, on average, 81% identical among *Apis/Nasonia* orthologs, the entire protein is far less conserved with an average of 58% The presence of clear orthologs across orders and about 300 myr suggests that particular proteins are serving important and distinct (as yet unknown) functions. The selective pressures and mechanisms that contribute to expansion and contraction of these gene families and act on different parts of each protein are likewise tantalizing issues.

The large number of CPR proteins found in mosquitoes but not other dipterans is explained by an analysis of the similarities among sequences. Cornman (2009) has analyzed the number of CPR sequences in 7 of the 12 species of *Drosophila* whose genomes are available. The number ranged from 100-105. Many of the genes are close to one another on a chromosome, indeed ~75% of the CPR genes are found in 15 tandem arrays (genes within 20 kb of one another). Within these tandem arrays, some genes occur in pairs of almost identical adjacent sequences. The situation in the mosquitoes is quite different. Here there are also tandem arrays (8 in *An. gambiae*), but within a single tandem array there may be up to16 almost identical sequences, referred to as sequence clusters. Only RR-2 genes have been found in sequence clusters and each has a distinct form of the R&R Consensus. The number of genes in sequence clusters is even larger in *Aedes* (Cornman and Willis, 2008). The genes in sequence clusters clearly account for much of the difference in numbers of CPR proteins between *An. gambiae* and *D. melanogaster*. *An. gambiae* has 102 RR-2 proteins; of which 69 belong to 8 sequence clusters, allowing for 1 gene to be representative of each sequence cluster, 102-61 = 41, close to the 34 RR-2 genes found in *D. melanogaster*.

The expression patterns of the *An. gambiae* sequence clusters reveal a probable explanation for their existence. All genes in sequence clusters are expressed in pharate stages, when the

mosquito is forming the cuticle of the next stage. Real-time RT-PCR revealed that many of the genes in sequence clusters have very high levels of mRNA abundance (Togawa et al., 2008). *D. melanogaster* larvae develop in a relatively predator-free environment whereas mosquito larvae are exposed not only to predators but to the risk of their larval environment drying out before they have completed development. Thus one could well imagine strong selective pressure for rapid development and for rapid cuticle formation in pharate stages of mosquitoes so that there is only a very brief period when the larva or pupa will not have a sufficiently thick exoskeleton to allow movement.

B. mori also has three distinct sequence clusters involving 15 of its 89 RR-2 genes (Futahashi et al., 2008). All three clusters have Consensus regions that are most closely related to AgamCPR70, suggesting less diversification in the genes that were amplified than in *Anopheles*. Here the number of distinct RR-2 genes is clearly higher than in the dipterans (89-12 = 77). There are insufficient data to be able to speculate why these three small clusters may have arisen. Futahashi et al. (2008) suggest that the large number of RR-2 genes may "allow for the possibility of divergence of body surface and scale structures to adapt to the environment in lepidopteran species." The data behind this claim came from the finding that RR-2 proteins contribute in precise ways to the formation of tubercles on early instars of the swallowtail butterfly *Papilio xuthus* (Futahashi and Fujiwara, 2008).

This discussion of the CPR family has focused on the Consensus. But the functional properties of the proteins depend also on the flanking sequences. It is interesting that in some cases orthologs (defined by high identity in the Consensus region) can have similar flanking regions and in other cases have quite different flanking regions (see discussion of resilin, Section 4.3). An attempt to analyze this issue can be found in Supplementary Table 5 of Cornman et al., 2008, which summarizes amino acid similarities in flanking regions of 22 CPR presumptive orthologs in *An. gambiae* and *D. melanogaster*. Although the Consensus region could begin as close as 6% from the N-terminus of the mature protein and as late as 90%, all but two of these ortholog pairs had comparably positioned Consensus regions. Given the importance of histidine and lysine residues in cross-linking, it was interesting to find large differences in the percent of one or both of these amino acids in half of these pairs. In addition, Cornman et al. (2008) identified a proline-rich region as commonly found in RR-1 proteins of *An. gambiae* adjacent to the Consensus, with the sequence GFQPQGxHxPxPPP, and the sequence RDGDVVKG was found in many RR-2 proteins. Some proteins have abundant histidines, frequently regularly spaced, many have a large proportion of glycine or alanine residues, others emphasize prolines and several have stretches of 5 or more glutamines. Obviously any understanding of the precise roles these proteins play will have to involve consideration of the entire protein.

2.2. CPF and CPFL families

The CPF family was first recognized by Andersen et al. (1997) in a total of six CPs from *Tenebrio* and *Locusta*. A stretch of fifty-one amino acids was given as its defining consensus: (AY)-(AP)-x(2)-(PA)-(PA)-//A-(LIV)-x-(SA)-(QS)-x-(SQ)-x-(IV)-(LV)-R-S-x-G-(NG)-x(3)-V-S-x-Y-(ST)-K-(TA)-(VI)-D-(TS)-(PA)-(YF)-S-S-V-x-K-x-D-x-R-(VI)-(TS)-N-x-(GA)//-(IVL). The family was named CPF after this 51-amino acid motif. Once again, just like with the CPR family, an early observation turned out to have defined a family of CPs. Here also, the original consensus has undergone considerable modification, this time being shortened.

The modification came about when Togawa et al. (2007) searched available protein sequences and came up with a sizeable array of proteins that matched part of the consensus. As a result, they concluded that a conserved motif was only 42-44 amino acids long, fortunately, still capable of being called CPF. That revised consensus resides in the region between the pair of //

marks. The same analysis also revealed the presence of another group of proteins that lacked the consensus but had considerably similarity at the carboxyl-end of the molecule. These proteins were placed in a new family called CPFL – for CPF-like. CPFs and/or CPFLs have been recognized in 7 orders of insects (Table 2).

At present there is no indication of the function these proteins may serve in cuticle, but they are definitely cuticular proteins as He et al. (2007) identified unique peptides from all 4 AgamCPFs and AgamCPFL1 and shared peptides from AgamCPFL2-7 in a proteomics analysis of cast cuticles. Chitin binding assays with recombinant AgamCPF1 and CPF3 proteins carried out in the same manner that had been used successfully by Rebers and Willis (2001) failed to detect chitin binding, while a positive control AgamCPR21 run at the same time did bind (Togawa et al., 2007). It would be premature to conclude that the CPF proteins do not bind chitin, given that the assay is not run under conditions that in any way mimic the normal environment in which these proteins interact with chitin. Furthermore, recombinant CPFs used in this study, unlike the CPR proteins, ended up in the inclusion bodies of *E. coli* during expression, and it took conditions established though using a Pierce Refolding Kit® to render them soluble and keep them soluble when dialyzed against binding buffer.

The best clue we have to the use to which these proteins are put came from expression analyses. All 4 of the AgamCPF genes had mRNA only during pharate stages, with *CPF1* and *2* being expressed in pharate larvae and pupae and *CPF3* and *4* in pharate adults (Togawa et al., 2007). The 7 *An. gambiae* CPFLs did not have detectable mRNA until later during pharate development and the message was still present immediately after the molt to the next stage (Togawa et al., 2007). Whether the CPFs are used to form epicuticle, the one cuticular region that lacks chitin, or just exocuticle remains to be resolved. The AgamCPFL genes had transcripts at similar levels in both the pharate and freshly ecdysed larvae, quite a distinct pattern from the CPFs.

2.3. Tweedle (TWDL) Family

This family was first identified in *D. melanogaster* and was named because of the resemblance of a mutant of one member to the corpulent Tweedledee in “Alice through the Looking-glass.” It was suggested that the cause of the mutation, named TweedleD, was a cuticle of insufficient thickness to maintain the normally slender shape of a *Drosophila* larva. Twenty-seven members of the Tweedle family were identified in *D. melanogaster* (Guan et al., 2006). The number is smaller in other species (Table 1), but the family is widespread with representatives already identified in the six insect orders for which sufficient sequence data are available (Table 2). It was not identified in Crustacea or Chelicerata. A neighbor-joining phylogeny revealed that 23 of the *D. melanogaster* TWDLs came from a *Drosophila* specific expansion. A smaller expansion with 6 members was seen for *An. gambiae*, with these plus an additional 6 members forming a mosquito specific expansion. The rest of the TWDLs did not cluster by species or order (Cornman and Willis 2009).

Rather than a continuous stretch of amino acids that define the group, there are four small regions, recognized by Guan et al. (2006) in *D. melanogaster*. These can easily be visualized in the diagrams in Figure 3. The two diagrams represent a continuous sequence that was separated for clarity. The sequences chosen were representatives of 5 orders, with only one representative from the expanded group in *Drosophila*; 9 *Ae. aegypti* sequences were used because they were found throughout the phylogeny. Sequences from the pea aphid *Acyrthosiphon* and the louse *Pediculus* were also used in constructing the sequence logo to illustrate the conservation of the conserved region across 300 myr. The identifying region has been identified as pfam03103, but its consensus does not appear to be very representative (Cornman and Willis 2009).

TWDL family members now join the CPRs as binding to chitin. This important discovery was based on the presence of BmorCPT1 in a chitin binding fraction of *B. mori* larval proteins and by direct measurement of the binding of a recombinant CPT1 protein to chitin beads (Tang et al., 2010).

2.4. CPLCA Family

A small family of CPs rich in alanine residues (13-26%) is present in Diptera but not in other orders. This family designation is not based on the alanine content which can be as high in other CPs, rather the key feature is the presence of the retinin domain (pfam04527/IPR007614). The *D. melanogaster* retinin protein that is expressed in the cornea (Kim et al., 2008) has only 10.6% alanine residues and is an outlier in a phylogeny of this group (Cornman and Willis, 2009). Two of the three members of this family in *An. gambiae* were represented by numerous peptides in a proteomics analysis of cast cuticles (He et al. 2007). Although a WebLogo is presented in Fig. 4, at present it is probably best to rely on the retinin domain to identify members of this family.

2.5. CPLCG Family

Two members of this family were first recognized in *D. melanogaster*, but a larger number turned up in a proteomics analysis of *An. gambiae* cuticles. The *Drosophila* proteins were named after their presence in the adult as Dacp1 and Dacp-2 (Qui and Hardin, 1995), but a better name is based on an abbreviation that recognized two glycines in a shared motif near the carboxyl-terminus. The signature, at residues 5-16 in the WebLogo in Fig.3, is G-x(2)-H-x-A-P-x(2)-G-H, but residues over 35 amino acids are well conserved.

Twelve of the 27 genes in *An. gambiae* are members of a sequence cluster (Group A) and the entire set is found in a 195 kb array on chromosome 3R. The genes in *Aedes aegypti* and *Culex pipiens* are also found in tandem arrays, but over a much greater distance. The organization of these sequence clusters presents a challenging scenario for their evolution and maintenance. The genes in Group A are not adjacent, and indeed are interspersed with genes from another CP family (CPLCW) and even by unrelated genes (Cornman and Willis, 2009).

Although Cornman and Willis (2009) reported that the CPLCG family was restricted to the Diptera, this is incorrect. A tblastn search using the 35 amino acid consensus given in Supplementary Information File 1 against “EST other” revealed numerous examples in the primitive wingless insect, *Lepismachilis y-signata*, from the order Archaeognatha, in the crustacean, *Daphnia pulex*, as well as in *Blattella germanica* and *T. castaneum*. Not only is the consensus well conserved (Supplementary Information File 2) but it lies at the carboxyl-terminus of the short proteins, just as it does in the Diptera.

2.6. CPLCW Family

This small family (9 members in *An. gambiae*) was named after its invariant tryptophan residue that can be seen at position 10 in the sequence logo (Fig. 3). The members of this family range in length from 106-175 amino acids and share 92-100% amino acid identity with one another in *An. gambiae*, where, in a neighbor-joining tree, they cluster into a distinct group relative to the groups found with the corresponding genes in *Aedes* and *Culex* (Cornman and Willis, 2009). The sequence logo was based on all three mosquito species, for so far the family appears to be restricted to mosquitoes. As mentioned above, this family in *An. gambiae* is interspersed in a tandem array with the sequences for the CPLCG genes. Yet the latter family is present in *Drosophila* that has no CPLCW genes. They are clearly distinct families for the average sequence similarity between two *An. gambiae* CPLCG and CPLCW proteins is 20%.

2.7. CPLCP Family

This family was first recognized in *An. gambiae* when peptides corresponding to four genes turned up in a proteomics analysis of proteins from cast larval head capsules and pupal cuticle (He et al., 2007). These are now named *AgamCPLCP8,10,11,12*. An additional 24 similar genes were identified in *An. gambiae*; these were all grouped into the CPLCP family. None of the 24 yet has proteomics support but their expression profiles match those of genes from other low complexity families known to be authentic CPs (Cornman and Willis, 2009). The prolines occur primarily as PV or PY frequently adjacent to one another. The 92 proteins used by Cornman and Willis (2009) in their analysis had 1785 instances of PV and 850 occurrences of PY. These pairs are common in many cuticular proteins but not in the density found in the CPLCPs. For example, a search for these residues among the *Bombyx* protein classified as CPGs (Futahashi et al., 2008) immediately revealed the 6 genes that had been classified as CPLCPs by Cornman and Willis (2009). All CPLCP members were reported to share features in addition to blocks with a high proline content (Cornman and Willis, 2009). One such feature was the presence of GLW[D/E], but this feature is restricted to mosquitoes. This tiny motif was reported to be flanked by a region rich in polar and acidic residues and one rich in glycine, tyrosine and histidine, with many occurrences of histidines near the C-terminus. These features are not as consistent as the high density of PV and PY. This gene family has been identified with small numbers present in other insect genomes (Table 1). Once again, both *Aedes* and *Culex* have larger families (19 and 24 members, respectively), with the majority being found in mosquito specific gene expansions (Cornman and Willis, 2009).

2.8. Glycine-rich protein Family

Before whole genome sequencing, several CP genes had been identified in *Bombyx* that were high in glycine content (Zhong et al., 2006). With the completion and annotation of the genome, it was appreciated that there were 28 such genes in the genomic data most with GGYGG or GGxGG repeats and these were named CPG for “cuticular proteins glycine-rich” (Futahashi et al., 2008). This grouping appears rather heterogeneous and a blastp search against all arthropod proteins revealed several interesting features. Details are in Supplementary Information File 3. Six of the genes coding for proteins with but 0-3 GGY motifs resemble genes in the CPLCP family of *An. gambiae* (Cornman and Willis, 2009), a family not yet identified when the *Bombyx* sequences were annotated. Only 10 had five or more GGY motifs; on BLAST searches, all 10 were lepidopteran specific, especially with matches to *Papilio xuthus* (Futahashi and Fujiwara, 2008). The only matches for 8 others were also only to other Lepidoptera. Three of the 4 remaining had paralogs in *D. melanogaster* that had not been identified as CPs, although three have been reported to be expressed in cuticle-secreting tissues (Supplementary Information File 3). GGY repeats are not a unique feature of this group of proteins. The Web site cuticleDB has numerous examples, with the first ones having been found in the earliest CP sequencing studies from locusts; some proteins have no other family signature, but many clearly belong to the CPR or CPF families. Futahashi et al. (2008) wisely only assigned proteins with no other family signatures to their glycine-rich protein family, and the high number of GGY repeats in several *Bombyx* CPGs is exceptional. At present, it appears that the glycine-rich family identified in *Bombyx* has but 18 members and is restricted to the Lepidoptera (Supplementary Information File 3).

2.9. Apidermin Family

Genes for three highly hydrophobic proteins, with at least a 30% alanine content, were spotted in an 11 kb genomic region of the *Apis mellifera* genome and named apidermin 1-3 (Kucharski et al., 2007). The mature proteins ranged in size from 6.1 to 9.2 kD. Analyses of temporal and spatial (for *apd-1*) expression revealed that the apidermins are expressed in cuticle forming epidermis with *apd-1* having the broadest expression predominantly in epidermis underlying

cuticle destined for sclerotization. In contrast, *apd-2* and *3* are expressed in tracheae and various parts of the digestive tract with *apd-3* also being expressed in external epidermis, including the eye where *apd-1* transcripts are not found. Three similar genes have been identified in *Nasonia*, but as presently annotated they are considerably larger, from 21.6 to 37 kD. Attempts to find homologs in other groups are not productive because of the absence of unique features defining the family.

2.10. CPAP1 and CPAP3 (Gasp-Obstructor) Families

One consistent characteristic of the CPs is that they rarely have a cysteine residue. Indeed the absence of cysteines is a hallmark of CPs (Andersen et al., 1995; Willis et al., 2005). Thus it was unexpected when a cuticular protein with 18 cysteine residues, organized into three repeats of the six-cysteine-containing ChtBD2 chitin-binding domain, separated by two spacers was described in *D. melanogaster*. The corresponding gene was shown to be expressed in embryonic tracheae and named *gasp* (Barry et al., 1999). ChtBD2 previously had been associated with proteins found in peritrophic matrices, named peritrophins. Then Behr and Hoch (2005) identified more *D. melanogaster* proteins with a similar structure to gasp and classified them as belonging to the obstructor multigene family. Their name obstructor came from the mutant of one that displayed "a barrier brake-down" phenotype, a term that was never defined. *In situ* hybridization revealed that these proteins were expressed in regions of the developing embryo that form cuticle, in tracheae and epidermis, once again inappropriate locations for peritrophins. Two distinct groups of obstructors differing in the lengths and sequences of spacers that separate adjacent ChtBD2s were identified, A-E and F-J. (*gasp* was classified as *obst-C*, although the original name remains on FlyBase.)

Now a comprehensive analysis of proteins with the ChtBD2 domain has been carried out in *T. castaneum* (Jasrapuria et al., 2010). Among the 42 proteins found, annotated and verified with sequencing of cDNAs, 13 were chitin metabolic enzymes (chitinases, chitin deacetylases, chitin synthases etc.); the remaining 29 had signal peptides indicating that they are secreted. Jawrapuria et al. (2010) have revised the consensus for ChtBD2 originally proposed by Tellam et al. (1999). It now is: $CX_{11\text{-}24}CX_5CX_{9\text{-}14}CX_{12\text{-}16}CX_{6\text{-}8}C$. Eleven proteins were very similar in sequence to previously identified peritrophins and were put in a family named PMP for **P**eritrophic **M**atrix **P**roteins. All were found to be expressed in the midgut that makes the peritrophic matrix.

Two other families with ChtBD2 were identified that were shown by RT-PCR to be expressed in cuticle-forming tissues. Seven *T. castaneum* genes (one coding for two proteins by alternative splicing) were paralogs of *gasp* (obstructors), none of these were expressed in the midgut and these have been given the family name CPAP3 (**C**uticular **P**roteins **A**nalogous to **P**eritrophins) with the 3 indicating that they have 3 ChtBD2 domains. Another *T. castaneum* family had 10 members, each with a single ChtBD2. This family was named CPAP1. A phylogenetic analysis carried out with a shortened sequence that excluded linker regions between the ChtBD2 domains (see Jasrapuria et al. 2010 for details). The three families, PMP, CPAP1 and CPAP3, formed distinct branches.

Similar proteins were found in numerous other insect orders (Table 2). It is hoped that the names CPAP1 and CPAP3 will be adopted to emphasize that these two distinct families are not peritrophins although they share its ChtBD2 domain and contribute to an interface with the outside world.

An exciting finding in the paper by Behr and Hoch (2005) was the identification of a paralog to their "obstructors" in *C. elegans*. This protein (now annotated NP_490942.1 = NM_058541.5) has a six-cysteine containing ChtBD2 domain at the C-terminus and a truncated version of this domain at the N-terminus, which is missing the first two cysteines of the revised

consensus for ChtBD2 (see above). Other *C. elegans* proteins, including NP_502145.2 (NM_069744.4), CEJ-1 (DQ340623.1) and B0280.5 with multiple ChtBD2 domains, the latter two with demonstrable chitin binding affinity have also been identified. Proteins orthologous to these *C. elegans* proteins are found in *Brugia malayi*, the filarial parasitic nematode. RNAi phenotypes with the *C. elegans* proteins, CEJ-1 and B0280.5 are consistent with defects in the eggshell, a chitin-bearing structure (Johnston et al., 2006). Hence, ChtBD2 domain-containing proteins are definitely present in nematodes. Of all the motifs that define CP families discussed in this review, it is the only one that is found outside arthropods. Nonetheless, precise features of these nematode proteins do not allow them to be placed as paralogs of either of the two arthropod CP families with the ChtBD2 domain, CPAP1 or CPAP3. It seems likely that at least some of these chitin binding proteins are paralogous to PMPs because both groups have ChtBD2s and mucin domains.

2.11. Miscellaneous cuticular proteins not assigned to families

Three proteins identified with proteomics in *Anopheles* could not be assigned to any CP family (Cornman and Willis, 2009). BcNCP1 discussed below (Section 4.4) is another “orphan” protein, but one that has turned out to have orthologs in many species. There are 34 proteins in *Bombyx* assigned the name CPH (cuticular protein hypothetical) because definitive evidence for their participation in cuticle structure had not been identified. One of these, BmorCPH1, is the ortholog of BcNCP1; BmorCPH32 appears to belong to the CPLCP family. Three (BmorCPH17, 30, 31) share features outside of the R&R Consensus with members of the CPR family in *An. gambiae*. Whether these represent proteins truncated in evolution or annotation remains to be learned.

3. Motifs and features frequently found associated with cuticular proteins

There are motifs or short stretches of amino acids that are commonly found in CPs. The ones that appear in more than one CP family are discussed below. Most were first recognized by Andersen et al. (1995), and their continuous presence as the number of CP sequences increases is testimony to their stability. Unfortunately, the function of none is known.

3.1. The 18 amino acid motif

Nakato et al. (1990) described a motif of 18 amino acids present in 3 copies in a *B. mori* cuticular protein they named PCP (now known as BmorCPH31). Subsequently Andersen (2000) described the same motif, sometimes as single, sometimes as multiple copies in cuticular proteins from several other species. Andersen suggested that since it had been conserved during several hundred million years of evolution, “it has an essential function.” Subsequently this motif has turned up frequently in CPs. It is present in most but not all sequences that have been assigned as RR-3. It is also present, frequently in more than one copy, in sequences that lack the R&R Consensus, such as BmorCPH31. A sequence logo based on 40 occurrences in 27 proteins from 5 orders of insect and two crustaceans is shown in Fig. 4. It is a modification of the original consensus that is used by cuticleDB: VxDTPEVAAAxAAHxAAH. This motif was not found in any of the ~250 Chelicerata CP sequences available on PubMed. Many of the proteins in which it occurs have frequent and regularly spaced histidines. Unfortunately, no one yet has evidence about or even has speculated on its function.

3.2. AAP[AVL]

Another feature that Andersen et al. (1995) commented on was the frequent occurrence of AAP[AV]. Work with CPs from *An. gambiae* revealed that this tiny motif could be expanded to AAP[AVL]. While some have suggested that the presence of this motif might be a diagnostic feature of CPs, it certainly is not. Andersen et al. (1995) pointed out that it is found in chorion proteins, and in a non-collagenous cuticle protein of *C. elegans*. When it is present in chorion

proteins of *D. melanogaster* (43/171 proteins), it is never found in more than two copies, and in *B. mori*, only 5/20 chorion proteins have the grouping and only as a single occurrence.

In an attempt to learn if AAP[AVL] is a diagnostic characteristic for CPs, I looked for it in all predicted proteins in *An. gambiae*. There were over 300 occurrences; many were not in CPs. But the presence of several repeats in a single sequence does seem to be restricted to CPs. Hence in the 233 CPs annotated in that species, 13 CPRs and 6 members of other families had 3 or more instances. Multiple instances of the motif are common features in the small number of *Locusta* and *Tenebrio* sequences that are known. Hence, multiple copies of this short motif appear to be a diagnostic feature for CPs. This feature has evidently been used to establish that "the most abundant house dust mite protein" (AAP57092 from *Dermatophagoides farinae*) is a CP for it has 8 of these repeats.

There has been speculation on the conformation that this motif might adopt. After an examination of it in a variety of non-insect proteins, Andersen et al (1995) concluded: "A relevant feature of the Ala-Ala-Pro-Ala motif appears to be a strong tendency to form turns; several conformations can be present in equilibrium, indicating low energy barriers between the conformations. When the sequence occurs regularly in a protein, as it does in many of the CPs as well as in other structural proteins, it can be suggested that the result will be proteins folded in a more or less regular helix, which is easily and reversibly deformed by external forces, thereby resembling elastin." Interestingly, there are none of these groupings in human or rat elastin, although long stretches of primarily alanines, interrupted by prolines are common. The AAP[AVL] grouping is not found in the two insect proteins, one clearly resilin, experimentally demonstrated to have elastic properties (Elvin et al., 2005; Lyons et al., 2007). These resilin, resilin-like proteins will be discussed in more detail in Section 4.3.

3.3. Absence of cysteine residues

One of the generalizations frequently made about CPs (Andersen et al., 1995; Willis et al., 2005) is that cysteine is absent in the secreted form of the protein. Andersen (2005) suggested that cystine and cysteine could react with ortho-quinones and interfere with sclerotization.

There are major exceptions for specific proteins or families (see BcNCP1 in Section 4.4) and the discussion of CPAP1 and CPAP3 (Section 2.10). Nonetheless, for the most abundant families of CPs the absence of cysteine is evident. In an attempt to formalize this generalization, all 101 CPR proteins in *D. melanogaster* were examined. The selection of this species was based on its excellent annotation, extensive EST data to verify the genes and the recent addition of data from 11 more species of *Drosophila*. Numerous proteins had cysteine residues within their predicted signal peptides. Only four had cysteines within the mature protein. Cpr65Aw had cys as its final residue. Clear orthologs were identified in three other species, *D. sechellia*, *D. simulans* and *D. erecta*. None had this final cys. Each ended QVEH, rather than the *D. melanogaster* end, QVEHSSRDRFGHC. Provocatively, the first serine residue was coded for by TCA, hence only a single base change would have been needed to convert that codon into the TAA stop codon found in the other species. For the other three proteins, the presence of cys residues is conserved among most or all of the other sequenced species of *Drosophila*, although other regions of the protein are not identical, so the retention of those cys residues is not trivial. DmelCPR65Az has two cysteines present as CHGC ending 21 amino acids before the C-terminus of the protein. DmelCpr76Bd has a single cys residue just 13 residues down from the start of the mature protein. The final protein, Dmell(3)mbn has two cys (CSGC) beginning 101 residues in from the N-terminus. None of the 26 members of the *D. melanogaster* TWDL family has cysteine in the mature protein.

Thus, for the CPR proteins and TWDLs and most other CP families, the presence of cysteine is sufficiently rare to serve as a warning that an annotation might have an error.

4. Provocative Cuticular Proteins

Four CPs have been selected for special mention, one (dumpy) because it is unusual in many respects, one (resilin) because it challenges our understanding of the recognition of orthologs and the other two because of special features.

4.1. dumpy

There is an enormous (2.5 MDa) CP, dumpy, that has been identified in *D. melanogaster* (Wilkin et al., 2000). In addition to its size, it has two features that are atypical for a CP. First, in addition to a predicted signal peptide, it is reported to be inserted into the cell membrane with a short C-terminal cytoplasmic tail. The extracellular domain is estimated to be about 1 μm in length. It is found at muscle insertion zones and genetic evidence established that it serves to organize the inner layers of the cuticle. In addition, it is found in tracheae where it plays a similar role. Its most atypical feature is that the protein has 11.5% cysteine residues in its 22,971 amino acids. And these cysteines are restricted to two regions, for there is an interior domain of about 4,000 amino acids (beginning at amino acid 3950) with only 3 cysteines. The cysteines are found in various repeated motifs, some are unique for this protein. Thus there are 308 epidermal growth factor repeats, each with 6 conserved cysteines and many with putative calcium binding sites. These are interspersed with 185 copies of a novel 21 amino acid module with 4 cysteines that has been named “dumpy”. The cys-free region is populated by about 30 copies of a novel motif named “pigsfeast.” One such region is shown here: PIGS TGGQVTEQTTSSPSEVRTTIGLEESTLPSRSTDRTTPSESPETPTTLPSDFITRPHSDQT TESTRDVPT TRPFEAST. Immediately before the transmembrane domain is a zona pellucida (ZP) domain, conventionally found in proteins that surround an ovum.

Conserved homologs of “dumpy” have been identified in several insect orders (Carmon et al., 2007). The data are summarized in Table 2 although the complete sequence has not been annotated in any genus except *Drosophila*. This is not surprising given its length and that fact that the gene in *D. melanogaster* has 81 exons. I could only identify the pigsfeast region in five *Drosophila* species, although proteins with long stretches of EGF domains were present in many insects and even in Crustacea and Chelicerata. Those interested in more information are referred to the 178 references listed on FlyBase, especially Carmon et al. (2007) and Wilkin et al. (2000).

Several other ZP domain bearing proteins are known from *D. melanogaster* that are intimately involved with cuticle formation, possibly being incorporated directly into the cuticle (Roch et al. 2003; Bokel et al. 2005). These are miniature (m), dusky (dy), dusky-like (dyl), papillote (pot), and piopio (pio). Interestingly, a *C. elegans* cuticular protein (cut-1), expressed only in dauer larvae, also has a ZP domain (Sebastiano et al. 1991; Roch et al. 2003).

4.2. Insect Proteins with 3 R&R Consensus regions

There are rare exceptions to the generalization that only one R&R Consensus is found in any given insect CPR protein. For example, *Nasonia* appears to have two such proteins, NvitCPR42 and NvitCPR58 (sequences available in Supplementary Information File 8), but these proteins and their possible orthologs in other species have not been studied with care. There is, however, one *An. gambiae* protein (AgamCPR144) with three R&R Consensus regions; orthologs are known in *Drosophila*, *Aedes*, *Bombyx*, *Tribolium*, and even *Pediculus*. These are annotated as: DmelCpr73D (CG9665), *Aedes* (XP_001652170.1), BmorCPR47 (BR000548.1), *Tribolum* (XP_969947.2), and *Pediculus* (XM_002432771.1). An interesting feature found in all the orthologs is that the first Consensus region is atypical in having 9 amino acids rather than 8 between the first two glycines of the original Consensus. The second glycine is present as an aspartic acid (D) residue in *Bombyx*, and in the ESTs for two other lepidopterans that appear

to represent the start of this unusual protein (*Heliothis virescens*, GT207140.1 and *Heliconius erato*, DT668941.3). Thus the first Consensus region has this form: G-x(9)-[G/D]-x(6)-Y-T-A-G-x(2)-G-[F/Y]. Each of the three Consensus regions is most similar to its corresponding region in the other species. The first Consensus region of AgamCPR144 is given in Supplementary Information File 1.

Only the *An. gambiae* transcript has experimental evidence verifying its sequence and revealing that the three consensus regions are spread over 4 exons (Cornman et al., 2008). Attempts to localize the transcript for this gene in *An. gambiae* tissues have been unsuccessful, but there are *in situ* data for *D. melanogaster* embryos available at the Fly Express facility at FlyBase. Tiny dots of hybridization, in a pattern unmatched by any other gene, are seen in stage 13-16 embryos attributed to the ventral epidermis. So the function of this unusual and highly conserved gene remains unknown.

4.3. Resilin

Weis-Fogh (1960) recognized that there is a rubber-like type of cuticle in prealar arms, wing hinges and elastic tendons of several insects. He named the protein found in these regions – resilin. Extensive information on this fascinating protein can be found in Andersen and Weis-Fogh (1964).

Lombardi and Kaplan (1993) obtained and sequenced 17 peptides from the prealar arms of *Periplaneta americana*. The entire sample had 35% glycine, what would be expected for resilin. Eventually, Ardell and Andersen (2001) were able to obtain a complete sequence for resilin. They accomplished this by using the sequences of peptides from locust resilin to search the predicted genes of *D. melanogaster* and got a match to *CG15920*, now named *resilin*. Importantly, three of the peptides used to identify the gene contributed to the R&R Consensus (Fig.5). They also noted that three of the *Periplaneta* resilin peptides also matched the *Drosophila* sequence; two matched the Consensus (Fig. 5).

Dmelresilin has been confirmed (http://flybase.org/reports/FBgn0034157.html) to have two alternatively spliced forms, only one (PA) has the complete R&R Consensus; the other (PB) is truncated so that 46 amino acids of the Consensus are lacking (Fig. 5). The R&R Consensus, of the RR-2 form, from *D. melanogaster* is highly conserved, ten out of the 12 additional species of *Drosophila* whose whole genomes have been sequenced have a match of 100%, the other two (*D. virilis* and *D. mojavensis*) differ by one amino acid each. The entire amino acid sequence of Dmelresilin PA is present with from 87-97% identity in four other species of *Drosophila*.

Dmelresilin has been demonstrated to bind chitin via its R&R Consensus region (Qin et al., 2009) and the authors speculate that a combination of bound and unbound resilin proteins contribute to the elastic properties. Alternatively, it may be that the two forms of resilin are used in different structures, for Bailey and Weis-Fogh (1961) report that the main-wing hinge of locusts had a “pad of pure rubber-like protein” in addition to a lamellar component with both chitin and resilin.

The matches to the locust and cockroach peptides are not the only criteria for classifying this *D. melanogaster* sequence as resilin. Elvin and his co-workers made a synthetic peptide coded by most of the first exon except for the signal peptide but including 17 copies of an imperfect repeat region (Elvin et al., 2005). They cross-linked this construct appropriately, did the proper physical measurements, and established that this protein had precisely the properties one would want in resilin. They made an artificial protein with 16 copies of a consensus of the repeat in the first exon (GGRPDSYGAPGGGN), cross-linked it, and it too had elastic properties. Qin

et al. (2009) also carried out various physical measurements with the complete *D. melanogaster* protein and confirmed its elastomeric properties.

What was presumed to be a homolog of *resilin* was identified in *Anopheles*, based on the EST BX619161 (Lyons et al., 2007). They did not comment that the corresponding gene (AGAP002367) lacked the R&R Consensus, the precise region that had been used to identify resilin in *D. melanogaster*. AGAP002367 also has repeats (AQTPSSQYGAP), and the artificial peptide created from 16 of them behaves just like resilin should (Lyons et al., 2007). Not only did the AGAP002367 sequence lack the R&R Consensus, but we had identified an *An.gambiae* gene (*AgamCPR152 = AGAP012487*) that appeared to be the ortholog of the *D. melanogaster resilin* gene based on its R&R Consensus having 74% identity (Cornman et al. 2008). AgamCPR152 lacks anything resembling the repeats that underlie the elastic properties of either Dmelresilin or AGAP002367. Furthermore, *D. melanogaster* has a different gene, *CG7709*, that appears to be the ortholog of *AGAP002367*, and this gene has been identified as a mucin, indeed its official name is *Muc91c* (Supplementary Information File 5).

So, now we have two *D. melanogaster* genes, one codes for a sequence that matched locust and cockroach resilin based on its R&R Consensus region and the other matches the *An. gambiae* gene that Elvin and co-workers identified as resilin-like based on repeats that resemble slightly those in Dmelresilin and confer elastic properties.

Further evidence that AgamCPR152 was correctly identified as an *An. gambiae* resilin homolog is that the protein in *Nasonia* with the best match to the Consensus of Dmelresilin and AgamCPR152 is NvitCPR25 (XP_001604687.1). This protein has some repeats that are similar to those in Dmelresilin (Supplementary Information File 5).

Another feature of resilins is their high proportion of glycine residues from 35-40% (Ardell and Andersen, 2001). Data shown in Fig. 5 indicate that Dmelresilin and NvitCPR25 and an *Apis* protein AmelCPR15 (~ XP_392701.2) fall within the correct range, the other proteins discussed above are far lower.

So, how should an ortholog be defined and more importantly what is the function of these proteins?

4.4. BcNCP1 and its orthologs

An atypical protein was isolated from cuticle of the cockroach, *Blaberus craniifer* by Jensen et al. (1997). It is unusual because it has 6 cysteine residues occurring as identically spaced pairs separated by five amino acids in three almost identical motifs of 16 amino acids (Figure 4). BcNCP1 has unambiguous orthologs in numerous other insect species of several orders and in Crustacea with most showing the same three instances of similar motifs (Table 2, Supplementary Information File 6). It would be of interest to learn more about how this atypical protein is used. Several clues are available. BCNCP1 was isolated from post-ecdysial nymphal abdominal cuticle; the *D. melanogaster* ortholog is expressed in larval tracheae (FlyAtlas) and in stage 13-16 embryos in the head dorsal epidermis and atrium. The *B. mori* ortholog, *CPH1*, is expressed in epidermis, wing discs and compound eyes (Futahashi et al. 2008).

5. CP sequence variation in natural populations

Ever since the first cuticle protein sequences became available concerns have been raised about whether two highly similar sequences represented two different genes or allelic variations of a single gene. When protein and RNA preparations came from pooled animals, there was no way of distinguishing between alleles and similar genes.

Two pre-genomic studies, however, had evidence for allelic variation in sequences obtained from lobsters (*Homarus americanus*) and mealworms (*Tenebrio molitor*). In the three papers that came from direct protein sequencing of cuticle from various regions of individual lobsters, Andersen and his co-workers obtained complete sequences for 23 CPs, and found 5 pairs of very similar sequences. There were no instances of more than two similar sequences, hence the data are consistent with the slight variations in these lobster proteins representing different alleles of a single gene (Krag et al., 1997; Andersen, 1998; Nousiainen et al., 1998). Further evidence for allelic variation came from a study in *Tenebrio* designed specifically to address this issue by analyzing CPs from individual animals (Haebel et al., 1995). There were three variants of protein TMl-F1 (a,b,c) and in an individual animal at most two variants were found. The analysis began with spots from 2D gels and when two variants were present, the spots were approximately equal in intensity. Hence, this is strong evidence for allelic variation for this gene.

On the other hand, a genomic clone isolated from *An. gambiae* had three closely related genes. Identity, at the amino acid level, between pairs ranged from 96-98% (Dotson et al., 1998). Here, similarity was clearly non-allelic.

One of the first lessons learned from whole genome sequences, where one could examine far longer stretches of DNA than were found in single cloned regions, was that a genome might have multiple genes with very similar sequences. Indeed as discussed above (Section 2.1), sequence clusters of RR-2 genes in *An. gambiae* and *B. mori* could have as many as 16 genes sharing considerable identity throughout their sequences and especially in the R&R Consensus region. An extreme case of almost identical genes occurs with the entire CPLCW family in *An. gambiae* discussed above (Section 2.6). Furthermore, even in species where sequence clusters have not been found, pairs of genes coding for almost identical proteins are common. Thus there is abundant evidence for distinct but similar genes.

There are two studies that looked at allelic variation in CPs in natural populations. The only study that sampled variation in animals sequenced directly from a natural population provided conclusive evidence for allelic variation in two *An. gambiae* proteins (White et al., 2007). A portion of chromosome 2L in *An. gambiae* exists in two forms, one 2La, appears to have come from *An. arabiensis*. It was first recognized because it forms a distinct inversion visible in chromosome preparations when paired with the normal region (2La+). In a study designed to probe differences between the two chromosomal forms, White et al. (2007) sequenced multiple copies of several genes in animals homozygous for one or the other of the two chromosomal forms. Two of the sequenced genes were *AgamCPR34* and *AgamCPR63*. The sequenced regions encompass the R&R Consensus and some flanking sequence. Both genes are classified as RR-2, the form with the most conserved Consensus region. Hence the data allow one to compare differences between the Consensus and its flanking sequences as well as to obtain information on variation in natural populations. Care was taken to design gene specific primers; the products were sequenced in both directions, and the sequences were reported to be error-free. The mosquitoes sequenced were collected in the field in Tibati, Cameroon. The data (shown in Supplementary Information File 7) reveal the variation in genomic sequence in a natural population. For each mosquito, there is the possibility of just two alleles for each gene. There was far more amino acid variation in AgamCPR34 than in AgamCPR63. Ten of 27 animals sequenced for *CPR34* were heterozygous for this gene, and about half (7/13 sites) of the heterozygosity was at sites where there was clearly polymorphism in the population for the homozygotes had different amino acids at those sites. Polymorphism was far lower for *AgamCPR63*, only two sites had a homozygous variant form, but there were more heterozygous individuals (16/32) and 12 of their alternative forms were at the variant sites. For both genes, variation was found within the R&R Consensus region.

These data revealed that there can be considerable variation within a natural population even for the Consensus region.

A second study examined genetic differences in 21 isogenic lines of *D. melanogaster* with 15 source populations coming from Africa, two from Asia, one from France, and 3 from the U.S. (Shapiro et al., 2007). The possibility of sequencing errors was acknowledged but was reported to be rare, but many uncalled residues (n) are seen in the sequences. Among the genes surveyed were two CPR genes, *DmelCpr65Az* (RR-1) and *DmelCcp84Ae* (RR-2). The genomic region of *Cpr65Az* that was sequenced spanned an intron that was spliced out for these comparisons. The coding region of *Cpr65Az* sequenced (128 amino acids) had two lines that each differed in two amino acids from the rest, neither was in the Consensus region (Supplementary Information File 7). The region sequenced for *Ccp84Ae* was entirely in an exon and here the only amino acid differences were in what were clearly repetitive stretches. The greater amount of genetic diversity found in the *Anopheles* study is intriguing.

6. Conclusions

This review has summarized what we know about the diversity of sequences found in arthropod cuticles. What is striking is how many families of proteins are found in multiple insect orders and even in other arthropods, but not beyond. Where did these proteins come from? What is special about their properties that accounts for their not being found in other instances where proteins and chitin interact to form protective surfaces? As more investigators turn to proteomic analyses, more forms of cuticular proteins are likely to appear. Will these be group specific like the apidermins and CPLCW families, or reveal new widely distributed groups?

The real need for the future is to go beyond identifying sequences and to begin learning precisely how they function. Correlations between various physiological functions or morphological forms have been found when individual CP genes are silenced, even when the specific protein being studied belonged to a sizeable family, i.e. TwdlD (Guan et al., 2006) and Gasp (Barry et al., 1999; Behr and Hoch, 2005). Hence it is reasonable to expect that RNAi analyses will provide correlations between individual sequences and precise functions. A major gap in our knowledge is how the proteins interact within the cuticle to give rise to its exquisite layered structure. We need to learn which proteins are actually bound to chitin and whether that interaction is by simple hydrogen bonds or more stable linkages. What proteins are involved in cross-linking and what residues are involved? Where and how do the non-chitin binding proteins fit in? While this review should certainly serve to help others to correctly classify protein sequences, it is hoped that it will serve to guide studies that go beyond protein identification and stimulate further investigation into precisely how these cuticular proteins contribute to the form and function of the arthropods who devote so many of their genes to their production.

Recently genes for putative CPs have been identified as major players in physiological phenomena as diverse as insecticide resistance (Vontas et al., 2007; Awolola et al., 2008; Zhang et al., 2008), drought tolerance (Zhang et al., 2008), resistance to heavy metals (Roelofs et al., 2009; Shaw et al., 2007), and even sibling species differentiation (Cassone et al., 2008). These studies were beyond the scope of this review but provide further evidence about the importance of CPs and the need to learn more about their structures and functions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

R. Scott Cornman, Toru Togawa and Hitoshi Tsujimoto provided valuable suggestions for this review. Subbaratnam Murthukrishnan aided my understanding of the CPAP1 and CPAP3 families. I am especially grateful to three anonymous referees who provided important insights, challenged assumptions, and recommended more details. Work on this review was supported by a grant from the National Institutes of Health (AI55624).

References

Andersen SO. Characterization of proteins from arthrodial membranes of the lobster, *Homarus americanus*. Comp Biochem Physiol A Mol Integr Physiol 1998;121:375–383. [PubMed: 10048190]

Andersen SO. Studies on proteins in post-ecdysial nymphal cuticle of locust, *Locusta migratoria*, and cockroach, *Blaberus craniifer*. Insect Biochem Mol Biol 2000;30:569–577. [PubMed: 10844249]

Andersen, SO. Cuticular sclerotization and tanning. In: Gilbert, LI.; Iatrou, K.; Gill, SS., editors. Comprehensive molecular insect science. Vol. 4. Elsevier B.V., Amsterdam; New York: 2005. p. 145-170.

Andersen SO, Hojrup P, Roepstorff P. Insect cuticular proteins. Insect Biochem Mol Biol 1995;25:153–176. [PubMed: 7711748]

Andersen SO, Rafn K, Roepstorff P. Sequence studies of proteins from larval and pupal cuticle of the yellow meal worm, *Tenebrio molitor*. Insect Biochem Mol Biol 1997;27:121–131. [PubMed: 9066122]

Andersen SO, Weis-Fogh T. Resilin. A rubber-like protein in arthropod cuticle. Adv Insect Physiol 1964;2:1–65.

Arakane Y, Begum K, Dixit R, Park Y, Specht CA, Merzendorfer H, Kramer KJ, Muthukrishnan S, Beeman RW. Analysis of functions of the chitin deacetylase gene family in *Tribolium castaneum*. Insect Biochem Mol Biol. 2009

Ardell DH, Andersen SO. Tentative identification of a resilin gene in *Drosophila melanogaster*. Insect Biochem Mol Biol 2001;31:965–970. [PubMed: 11483432]

Awolola TS, Oduola OA, Strode C, Koekemoer LL, Brooke B, Ranson H. Evidence of multiple pyrethroid resistance mechanisms in the malaria vector *Anopheles gambiae sensu stricto* from. Nigeria Trans R Soc Trop Med Hyg 2009;103:1139–1145.

Bailey K, Weis-Fogh T. Amino acid composition of a new rubber-like protein, resilin. Biochimica et. Biophysica Acta 1961;48:452–459.

Barry MK, Triplett AA, Christensen AC. A peritrophin-like protein expressed in the embryonic tracheae of *Drosophila melanogaster*. Insect Biochem Mol Biol 1999;29:319–327. [PubMed: 10333571]

Behr M, Hoch M. Identification of the novel evolutionary conserved *obstructor* multigene family in invertebrates. FEBS Lett 2005;579:6827–6833. [PubMed: 16325182]

Bokel C, Prokop A, Brown NH. Papillote and Piopio: *Drosophila* ZP-domain proteins required for cell adhesion to the apical extracellular matrix and microtubule organization. J Cell Sci 2005;118:633–642. [PubMed: 15657084]

Bouhin H, Charles JP, Quennedey B, Delachambre J. Developmental profiles of epidermal mRNAs during the pupal-adult molt of *Tenebrio molitor* and isolation of a cDNA clone encoding an adult cuticular protein: effects of a juvenile hormone analogue. Dev Biol 1992;149:112–122. [PubMed: 1728581]

Carmon A, Wilkin M, Hassan J, Baron M, MacIntyre R. Concerted evolution within the *Drosophila dumpy* gene. Genetics 2007;176:309–325. [PubMed: 17237523]

Cassone BJ, Mouline K, Hahn MW, White BJ, Pombi M, Simard F, Costantini C, Besansky NJ. Differential gene expression in incipient species of *Anopheles gambiae*. Mol Ecol 2008;17:2491–2504. [PubMed: 18430144]

Charles JP, Bouhin H, Quennedey B, Courrent A, Delachambre J. cDNA cloning and deduced amino acid sequence of a major, glycine-rich cuticular protein from the coleopteran *Tenebrio molitor*. Temporal and spatial distribution of the transcript during metamorphosis. Eur J Biochem 1992;206:813–819. [PubMed: 1606964]

Cornman RS. Molecular evolution of *Drosophila* cuticular protein genes. PLoS One 2009;4(12):e8345. [PubMed: 20019874]

Cornman RS, Togawa T, Dunn WA, He N, Emmons AC, Willis JH. Annotation and analysis of a large cuticular protein family with the R&R Consensus in *Anopheles gambiae*. BMC Genomics 2008;9:22. [PubMed: 18205929]

Cornman RS, Willis JH. Extensive gene amplification and concerted evolution within the CPR family of cuticular proteins in mosquitoes. Insect Biochem Mol Biol 2008;38:661–676. [PubMed: 18510978]

Cornman RS, Willis JH. Annotation and analysis of low-complexity protein families of *Anopheles gambiae* that are associated with cuticle. Insect Mol Biol 2009;18:607–622. [PubMed: 19754739]

Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: A sequence logo generator. Genome Research 2004;14:1188–1190. [PubMed: 15173120]

Dotson EM, Cornel AJ, Willis JH, Collins FH. A family of pupal-specific cuticular protein genes in the mosquito *Anopheles gambiae*. Insect Biochem Mol Biol 1998;28:459–472. [PubMed: 9718679]

Edgecombe GD. Palaeontological and molecular evidence linking arthropods, onychophorans, and other Ecdysoza. Evolution: Education and Outreach 2009;2:178–190.

Elvin CM, Carr AG, Huson MG, Maxwell JM, Pearson RD, Vuocolo T, Liyou NE, Wong DC, Merritt DJ, Dixon NE. Synthesis and properties of crosslinked recombinant pro-resilin. Nature 2005;437:999–1002. [PubMed: 16222249]

Futahashi R, Fujiwara H. Identification of stage-specific larval camouflage associated genes in the swallowtail butterfly, *Papilio xuthu*s. Dev Genes Evol 2008;218:491–504. [PubMed: 18712529]

Futahashi R, Okamoto S, Kawasaki H, Zhong YS, Iwanaga M, Mita K, Fujiwara H. Genome-wide identification of cuticular protein genes in the silkworm, *Bombyx mori*. Insect Biochem Mol Biol 2008;38:1138–1146. [PubMed: 19280704]

Guan X, Middlebrooks BW, Alexander S, Wasserman SA. Mutation of TweedleD, a member of an unconventional cuticle protein family, alters body shape in *Drosophila*. Proc Natl Acad Sci U S A 2006;103:16794–16799. [PubMed: 17075064]

Haebel S, Jensen C, Andersen SO, Roepstorff P. Isoforms of a cuticular protein from larvae of the meal beetle, *Tenebrio molito*r, studied by mass spectrometry in combination with Edman degradation and two-dimensional polyacrylamide gel electrophoresis. Protein Sci 1995;4:394–404. [PubMed: 7795523]

Hamodrakas SJ, Willis JH, Iconomidou VA. A structural model of the chitin-binding domain of cuticle proteins. Insect Biochem Mol Biol 2002;32:1577–1583. [PubMed: 12530225]

He N, Botelho JM, McNall RJ, Belozerov V, Dunn WA, Mize T, Orlando R, Willis JH. Proteomic analysis of cast cuticles from *Anopheles gambiae* by tandem mass spectrometry. Insect Biochem Mol Biol 2007;37:135–146. [PubMed: 17244542]

Honeybee Genome Sequencing Consortium. Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature 2006;443:931–949. [PubMed: 17073008]

Iconomidou VA, Willis JH, Hamodrakas SJ. Unique features of the structural model of ‘hard’ cuticle proteins: implications for chitin-protein interactions and cross-linking in cuticle. Insect Biochem Mol Biol 2005;35:553–560. [PubMed: 15857761]

Iijima M, Hashimoto T, Matsuda Y, Nagai T, Yamano Y, Ichi T, Osaki T, Kawabata S. Comprehensive sequence analysis of horseshoe crab cuticular proteins and their involvement in transglutaminase-dependent cross-linking. FEBS J 2005;272:4774–4786. [PubMed: 16156796]

Ikeya T, Persson P, Kono M, Watanabe T. The DD5 gene of the decapod crustacean *Penaeus japonicus* encodes a putative exoskeletal protein with a novel tandem repeat structure. Comp Biochem Physiol B Biochem Mol Biol 2001;128:379–388. [PubMed: 11250533]

Jasrapuria S, Arakane Y, Osman G, Kramer KJ, Beeman RW, Muthukrishnan S. Genes encoding proteins with peritrophin A-type chitin binding domains in *Tribolium castaneum* are grouped into three distinct families based on phylogeny, expression and function. Insect Biochem Molec Biol. 2010 this issue.

Jensen UG, Rothmann A, Skou L, Andersen SO, Roepstorff P, Hojrup P. Cuticular proteins from the giant cockroach, *Blaberus craniifer*. Insect Biochem Mol Biol 1997;27:109–120. [PubMed: 9066121]

Johnston WL, Krizus A, Dennis JW. The eggshell is required for meiotic fidelity, polar-body extrusion and polarization of the *C. elegans* embryo. BMC Biology 2006;4:35. [PubMed: 17042944]

Kim E, Choi Y, Lee S, Seo Y, Yoon J, Baek K. Characterization of the *Drosophila melanogaster retinin* gene encoding a cornea-specific protein. Insect Mol Biol 2008;17:537–543. [PubMed: 18828839]

Klein SL, Strausberg RL, Wagner L, Pontius J, Clifton SW, Richardson P. Genetic and genomic tools for *Xenopus* research: The NIH *Xenopus* initiative. Dev Dyn 2002;225:384–391. [PubMed: 12454917]

Kragh M, Molbak L, Andersen SO. Cuticular proteins from the lobster, *Homarus ame*ricanus. Comp Biochem Physiol B Biochem Mol Biol 1997;118:147–154. [PubMed: 9418004]

Kucharski R, Maleszka J, Maleszka R. Novel cuticular proteins revealed by the honey bee genome. Insect Biochem Mol Biol 2007;37:128–134. [PubMed: 17244541]

Lombardi EC, Kaplan DL. Preliminary characterization of resilin isolated from the cockroach, *Periplaneta americana*. Materials Research Society Symposium Proceedings 1993;292:3–7.

Lyons RE, Lesieur E, Kim M, Wong DC, Huson MG, Nairn KM, Brownlee AG, Pearson RD, Elvin CM. Design and facile production of recombinant resilin-like polypeptides: gene construction and a rapid protein purification method. Protein Eng Des Sel 2007;20:25–32. [PubMed: 17218334]

Nakato H, Toriyama M, Izumi S, Tomino S. Structural and expression of mRNA for a pupal cuticle protein of the silkworm, *Bombyx mori*. Insect Biochem 1990;20:667–678.

Nasonia Genome Working Group. Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. Science 2010;327:343–348. [PubMed: 20075255]

Nousiainen M, Rafn K, Skou L, Roepstorff P, Andersen SO. Characterization of exoskeletal proteins from the American lobster, *Homarus americanus*. Comp Biochem Physiol B Biochem Mol Biol 1998;119:189–199. [PubMed: 9530820]

Okamoto S, Futahashi R, Kojima T, Mita K, Fujiwara H. Catalogue of epidermal genes: genes expressed in the epidermis during larval molt of the silkworm *Bombyx mori*. BMC Genomics 2008;9:396. [PubMed: 18721459]

Qin G, Lapidot S, Numata K, Hu X, Meirovitch S, Dekel M, Podoler I, Shoseyov O, Kaplan DL. Expression, cross-linking, and characterization of recombinant chitin binding resilin. Biomacromolecules 2009;10:3227–3234. [PubMed: 19928816]

Qui J, Hardin PE. Temporal and spatial expression of an adult cuticle protein gene from *Drosophila* suggests that its protein product may impart some specialized cuticle function. Dev Biol 1995;167:416–425. [PubMed: 7875368]

Rebers JE, Riddiford LM. Structure and expression of a *Manduca sexta* larval cuticle gene homologous to *Drosophila* cuticle genes. J Mol Biol 1988;203:411–423. [PubMed: 2462055]

Rebers JE, Willis JH. A conserved domain in arthropod cuticular proteins binds chitin. Insect Biochem Mol Biol 2001;31:1083–1093. [PubMed: 11520687]

Roch F, Alonso CR, Akam M. *Drosophila miniature* and *dusky* encode ZP proteins required for cytoskeletal reorganisation during wing morphogenesis. J Cell Science 2003;116:1199–1207. [PubMed: 12615963]

Roeding F, Borner J, Kube M, Klages S, Reinhardt R, Burmester T. A 454 sequencing approach for large scale phylogenomic analysis of the common emperor scorpion (*Pandinus imperator*). Mol Phylogenet Evol 2009;53:826–834. [PubMed: 19695333]

Roeding F, Hagner-Holler S, Ruhberg H, Ebersberger I, von Haeseler A, Kube M, Reinhardt R, Burmester T. EST sequencing of *Onychophora* and phylogenomic analysis of Metazoa. Mol Phylogenet Evol 2007;45:942–951. [PubMed: 17933557]

Roelofs D, Janssens TK, Timmermans MJ, Nota B, Marien J, Bochdanovits Z, Ylstra B, Van Straalen NM. Adaptive differences in gene expression associated with heavy metal tolerance in the soil arthropod *Orchesella cincta*. Mol Ecol 2009;18:3227–3239. [PubMed: 19566677]

Schmidt-Rhaesa A, Bartolomaeus T, Lemburg C, Ehlers U, Garey JR. The position of the Arthropoda in the phylogenetic system. J Morphol 1998;238:263–285.

Schneider TD, Stephens RM. Sequence Logos: A New Way to Display Consensus Sequences. Nucleic Acids Res 1990;18:6097–6100. [PubMed: 2172928]

Sebastiano M, Lassandro F, Bazzicalupo P. *cut-1*, a *Caenorhabditis elegans* gene coding for a dauer-specific noncollagenous component of the cuticle. Develop Biol 1991;146:519–530. [PubMed: 1864469]

Shapiro JA, Huang W, Zhang C, Hubisz MJ, Lu J, Turissini DA, Fang S, Wang HY, Hudson RR, Nielsen R, Chen Z, Wu CI. Adaptive genic evolution in the *Drosophila* genomes. Proc Natl Acad Sci U S A 2007;104:2271–2276. [PubMed: 17284599]

Shaw JR, Colbourne JK, Davey JC, Glaholt SP, Hampton TH, Chen CY, Folt CL, Hamilton JW. Gene response profiles for *Daphnia pulex* exposed to the environmental stressor cadmium reveals novel crustacean metallothioneins. BMC Genomics 2007;8:477. [PubMed: 18154678]

Tang L, Liang J, Zhan Z, Xiang Z, He N. Identification of the chitin binding proteins from the larval proteins of silkworm, *Bombyx mori*. Insect Biochem Molec Biol. 2010 this issue.

Tellam RL, Wijffels G, Willadsen P. Peritrophic matrix proteins. Insect Biochem Mol Biol 1999;29:87–101. [PubMed: 10196732]

Togawa T, Dunn WA, Emmons AC, Willis JH. CPF and CPFL, two related gene families encoding cuticular proteins of *Anopheles gambiae* and other insects. Insect Biochem Mol Biol 2007;37:675–688. [PubMed: 17550824]

Togawa T, Dunn WA, Emmons AC, Nagao J, Willis JH. Developmental expression patterns of cuticular protein genes with the R&R Consensus from *Anopheles gambiae*. Insect Biochem Mol Biol 2008;38:508–519. [PubMed: 18405829]

Togawa T, Nakato H, Izumi S. Analysis of the chitin recognition mechanism of cuticle proteins from the soft cuticle of the silkworm, *Bombyx mori*. Insect Biochem Mol Biol 2004;34:1059–1067. [PubMed: 15475300]

Vontas J, David JP, Nikou D, Hemingway J, Christophides GK, Louis C, Ranson H. Transcriptional analysis of insecticide resistance in *Anopheles stephensi* using cross-species microarray hybridization. Insect Mol Biol 2007;16:315–324. [PubMed: 17433071]

Weis-Fogh T. A rubber-like protein in insect cuticle. Journal of Experimental Biology 1960;37:889–907.

White BJ, Hahn MW, Pombi M, Cassone BJ, Lobo NF, Simard F, Besansky NJ. Localization of candidate regions maintaining a common polymorphic inversion (2La) in *Anopheles gambiae*. PLoS Genet 2007;3:e217. [PubMed: 18069896]

Wilkin MB, Becker MN, Mulvey D, Phan I, Chao A, Cooper K, Chung HJ, Campbell ID, Baron M, MacIntyre R. *Drosophila dumpy* is a gigantic extracellular protein required to maintain tension at epidermal-cuticle attachment sites. Curr Biol 2000;10:559–567. [PubMed: 10837220]

Willis, JH.; Iconomidou, VA.; Smith, RF.; Hamodrakas, SJ. Cuticular proteins. In: Gilbert, LI.; Iatrou, K.; Gill, SS., editors. Comprehensive molecular insect science. Vol. 4. Elsevier B.V., Amsterdam; New York: 2005. p. 79-109.

Wootton JC, Federhen S. Statistics of Local Complexity in Amino-Acid-Sequences and Sequence Databases. Comput Chem 1993;17:149–163.

Zhang J, Goyer C, Pelletier Y. Environmental stresses induce the expression of putative glycine-rich insect cuticular protein genes in adult *Leptinotarsa decemlineata* (Say). Insect Mol Biol 2008;17:209–216. [PubMed: 18477239]

Zhong YS, Mita K, Shimada T, Kawasaki H. Glycine-rich protein genes, which encode a major component of the cuticle, have different developmental profiles from other cuticle protein genes in *Bombyx mori*. Insect Biochem Mol Biol 2006;36:99–110. [PubMed: 16431278]

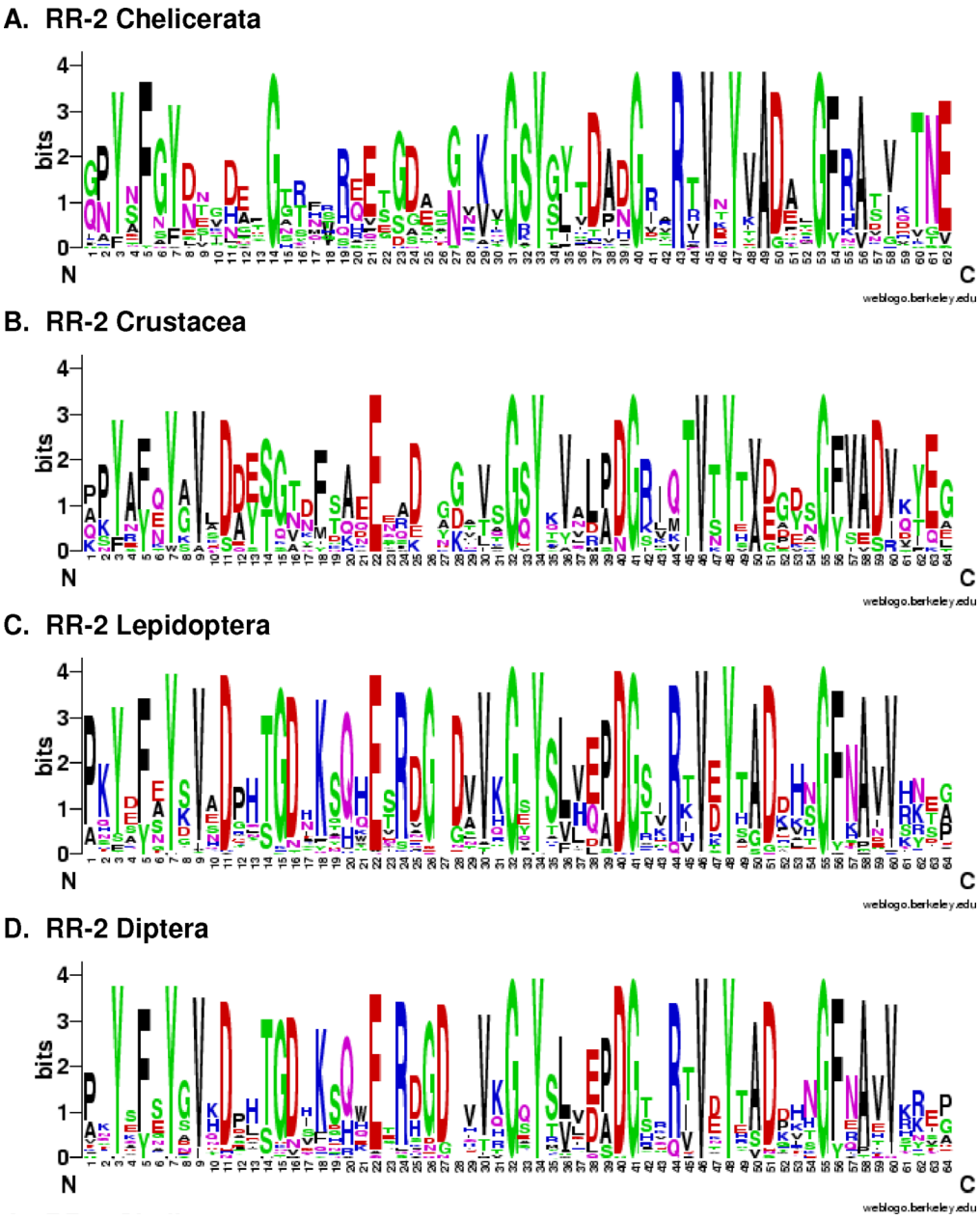

Fig.1.

Comparisons of RR-2 Consensus regions from 4 different groups of arthropods. WebLogos were constructed at <http://weblogo.berkeley.edu/logo.cgi> (Schneider and Stephens, 1990; Crooks et al., 2004). Details on sequences used are in Supplementary Information File 8. A. Chelicerata based on 33 sequences. B. Crustacea based on 16 sequences. C. Lepidoptera based on 87 *B. mori* sequences. D. Diptera based on 101 *An. gambiae* sequences.

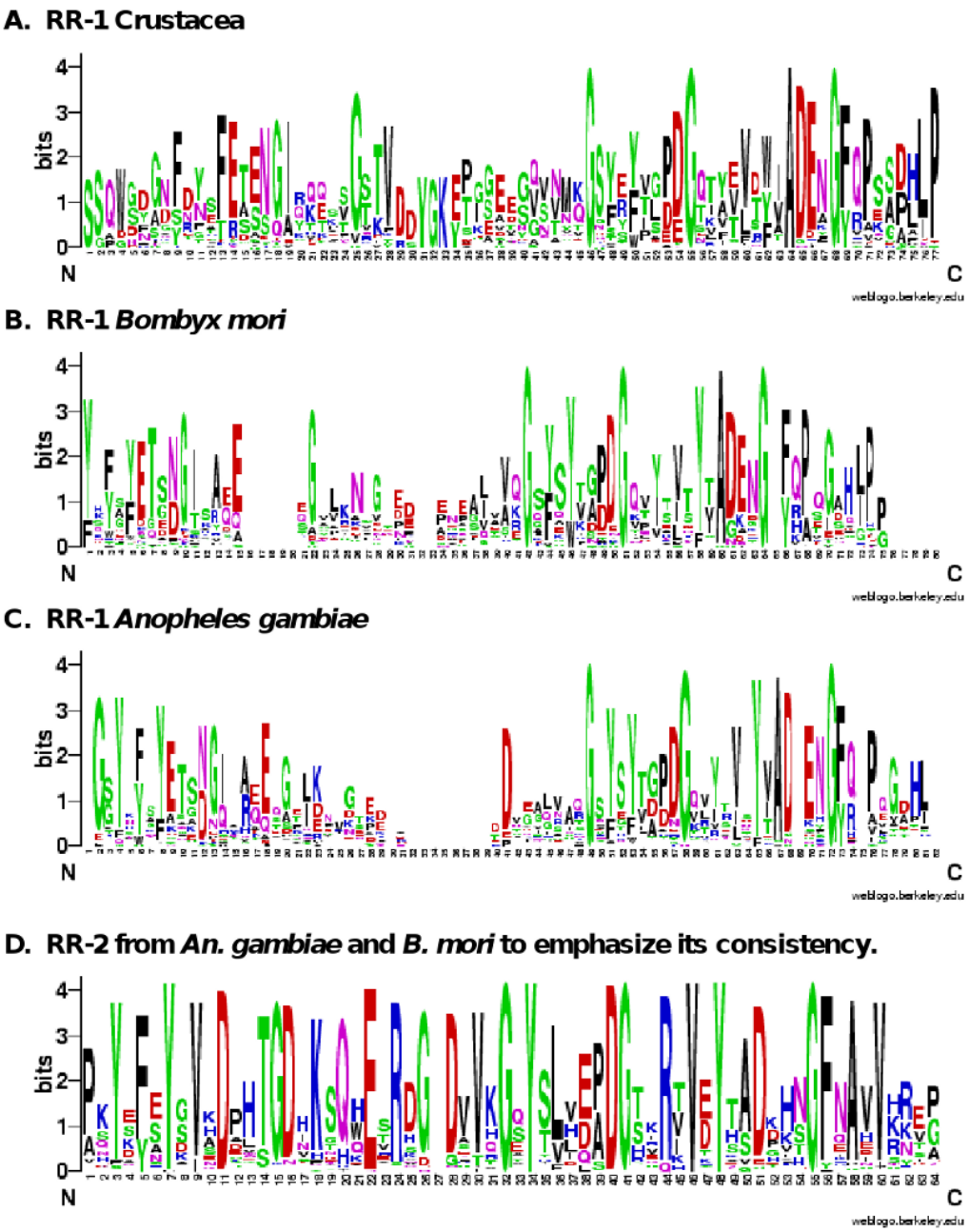

Fig. 2.

WebLogos (see Fig. 1) of RR-1 Consensus regions compared to RR-2. Details on sequences used are in Supplementary Information File 8. A. Crustacea based on 48 sequences. B. Representative lepidopteran, *Bombyx mori*, 52 sequences. C. Representative dipteran, *Anopheles gambiae*, 51 sequences. D. RR-2 sequences from combination of data used for panels C and D of Fig. 1.

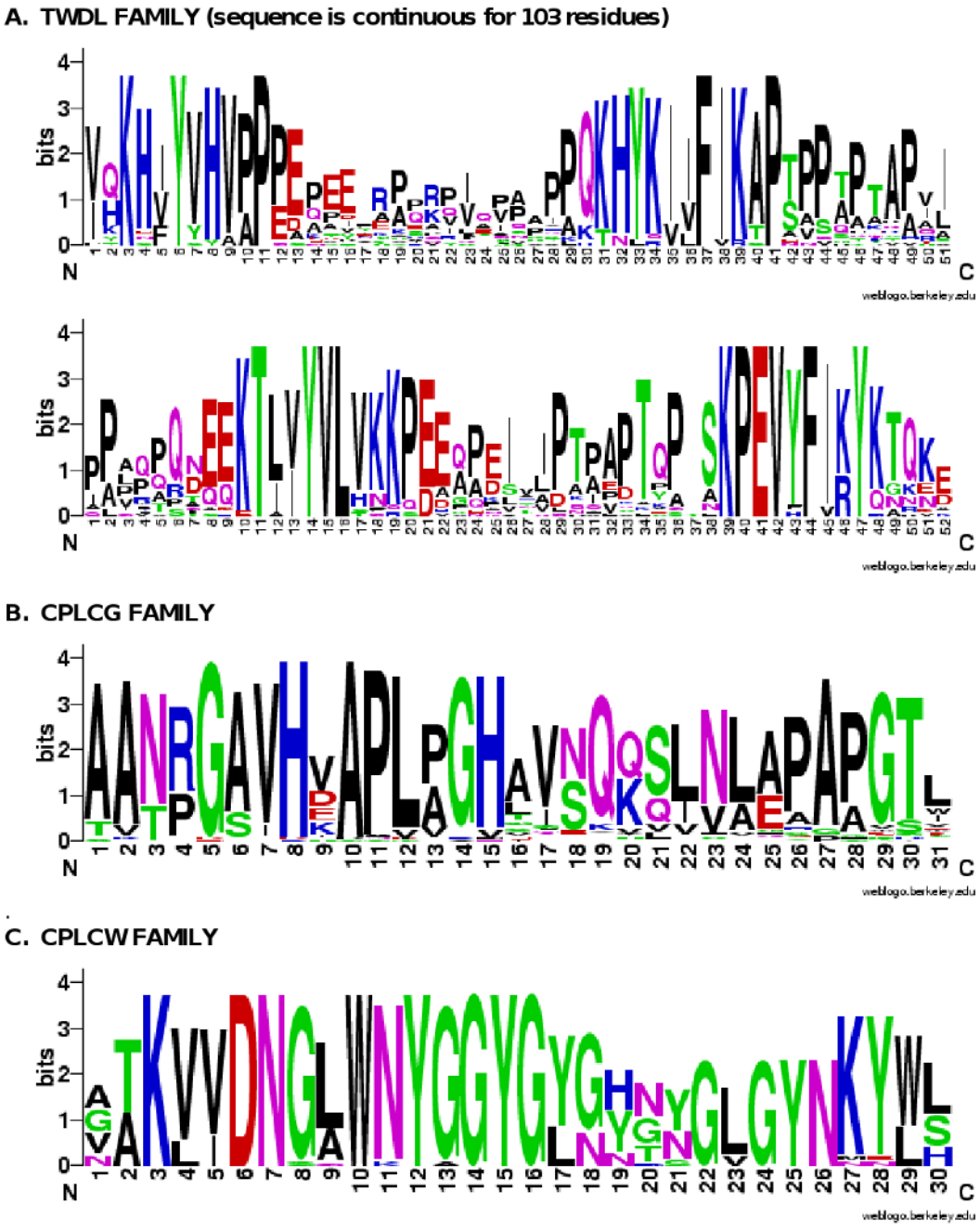

Fig. 3.

WebLogos (see Fig. 1) for three cuticular protein families. Details on sequences used are in Supplementary File 8. A. TWDL Family. Twenty-four sequences from 8 species in 6 orders of insects were used to create this figure. The continuous sequence was split to facilitate recognition of the four conserved regions. B. CPLCG Family. Note the highly conserved GHPG at residues 5, 8,11,14. Eighty-six sequences from dipterans were used. This is a shortened version of the WebLogo that appears in Cornman and Willis (2009). See Supplementary Information File 2 for WebLogo of CPLCGs from non-dipterans. C. CPLCW Family. The 27 CPLCW sequences of this mosquito-restricted family were used. Unlike other WebLogos, the alignment for this one required gaps of 5, 6, or 8 amino acids between position 18 and 27 to

accommodate the longer *Ae. aegypti* sequences. This is a shortened version of the WebLogo that appears in Cornman and Willis (2009).

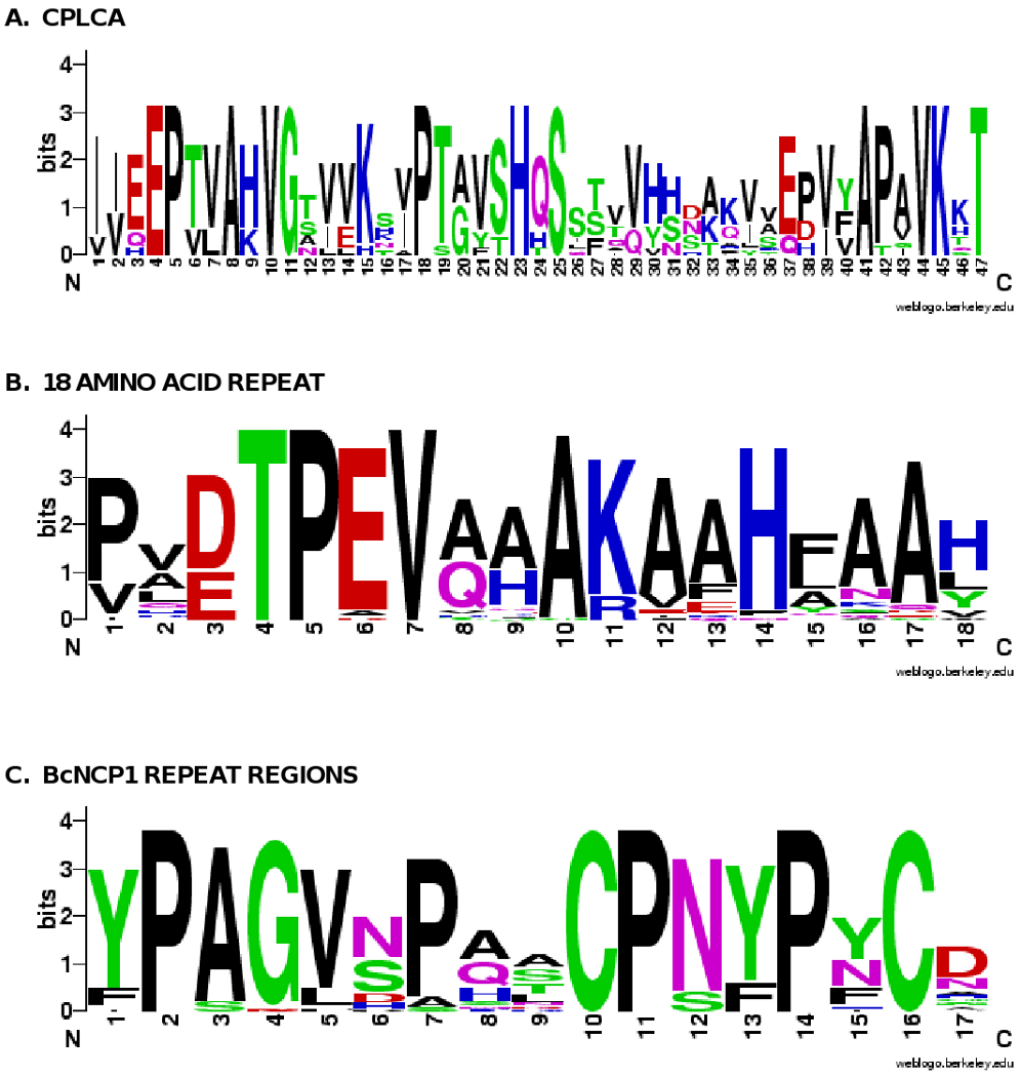

Fig. 4.

WebLogos (see Fig. 1). See Supplementary Information File 8 for details on sequences used. A. CPLCA Family. The WebLogo is based on three sequences from each of four species, *An. gambiae*, *Ae. agegypi*, *C. pipiens* and *D. melanogaster* that had the closest match to AgamCPLCA1. This region corresponds to the retinin domain. This is a shortened version of the WebLogo that appears in Cornman and Willis (2009). B. The 18 amino acid repeat from 40 sequences from 26 proteins from 5 insect orders and two crustaceans. C. BcNCP1, 29 repeat regions from orthologs in 11 species in 8 insect orders and Crustacea. Only one species of *Drosophila* was used.

A. Alignment of Relevant "Resilin" Peptides

```
1)   YEFSYEVSDAETGTEFG      DGDVAQGSYNVLL      QVVDYEADEG
2) AKYEFNYQVEDAPSGLSFGHSEMRDGDFTTGQYNVLLPDGRKQIVEYEADQQGYRPQIRYEG
3)    EFSYDVNDASTGTEF     DGDVAQGSY
4)                                              DIVEYEADQQGYRPQIRYEG
5) AKYEFSYEVDEEHTDLSFGHEEMRDGDYTTGKYNVLLPDGRRQIVEYEADHKGYRFKITYEG
```

1) *Locusta* resilin peptides
2) Stretch of Dmelresilin PA with critical residues of R&R Consensus highlighted
3) *Periplaneta* resilin peptides
4) Only region from Dmelresilin PB that matches R&R Consensus
5) AgamCPR152 with amino acids identical to Dmelresilin highlighted – 74% identity

B. Percent glycine in various proteins that have resilin-like properties. Proteins with the R&R Consensus are shown in bold.

Dmelresilin PA	35%
AmelCPR15	34%
NvitCPR25	41%
AgamCPR152	12%
AGAP002367	19%
DmelMuc91c	10%

Fig. 5.

Comparison of sequences discussed in relation to resilin. References are in the text. A. Alignment of peptides derived from various proteins as indicated on the figure. B. Comparison of glycine content in various resilin and resilin-like proteins. The first four have the R&R Consensus, the last two lack that region. Additional identifiers for these proteins are: Dmelresilin (CG15920) and AgamCPR152 (AGAP012487-PA**).** Manual annotation modified the sequences for both AmelCPR15 and NvitCPR25. The modified sequences are given in Supplementary Material 8.

TABLE 1

Number of genes in different CP families in species with manual annotation of CPs in whole genome data

	CPR	CPF + CPFL	TWDL	CPLCA	CPLCG	CPLCW	CPLCP	GLY-RICH	APIDERMIN	CPAP3 (OBSTRUCTOR)	CPAP1	OTHER	TOTAL
section of paper	*2.1*	*2.2*	*2.3*	*2.4*	*2.5*	*2.6*	*2.7*	*2.8*	*2.9*	*2.10*	*2.10*	*2.11*	
An. gambiae	156	11	12	3	27	9	4 +23?	0		7		11	240+
B. mori	148	5	4	0	0	0	7	18*		1		34	221
D. melanogaster	101	3	27	11	0	0	5	0		6	2		
A. mellifera	32	3	2	0	0	0	2	0	3	5			
N. vitripennis	62		2	0	0	0	3	0	3	6			
T. castaneum			3	0	2	0	4	0		7	10		

Souces: Futahashi et al. 2008; Cornman et al. 2008, Cornman and Willis, 2009, Togawa et al. 2007, Jasrapuria et al. 2010.

*Gly-Rich family from *Bombyx* is really a composite of possibly 3 families (see text). The 6 that have been identified as CPLCPs were deleted from this number. Only the 18 restricted to lepidoptera that have several GGY repeats were included.

Empty boxes mean data not available. A ? indicates that CP status of these genes is uncertain (see text).

TABLE 2

Presence of cuticular protein families and features in different groups of Arthropods

	Subphylum: Atelocerata										Subphylum: Crustacea	Subphylum: Chelicerata
	Diptera		Lepido-	Coleo-	Hymeno-	Hemi-	Ortho-	Dictyo-	Phthira-	Collembola		
	Brachycera	Nematocera										
R&R Consensus												
RR-1	+	+	+	+	+	+	+	+	+	+	+	no
RR-2	+	+	+	+	+	+	+	+	+	+	+	+
CPF/CPFL	+	+	+	+	+	+	+	id	+	id		
TWDL	+	+	+	+	+	+	id	id	+	id	no	no
CPLCA	+	+	no	no	no	no	no	no	no	id	no	no
CPLCG	+	+	no	+	no	no	no	+	no	no	+	no
CPLCW	no	+	no	no	no	no	no	no	no	no	no	no
CPLCP	+	+	+	+	+				+			
GPG			+						no			
apidermin	no	no	no	no	+	no	id	no	no	no	no	no
CPAP 1	+	no				+			+		+	probably
CPAP3	+	+	+	+	+							
BcNCP1 orthologs	+	+	+	+	no	+	id	+	+	+	no	no
18 aa motif	+	+	+	+	+	+	+	+	+	id	+	no
CP with >3 AAP [AVL]	+	+	+	+	+	+	+	id	+	id	id	+
dumpy	+	+	+	+	+							

Final syllable ptera was removed from names of most insect orders.

RR-3 is not well defined, so it was not included.

Data were obtained from BLAST searches in addition to analyses in: Togawa et al., 2007; Futahashi et al., 2008; Cornman et al. 2008, Cornman and Willis, 2009; Carmon et al., 2007, Jasrapuria et al., 2010.

id = insufficient data available to record absence; empty boxes indicate that motifs were insufficiently well defined to allow a search.