Insect Biochemistry and Molecular Biology 40 (2010) 189–204

Contents lists available at ScienceDirect

Insect Biochemistry and Molecular Biology

journal homepage: www.elsevier.com/locate/ibmb

Review

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

ARTICLE INFO

Article history:
Received 7 December 2009
Received in revised form
9 February 2010
Accepted 10 February 2010

Keywords:
R&R Consensus
Whole genome sequences
Anopheles gambiae
Bombyx mori
Cuticle

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.

© 2010 Elsevier Ltd. All rights reserved.

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 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 online 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

* Tel.: +1 706 542 0802; fax: +1 706 542 4271.
E-mail address: jhwillis@cb.uga.edu

0965-1748/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ibmb.2010.02.001

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[a]		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		

Sources: Futahashi et al. 2008; Cornman et al. 2008, Cornman and Willis, 2009, Togawa et al. 2007, Jasrapuria et al., 2010.
Empty boxes mean data not available. ? indicates that CP status of these genes is uncertain (see text).
[a] 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.

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 "link-outs" 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

Download English Version:

https://daneshyari.com/en/article/1982607

Download Persian Version:

https://daneshyari.com/article/1982607

Daneshyari.com