Insect Biochemistry and Molecular Biology 62 (2015) 100–113

Contents lists available at ScienceDirect

Insect Biochemistry and Molecular Biology

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

Annotation and expression analysis of cuticular proteins from the tobacco hornworm, *Manduca sexta*

Neal T. Dittmer [a, *], Guillaume Tetreau [b], Xiaolong Cao [c], Haobo Jiang [d], Ping Wang [b], Michael R. Kanost [a]

[a] Department of Biochemistry and Molecular Biophysics, 141 Chalmers Hall, Kansas State University, Manhattan, KS 66506, USA
[b] Department of Entomology, Cornell University, New York State Agricultural Experiment Station, Geneva, NY 14456, USA
[c] Department of Biochemistry and Molecular Biology, Oklahoma State University, Stillwater, OK 74078, USA
[d] Department of Entomology and Plant Pathology, Oklahoma State University, Stillwater, OK 74078, USA

ARTICLE INFO

Article history:
Received 25 October 2014
Received in revised form
19 December 2014
Accepted 29 December 2014
Available online 8 January 2015

Keywords:
Cuticular proteins
Rebers & Riddiford
RNAseq
Orthology
Manduca sexta

ABSTRACT

The insect cuticle is a unique material that covers the exterior of the animal as well as lining the foregut, hindgut, and tracheae. It offers protection from predators and desiccation, defines body shape, and serves as an attachment site for internal organs and muscle. It has demonstrated remarkable variations in hardness, flexibility and elasticity, all the while being light weight, which allows for ease of movement and flight. It is composed primarily of chitin, proteins, catecholamines, and lipids. Proteomic analyses of cuticle from different life stages and species of insects has allowed for a more detailed examination of the protein content and how it relates to cuticle mechanical properties. It is now recognized that several groups of cuticular proteins exist and that they can be classified according to conserved amino acid sequence motifs. We have annotated the genome of the tobacco hornworm, *Manduca sexta*, for genes that encode putative cuticular proteins that belong to seven different groups: proteins with a Rebers and Riddiford motif (CPR), proteins analogous to peritrophins (CPAP), proteins with a tweedle motif (CPT), proteins with a 44 amino acid motif (CPF), proteins that are CPF-like (CPFL), proteins with an 18 amino acid motif (18 aa), and proteins with two to three copies of a C-X_5-C motif (CPCFC). In total we annotated 248 genes, of which 207 belong to the CPR family, the most for any insect genome annotated to date. Additionally, we discovered new members of the CPAP family and determined that orthologous genes are present in other insects. We established orthology between the *M. sexta* and *Bombyx mori* genes and identified duplication events that occurred after separation of the two species. Finally, we utilized 52 RNAseq libraries to ascertain gene expression profiles that revealed commonalities and differences between different tissues and developmental stages.

1. Introduction

The insect exoskeleton (cuticle) is a remarkable extracellular structure secreted by epidermal cells that serves as the outer body covering. It helps to protect the insect from environmental stresses such as predators, parasites, abrasions, desiccation, and UV radiation. It also functions as an attachment site for internal muscles and organs, and is instrumental to locomotion and flight. The bulk of the cuticle is made primarily of a network of chitin embedded in a proteinaceous matrix, with water, catecholamines, and some lipids (Moussian, 2013). Despite what may seem like a limited pallet of materials, insects are able to synthesize cuticles that differ widely with respect to physical properties. Measurements of hardness have differed by more than 30 fold, while the stiffness of cuticle has been shown to vary by more than seven orders of magnitude (Klocke and Schmitz (2011) and references within; Vincent and Wegst, 2004). Investigations of the reasons behind these differences have focused on the role of dehydration and chemical cross-linking (Andersen, 2010; Sugumaran, 2010; Vincent, 2009). However, it has been apparent for many years that differences exist between proteins extracted from different types of cuticle (hard versus soft), different developmental stages (i.e. larva, pupa, adult), as well as different time points (pre-molt versus post-molt) (Andersen et al., 1986, 1995a; Andersen and Højrup, 1987; Cox

* Corresponding author. Tel.: +1 785 532 4033.
E-mail address: ndittmer@k-state.edu (N.T. Dittmer).

http://dx.doi.org/10.1016/j.ibmb.2014.12.010
0965-1748/

and Willis, 1985; Dittmer et al., 2012; Jensen et al., 1997; Kiely and Riddiford, 1985; Missios et al., 2000). Obtaining sequence information of these proteins was limited as it required either solubilization, purification, digestion, and sequencing of peptide fragments, or screening of cDNA expression libraries with antibodies raised against cuticle extracts in order to identify the proteins involved. Now, the emergence of large scale genomic, proteomic, and transcriptomic analyses has allowed for a renewed look at the importance of the protein content.

The first insect genome sequenced was that of *Drosophila melanogaster* (Adams et al., 2000); there are now 112 insect genome sequences available at NCBI. An analysis of 12 genomes by Ioannidou et al. (2014) suggested that genes encoding structural cuticular proteins (CP) represent on average 1% of the total protein-coding genes in insects. The total number of genes varied from 63 for *Pediculus humanus* to 301 for *Aedes aegypti*. These numbers, as well as those given for other insects, likely represent a minimum as most genomes have not gone through a rigorous manual annotation and rely on homology to other known CPs. Combining proteomics with genomics has been used to identify cuticular proteins for several insect species (Bae et al., 2011; Carrasco et al., 2011; Dittmer et al., 2012; Fu et al., 2011; He et al., 2007). Similarly, genomics and transcriptomics have contributed valuable information on CP gene expression, discerning variations in the timing (pre- or post-molt), developmental stage (larval, pupal, adult), and relative expression levels of these genes (Cornman and Willis, 2009; Dittmer et al., 2012; Futahashi et al., 2008; Gallot et al., 2010; Liang et al., 2010; Okamoto et al., 2008; Suetsugu et al., 2013; Togawa et al., 2007, 2008).

Many CPs can be classified by the conserved sequence motifs they contain (Ioannidou et al., 2014; Willis et al., 2012). The largest group is known as CPs with the Rebers and Riddiford motif (CPR) that contain a core 28 amino acid sequence that is now recognized as part of a larger 63 amino acid consensus sequence (pfam00379) (Rebers and Riddiford, 1988; Willis et al., 2012). Variations in the extended consensus have been recognized, and CPs are often classified into one of three sub types: RR-1, RR-2, and RR-3. The number of CPR genes can vary greatly among species but many insects have more than 100 (Ioannidou et al., 2014). Additional groups having conserved sequence motifs include CPs analogous to peritrophins (CPAP) (Jasrapuria et al., 2010), CPs with a 44 amino acid motif (CPF) (Andersen et al., 1997; Togawa et al., 2007), CPF-like (CPFL) (Togawa et al., 2007), CPs with a Tweedle motif (CPT) (Guan et al., 2006), CPs with two or three repeats of C-X_5-C motif (CPCFC) (Jensen et al., 1997; Willis et al., 2012), and CPs with an 18 amino acid motif (Andersen, 2000; Nakato et al., 1990). Additionally, low-complexity proteins can be found in the cuticle that are rich in glycine or contain repeats of AAP(A/V), P(V/Y), GYGL, or GLLG (Willis et al., 2012). However, since these proteins lack any other distinctive sequences, presence of these repeats alone is not proof enough of their location in the cuticle. Excellent reviews on CPs can be found in Willis (2010) and Willis et al. (2012).

The goal of this research was to annotate the CP genes in the genome of the tobacco hornworm, *Manduca sexta*, with respect to the groups described above. These seven groups (CPR, CPAP, CPF, CPFL, CPT, CPCFC, and 18 aa motif) were chosen as they all contain conserved sequences previously shown to be present in known cuticular structural proteins and, therefore, can serve as a diagnostic feature. We compared the CP genes in *M. sexta* with those of *Bombyx mori* (Futahashi et al., 2008) in order to identify orthologs. Finally, we utilized 52 RNAseq libraries prepared from various tissues and developmental stages to look for patterns of coordinated expression among the CP genes. This analysis offers new insights into the CPs present in cuticle synthesized at different times and developmental stages.

2. Materials and methods

2.1. CP gene annotation

To identify putative CP genes, the *M. sexta* genome and first official gene set (OGS1), available from the Agricultural Pest Genomics Resource Database (www.agripestbase.org), were searched by tBLASTn (Altschul et al., 1997) with the following sequences: for the CPR family, the consensus sequence GxFxYxxPDGxxxxVxYx-ADENGYQPxGAHLP was used to identify the RR-1 subtype, and EYDAxPxYxFxYxVxDxHTGDxKSQxExRDGDVVxGxYSL-xExDGxxRTVxYTADxxNGFNAVVxxE was used to identify the RR-2 subtype (Figure 3 in Willis et al., 2005); classification into the appropriate subfamily was confirmed by the use of a profile hidden Markov model that discriminates between the two subtypes, available at the cuticleDB website (biophysics.biol.uoa.gr/cuticleDB/, Karouzou et al., 2007). For the CPAP family, the *Tribolium castaneum* CPAP1-D (GenBank ACY95469) and CPAP3-A1 (GenBank ACY95475) sequences were used to identify CPAP family members. CPF genes were identified using the most highly conserved portion of the 44 amino acid motif, VSxYSKAVDTPFSSVRKxDxRIVNxA (derived from Fig. 1B in Togawa et al., 2007). CPFL genes were identified with the sequence LxYSAAPAVSHVAYxGxGxxYGW (derived from Figure 3 in Togawa et al., 2007). For Tweedle genes, the 100 amino acid weblogo sequence from Figure 3A in Willis (2010) was used to identify homologs. The sequence YPAGVN-PAACPNYPYCD was used to identify members of the CPCFC family, and PVDTPEVAAAKAAHFAAH was used to identify genes encoding CPs with the 18 amino acid motif (Figure 4C and B in Willis, 2010).

For all genes except those of the CPAP family, names were assigned based on putative orthology to *B. mori* homologs. Reciprocal BLAST was performed to confirm that the *B. mori* protein identified as the top hit to an *M. sexta* query identified the same *M. sexta* protein as the top hit when it was used as the query sequence. Orthology was further established through the use of microsynteny. When the genes flanking a CP gene were homologous between *B. mori* and *M. sexta*, the CP genes were considered to be orthologous; identification of genes surrounding *B. mori* CP genes was inferred from the Gene Report page at NCBI associated with that particular CP gene. Because the CPAP genes in *B. mori* have not been annotated yet, naming of the *M. sexta* CPAP genes was based on a phylogenetic analysis of homologous proteins from several insect species; a detailed description of this analysis can be found in Tetreau et al. (2015).

2.2. Phylogenetic analysis

Phylogenetic analysis was performed using the corresponding protein sequence of selected genes from the CPR, CPAP, and CPFL groups. Details of the CPAP analysis can be found in a companion paper (Tetreau et al., 2015). Analysis was performed using MEGA software (v5.2.1; Tamura et al., 2011). Sequences were aligned globally using the ClustalW program in MEGA and then adjusted manually by eye. For the CPR group, the extended RR domain (pfam00379) was used; RR-1 and RR-2 subgroups were treated separately. For the CPFL group, the entire sequence of the *M. sexta* and *B. mori* proteins were used. Trees were constructed by the neighbor-joining method with a Poisson correction model. Gaps were treated by the pairwise deletion method and statistical analysis was performed by the bootstrap method using 1000 repetitions.

2.3. CP gene expression

Fifty-two RNAseq libraries had been prepared from various tissues and developmental stages as part of the Manduca Genome