Insect Biochemistry and Molecular Biology 38 (2008) 1138–1146

Contents lists available at ScienceDirect

Insect Biochemistry and Molecular Biology

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

Genome-wide identification of cuticular protein genes in the silkworm, *Bombyx mori*

Ryo Futahashi [a,b], Shun Okamoto [a], Hideki Kawasaki [c], Yang-Sheng Zhong [c,1], Masashi Iwanaga [c], Kazuei Mita [b], Haruhiko Fujiwara [a,*]

[a] Department of Integrated Biosciences, Graduate School of Frontier Sciences, The University of Tokyo, Bioscience Building 501, Kashiwa, Chiba 277-8562, Japan
[b] National Institute of Agrobiological Sciences, Owashi 1-2, Tsukuba 305-8643, Japan
[c] Faculty of Agriculture, Utsunomiya University, 350 Mine, Utsunomiya, Tochigi 321-8505, Japan

ARTICLE INFO

Article history:
Received 31 October 2007
Received in revised form 1 April 2008
Accepted 6 May 2008

Keywords:
Bombyx mori
Cuticle
Cuticular protein
R&R consensus
EST library
Whole genome sequence

ABSTRACT

Many kinds of cuticular proteins are found in a single insect species and their numbers and features are diversified among insects. Because there are so many cuticular proteins and so much sequence variation among them, an overview of cuticular protein gene is needed. Recently, a complete silkworm genome sequence was obtained through the integration of data from two whole genome sequence projects performed independently in 2004. To identify cuticular protein genes in the silkworm *Bombyx mori* exhaustively, we searched both the *Bombyx* whole genome sequence as well as various EST libraries, and found 220 putative cuticular protein genes. We also revised the annotation of the gene model, and named each identified cuticular protein based on its motif. The phylogenetic tree of cuticular protein genes among *B. mori*, *Drosophila melanogaster*, and *Apis mellifera* revealed that duplicate cuticular protein clusters have evolved independently among insects. Comparison of EST libraries and northern blot analyses showed that the tissue- and stage-specific expression of each gene was intricately regulated, even between adjacent genes in the same gene cluster. This study reveals many novel cuticular protein genes as well as insights into cuticular protein gene regulation.

1. Introduction

Insect cuticle is composed of many kinds of cuticular proteins together with chitin. The numbers and features of cuticular proteins are diversified among insects, whereas chitin is a uniform polymer of *N*-acetylglucosamine. Because a majority of excreted cuticle components are cross-linked and therefore inextractable (Andersen et al., 1995), and because the amino acid sequences of cuticular protein genes are not well conserved among insects, the overview of cuticular protein genes is still largely unknown. The majority of cuticular proteins have the Rebers and Riddiford Consensus (R&R Consensus), which in an extended form is known to bind chitin (Rebers and Willis, 2001; Togawa et al., 2004; Willis et al., 2005). Proteins with R&R Consensus can be split into three groups, RR-1, RR-2, and RR-3, with some correlation to the type or region of the cuticle. Recently, other motifs of cuticular proteins have been reported. In *Drosophila melanogaster*, the Tweedle motif was found by identification of a body shape mutant (Guan et al., 2006). Because Tweedle proteins are predicted to form β-strands, and because a barrel structure formed by multiple β-strands provides an interface for aromatic residues to stack with and bind to chitin (Iconomidou et al., 1999; Hamodrakas et al., 2002), studies have postulated that Tweedle proteins interact directly with chitin (Guan et al., 2006). A motif of 51 amino acids was described (Andersen et al., 1997) but more recently Togawa et al. (2007) found that when more species were examined, the conserved motif was no more than 44 amino acids (cuticular protein with a 44 amino acid motif, CPF). Two proteins with this motif did not bind chitin in their assay. They also reported CPF-like proteins (CPFL), which lack the conserved 44 aa residues but their C-terminal regions are similar. It is also known that some cuticular proteins do not possess these motifs, and cuticular proteins devoid of the above mentioned motifs were also isolated and sequenced directly from cuticles (Andersen et al., 1995; Willis et al., 2005; He et al., 2007).

Recently, comprehensive identifications of cuticular proteins with an R&R motif have been attempted in *Drosophila melanogaster* (Karouzou et al., 2007) and *Apis mellifera* (Honeybee Genome Sequencing Consortium, 2006) based on their genome sequences. In the studied genomes, 101 and 28 genes with the R&R motif were

* Corresponding author. Department of Integrated Biosciences, Graduate School of Frontier Sciences, The University of Tokyo, Bioscience Building 501, Kashiwa, Chiba 277-8562, Japan. Tel.: +81 47136 3659; fax: +81 47136 3660.
E-mail address: haruh@k.u-tokyo.ac.jp (H. Fujiwara).
[1] Present address: Department of Sericulture, South China Agricultural University, China.

0965-1748/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ibmb.2008.05.007

found, respectively. These studies imply that the composition of cuticular protein genes vary among insect taxa. In 2004, Japanese and Chinese groups independently published the *Bombyx mori* genome draft sequences (Mita et al., 2004; Xia et al., 2004). Recently, these two data sets were merged and assembled through collaboration between China and Japan (The International Silkworm Genome Sequencing Consortium, in preparation), which has helped us to screen for genes of interest on a genome-wide scale. They also constructed gene model, which consists of all the genes predicted by gene finder BGF (Li et al., 2005) by pre-filtering classifiable transposable elements. Furthermore, *Bombyx* is suitable for the study of tissue specificity because its tissue size is relatively large, and it is easy to construct a tissue-specific cDNA library. Many EST libraries from various tissues have now been constructed (Mita et al., 2003; Kawasaki et al., 2004; Kinjoh et al., 2007; Okamoto et al., in press).

In the silkworm, *B. mori*, 28 cuticular protein genes have been already reported before merging the two genome data sets (Supplementary Table 1; Nakato et al., 1990, 1997; Takeda et al., 2001; Suzuki et al., 2002; Sawada et al., 2003; Noji et al., 2003; Zhong et al., 2006; Togawa et al., 2007). Here we performed a genome-wide screen for cuticular protein genes in the silkworm, *B. mori*. To identify cuticular protein genes exhaustively, we searched both the *Bombyx* whole genome sequence and EST libraries using not only the known motif described above, but also using the following criteria: (1) N-terminal signal peptide, (2) simple repeat sequence (GGX or AAP(A/V)), and (3) sequence similarity to known cuticular proteins. We found 220 putative cuticular protein genes (RR-1 56, RR-2 89, RR-3 3, Tweedle 4, CPF 1, CPFL 4, glycine-rich 29, and 34 other genes). We revised the annotation of the gene model in The International Silkworm Genome Sequencing Consortium (in preparation) and named each identified cuticular protein based on its motif. Phylogenetic analysis of RR-1 and RR-2 proteins among *B. mori*, *D. melanogaster*, and *A. mellifera* suggested that duplicate cuticular protein clusters have evolved independently among insect taxa. Comparison of EST libraries revealed that motif differences correlate to the tissue specificity. However, the expression of each cuticular protein does not correlate based on chromosomal location, and often differed between the adjacent genes. Through northern blot analysis we also found that stage-specific expression of cuticular protein in the wing disc also varied between adjacent genes. These results indicate that expression of each cuticular protein is regulated intricately, even in the same clusters.

2. Materials and methods

2.1. Prediction of cuticular protein genes in genome-wide scale

First we identified genes that might code for cuticular proteins with the following known motif, R&R Consensus (Rebers and Riddiford, 1988; Iconomidou et al., 1999), Tweedle motif (Guan et al., 2006), and 44 aa residues (Togawa et al., 2007) by using tBLASTn search both the whole genome sequence (The International Silkworm Genome Sequencing Consortium, in preparation) and various EST libraries (Mita et al., 2003; Kinjoh et al., 2007; Okamoto et al., in press). R&R Consensus is also confirmed by using a tool based on profile hidden Markov models (HMMs) in cuticleDB website <http://bioinformatics2.biol.uoa.gr/cuticleDB/index.jsp> (Karouzou et al., 2007), which is capable of discriminating between RR-1 and RR-2 cuticular proteins. In addition to these known motifs, we also predicted putative cuticular protein genes using several criteria as follows: (1) N-terminal signal peptide, (2) simple repeat sequence (GGX or AAP(A/V)), and (3) sequence similarity to known cuticular proteins.

2.2. Naming protocol

We adopted a simple nomenclature with their motif or characteristics such as *BmorCPT1* (***B. mori c****uticular* ***p****rotein* ***T****weedle motif* ***1***). Naming of cuticular protein was referred as other insect species such as *D. melanogaster* (Karouzou et al., 2007), *Anopheles gambiae* (He et al., 2007), and *A. mellifera* [cuticleDB website <http://bioinformatics2.biol.uoa.gr/cuticleDB/index.jsp>; Magkrioti et al., 2004]. For each type of cuticular protein except for CPG, we have numbered the genes in the order on which they appear on the chromosome. As for CPG, because CPG1 has been already reported (Suzuki et al., 2002), we first numbered CPG genes in the gene cluster including CPG1 (see Supplementary Table 1). 28 *Bombyx* cuticular protein genes have been published previously or available on NCBI database. (Most of these proteins are found in cuticleDB website described above.) We searched for the genomic location of these 28 previously identified genes. In 26 cases (except for *BmLCP18* and *BMWCP11*), the genomic sequence was not identical to the previously published sequence (see Supplementary Table 4). For example, the amino acid sequences are not identical between CPFLa1 and BmorCPFL3, and between CPFLb and BmorCPFL4. In *Bombyx* genome project, p50 (Dazao) strain was used (Mita et al., 2004; Xia et al., 2004), however, p50 strain has not been used to identify the cuticular protein genes in the previous study (see Supplementary Table 4). Because cuticular protein genes often are very similar and may code for proteins with identical amino acid sequences (Takeda et al., 2001; He et al., 2007; Karouzou et al., 2007), we could not conclude that the previously reported genes are the same as the genes identified in this study using the whole genome sequence. Takeda et al. (2001) reported two sets of very similar genes, (*BMWCP1a* and *BMWCP1b*) and (*BMWCP7a* and *BMWCP7b*). We found two different genomic regions corresponded to *BMWCP1a* and *BMWCP1b*, but the same genomic region corresponded to *BMWCP7a* and *BMWCP7b*. Moreover, deduced amino acid sequences of genome sequence are not identical with those of BMWCP1a, BMWCP1b, BMWCP7a and BMWCP7b. In *D. melanogaster*, the copy number of cuticular protein genes at band 65A varied even among strains (Charles et al., 1997), suggesting gene duplication may occur even among strains. Therefore we named all cuticular protein found in genome sequence to avoid confusing among strain differences. The corresponding genome regions (most similar genes) of known genes are denoted in Supplementary Table 1. *Bombyx* strain used for the previous study identifying *Bombyx* cuticular protein and each EST library is shown in Supplementary Tables 4 and 5.

2.3. Phylogenetic analysis

Sequences were aligned using Clustal_X (Thompson et al., 1997). Phylogenetic trees were constructed by the neighbor-joining method with the MEGA4 program (Tamura et al., 2007). The confidence of the various phylogenetic lineages was assessed by the bootstrap analysis. Amino acid sequences of cuticular protein genes of *D. melanogaster* and *A. mellifera* were obtained from cuticleDB website described above.

2.4. Comparison of EST libraries

In *B. mori*, various EST libraries have been constructed (Mita et al., 2003; Kawasaki et al., 2004; Kinjoh et al., 2007; Okamoto et al., in press). We searched 220 cuticular protein genes in these libraries and count the total clone numbers. Expression of several genes was also confirmed by reverse transcriptional-polymerase chain reaction (Okamoto et al., in press). Some libraries are constructed from another strain and nucleic acid sequence is slightly