Insect Biochemistry and Molecular Biology 42 (2012) 133–147

Contents lists available at SciVerse ScienceDirect

Insect Biochemistry and Molecular Biology

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

Comparative analysis of the UDP-glycosyltransferase multigene family in insects

Seung-Joon Ahn [1], Heiko Vogel, David G. Heckel*

Department of Entomology, Max Planck Institute for Chemical Ecology, Jena 07745, Germany

ARTICLE INFO

Article history:
Received 30 June 2011
Received in revised form
26 November 2011
Accepted 28 November 2011

Keywords:
UDP-glycosyltransferase
Helicoverpa armigera
Bombyx mori
Multigene family
Phylogenetic analysis
Detoxification

ABSTRACT

UDP-glycosyltransferases (UGT) catalyze the conjugation of a range of diverse small lipophilic compounds with sugars to produce glycosides, playing an important role in the detoxification of xenobiotics and in the regulation of endobiotics in insects. Recent progress in genome sequencing has enabled an assessment of the extent of the UGT multigene family in insects. Here we report over 310 putative UGT genes identified from genomic databases of eight different insect species together with a transcript database from the lepidopteran *Helicoverpa armigera*. Phylogenetic analysis of the insect UGTs showed Order-specific gene diversification and inter-species conservation of this multigene family. Only one family (UGT50) is found in all insect species surveyed (except the pea aphid) and may be homologous to mammalian UGT8. Three families (UGT31, UGT32, and UGT305) related to Lepidopteran UGTs are unique to baculoviruses. A lepidopteran sub-tree constructed with 40 *H. armigera* UGTs and 44 *Bombyx mori* UGTs revealed that lineage-specific expansions of some families in both species appear to be driven by diversification in the N-terminal substrate binding domain, increasing the range of compounds that could be detoxified or regulated by glycosylation. By comparison of the deduced protein sequences, several important domains were predicted, including the N-terminal signal peptide, UGT signature motif, and C-terminal transmembrane domain. Furthermore, several conserved residues putatively involved in sugar donor binding and catalytic mechanism were also identified by comparison with human UGTs. Many UGTs were expressed in fat body, midgut, and Malpighian tubules, consistent with functions in detoxification, and some were expressed in antennae, suggesting a role in pheromone deactivation. Transcript variants derived from alternative splicing, exon skipping, or intron retention produced additional UGT diversity. These findings from this comparative study of two lepidopteran UGTs as well as other insects reveal a diversity comparable to this gene family in vertebrates, plants and fungi and show the magnitude of the task ahead, to determine biochemical function and physiological relevance of each UGT enzyme.

1. Introduction

Glycoside conjugation is one of the most important metabolic pathways for biotransformation of a number of lipophilic xenobiotics and endobiotics. UDP-glycosyltransferases (UGTs) catalyze the conjugation of a sugar donated by a UDP-glycoside to a typically lipophilic molecule, generating water-soluble products which can be easily excreted as well as stably managed, therefore protecting the cellular system from being damaged by toxic foreign compounds and regulating internal molecules more easily (Bock, 2003).

UGTs are membrane-bound proteins located in the endoplasmic reticulum (ER) facing the lumen in animals. The UGT protein structure is divided into two main parts: the N-terminal aglycone substrate binding domain and the C-terminal UDP-glycoside binding domain. The N-terminal end of the animal UGTs has a signal peptide mediating the integration of the protein precursor into the ER compartment. The signal peptide is subsequently cleaved and then the protein is *N*-glycosylated. The mature protein is retained in the ER membrane by its hydrophobic transmembrane domain at the C-terminal end, followed by a short cytoplasmic tail (Magdalou et al., 2010).

UGTs are ubiquitous in all free-living organisms from bacteria to fungi, plants and animals. Most baculovirus genomes encode the enzyme ecdysteroid UDP-glycosyltransferase (EGT) which regulates the development of the host insect by glycosylating and inactivating ecdysteroid hormones (O'Reilly, 1995). Sterol-UGTs have also been found in a small dsRNA virus infecting

* Corresponding author. Tel.: +49 3641 571501; fax: +49 3641 571502.
E-mail address: heckel@ice.mpg.de (D.G. Heckel).
[1] Present address: Horticultural and Herbal Crop Environment Division, National Institute of Horticultural and Herbal Science, Rural Development Administration, Suwon 441-440, Republic of Korea.

0965-1748/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ibmb.2011.11.006

Phytophthora and two plant viruses in the genus *Endornavirus* (Hacker et al., 2005). In plants, a variety of soluble cytosolic UGTs play an important role in the modification of secondary metabolites, thereby enhancing their solubility and stability, and determining their bioactivity (Bowles et al., 2005). In vertebrates, membrane-bound UGTs are regarded as major members of the phase II drug metabolizing enzymes, conjugating a large number of xenobiotics as well as endobiotics, such as bilirubin and steroid hormones with UDP-glucuronic acid (Bock, 2003). In insects, the glycosylation of small lipophilic compounds has been considered as a minor enzymatic detoxification mechanism for half a century (Ahmad et al., 1986; Brattsten, 1988; Després et al., 2007; Smith, 1962). Insect UGT enzyme activity has been investigated in the housefly *Musca domestica* (Morello and Repetto, 1979), in the fruitfly *Drosophila melanogaster* (Real et al., 1991), in the tobacco hornworm *Manduca sexta* (Ahmad and Hopkins, 1992), in the silkworm *Bombyx mori* (Luque et al., 2002), and in other insects (Ahmad and Hopkins, 1993b). These biochemical studies have shown that the insect UGT enzymes typically use UDP-glucose as the main sugar donor unlike vertebrate UGTs which mainly utilize UDP-glucuronic acid. However, both insect and vertebrate UGTs are known to be bound to the endoplasmic reticulum in a similar manner. Enzyme activities of the insect UGTs are detected in the fat body, midgut and other tissues (Ahmad and Hopkins, 1993b), and are directed towards a variety of plant allelochemicals (Ahmad and Hopkins, 1993a; Luque et al., 2002; Sasai et al., 2009). Interestingly, the enzymes were also detected in the antenna of *D. melanogaster* (Wang et al., 1999). In addition, many endogenous compounds, like ecdysteroid hormones (Svoboda and Weirich, 1995) and cuticle tanning precursors (Ahmad et al., 1996; Hopkins and Kramer, 1992) are glycosylated by UGT enzymes. Furthermore, dietary flavonoids have been shown to be sequestered as glucose conjugates to impart color to the wings in a lycaenid butterfly (Wiesen et al., 1994) or in *B. mori* to be glycosylated to produce a green color in the cocoon with UV-shielding properties (Daimon et al., 2010). A UGT enzyme was recently shown to catalyze the final step in synthesis of cyanogenic glucosides by the Burnet moth *Zygaena filipendulae* (Jensen et al., 2011). These findings suggest multiple roles of the insect UGT enzymes in detoxification, olfaction, endobiotic modulation, and sequestration.

Many gene families involved in detoxification and metabolism in insects, like cytochrome P450s (P450s), carboxyl/cholinesterases (CCEs), and glutathione transferases (GSTs) have been identified from genome sequences of *D. melanogaster* (Tijet et al., 2001), *Aedes aegypti* (Strode et al., 2008), *Acyrthosiphon pisum* (Ramsey et al., 2010), *Nasonia vitripennis* (Oakeshott et al., 2010), *Apis mellifera* (Claudianos et al., 2006), and *B. mori* (Tsubota and Shiotsuki, 2010; Yu et al., 2008). The molecular identities of the insect UGTs are, however, relatively unknown compared to the other detoxification gene families in insects or to the vertebrate UGT gene families. Although the UGT gene families in *D. melanogaster* (Luque and O'Reilly, 2002) and in *B. mori* (Huang et al., 2008) have been reported, they have not been compared in the context of the diversity of other insects.

The cotton bollworm *Helicoverpa armigera* Hübner (Lepidoptera: Noctuidae) is a polyphagous herbivore feeding on more than 100 different species belonging to about 46 different plant families, and it is regarded as a serious agricultural pest in many regions in the world (Fitt, 1989). Thus it has been extensively studied in order to understand its ecology and physiology especially in relation to host plant adaptation and insecticide resistance (Heckel, 2010). Recently the *H. armigera* CCE gene family was described based on transcriptome analysis (Teese et al., 2010), which is likely incomplete based on comparison to the *Bombyx* genome; no other gene families related to detoxification have yet been described from this species.

In this study, we identified all the putative UGT genes from our *H. armigera* cDNA libraries and from *B. mori* genome databases, as well as from another seven insect genomes available in NCBI. Here we describe the genomic and phylogenetic analyses of the *H. armigera* and *B. mori* UGTs in the context of the insect UGT diversity, and make predictions of the UGT protein structure by comparing several important domains and residues with biochemically characterized human UGT sequences.

2. Materials and methods

2.1. UGT gene sequences from H. armigera

Putative UGT gene sequences were identified from a public database (NCBI GenBank) and from a locally generated EST database together with 454-pyrosequencing data of *H. armigera*. The *H. armigera* cDNA libraries in the Department of Entomology, Max Planck Institute for Chemical Ecology were produced from adult, pupa or various larval tissues (gut, fat body, salivary glands, hemocytes, integument, or rest of body) of the TWB strain originally from Toowoomba, Queensland, Australia. In total, 353 cDNA contigs were retrieved by tblastn searches, and the collected sequences were assembled by using Sequencher (Gene Codes Corporation, MI, USA). To obtain full-length sequences, RNA Ligase-Mediated Rapid Amplification of cDNA Ends (RLM-RACE) was carried out employing the GeneRacer Kit (Invitrogen) according to the manufacturer's instructions. Gene-specific primers were designed from the partial cDNA sequences assembled. Total RNA was extracted from a fifth-instar larva using Trizol (Invitrogen) followed by purification with an RNeasy kit (Qiagen). The RACE products were analyzed by 1.5% agarose gel electrophoresis, the excised bands were purified using QIAquick spin columns (QIAGEN) and sequenced with a DNA analyzer (Applied Biosystems, USA).

2.2. BLAST searches of insect UGT sequences from public databases

The *B. mori* UGT sequences were collected from KAIKObase (Shimomura et al., 2009) and SilkDB (Duan et al., 2010) based on the gene names in a previous report (Huang et al., 2008), which however did not give the actual sequences. BLAST searches of unannotated genomic contigs in NCBI were also performed to obtain additional sequences. Four UGT sequences reported from the Six-spot burnet moth *Z. filipendulae* were also included (Zagrobelny et al., 2009). In addition, non-lepidopteran UGT sequences were retrieved from seven representative insect species, of which genome sequences have been either completed or at least assembled. Among them are *Drosophila melanogaster*, *Anopheles gambiae*, *Aedes aegypti*, *Acyrthosiphon pisum*, *Apis mellifera*, *Tribolium castaneum*, and *N. vitripennis*. BLAST searches were performed against assembled RefSeq as well as against WGS databases from each genome. Individual genes were manually annotated using genomic sequences together with EST sequences. When no cDNA sequences were available, the open reading frame was predicted from genomic sequence by comparison with known UGT sequences. In addition, UGT sequences from several baculoviruses were obtained by BLAST searches of NCBI.

2.3. Nomenclature

According to the current UGT nomenclature guidelines (Mackenzie et al., 1997; Ross et al., 2001), families are defined at 40% amino acid sequence identity (aaID) or greater and subfamilies defined at 60% aaID or greater. Multiple sequence alignment was performed by using CLUSTAL and adjusted manually. Preliminary grouping was done using the program H-CD-HIT (Li and Godzik,