Molecular Phylogenetics and Evolution 76 (2014) 127–133

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Evolutionary diversification of aminopeptidase N in Lepidoptera by conserved clade-specific amino acid residues

Austin L. Hughes *

Department of Biological Sciences, University of South Carolina, Columbia, SC 29208, USA

ARTICLE INFO

Article history:
Received 14 October 2013
Revised 3 March 2014
Accepted 14 March 2014
Available online 24 March 2014

Keywords:
Aminopeptidase N
Convergent/parallel evolution
Diptera
Lepidoptera, Multi-gene family

ABSTRACT

Members of the aminopepidase N (APN) gene family of the insect order Lepidoptera (moths and butterflies) bind the naturally insecticidal Cry toxins produced by the bacterium *Bacillus thuringiensis*. Phylogenetic analysis of amino acid sequences of seven lepidopteran APN classes provided strong support for the hypothesis that lepidopteran APN2 class arose by gene duplication prior to the most recent common ancestor of Lepidoptera and Diptera. The Cry toxin-binding region (BR) of lepidopteran and dipteran APNs was subject to stronger purifying selection within APN classes than was the remainder of the molecule, reflecting conservation of catalytic site and adjoining residues within the BR. Of lepidopteran APN classes, APN2, APN6, and APN8 showed the strongest evidence of functional specialization, both in expression patterns and in the occurrence of conserved derived amino acid residues. The latter three APN classes also shared a convergently evolved conserved residue close to the catalytic site. APN8 showed a particularly strong tendency towards class-specific conserved residues, including one of the catalytic site residues in the BR and ten others in close vicinity to the catalytic site residues. The occurrence of class-specific sequences along with the conservation of enzymatic function is consistent with the hypothesis that the presence of Cry toxins in the environment has been a factor shaping the evolution of this multi-gene family.

1. Introduction

Gene duplication, followed by functional divergence of duplicated gene copies, is believed to be an important mechanism in the origin of new adaptive phenotypes (Hughes, 1994). In the genomes of multi-cellular eukaryotes, many genes belong to multi-gene families that have originated as a result of numerous gene duplication events over evolutionary history (Friedman and Hughes, 2001). Phylogenetic methods can be used to reconstruct the history of gene duplication within multi-gene families; and by comparing the phylogenies of genes with organismal phylogenies, it is possible to determine the timing of gene duplication relative to major cladogenetic events, thereby shedding light on the evolution of novel adaptive strategies.

Aminopeptidase N (APN) designates a class of zinc metalloproteinases that preferentially cleave single uncharged amino acids from the N-terminus of polypeptides, which are encoded in insects by the members of a multi-gene family (Nakanishi et al., 2002; Albiston et al., 2004; Piggott and Ellar, 2007; Crava et al., 2010). APNs expressed in the midgut of members of the insect order Lepidoptera (moths and butterflies) have been intensely investigated because they bind the Cry toxins produced by the bacterium *Bacillus thuringiensis* (Bt). These toxins are harmless to mammals and thus have potential as natural insecticides (Knight et al., 1994; Bravo et al., 2007). There do not appear to be any published estimates of the time over which *B. thuringiensis* and the insects it infects have co-evolved. However, the extensive diversity of Bt Cry proteins and their toxicity to insects of several different orders suggest a long co-evolutionary history, perhaps hundreds of millions of years (Schnepf et al., 1998; de Maagd et al. 2001). APNs are encoded by the members of a multi-gene family in Lepidoptera (Crava et al., 2010); thus, it is possible that selection arising from Cry toxins has been a factor in the evolution of the insect APN multi-gene family.

The individual susceptibilities of different APN family members to Cry toxins have been investigated experimentally in certain lepidopteran species. For example, in the silkworm *Bombyx mori*, protein fragments corresponding to the toxin-binding regions of four different APN family members were found to bind the Bt toxins Cry1Aa and Cry1ab *in vitro* (Nakanishi et al., 2002). However, in the brush border membrane vesiscle of *B. mori*, only one intact silkworm APN (APN1) showed detectable binding of the same Bt

* Address: Department of Biological Sciences, University of South Carolina, Coker Life Sciences Bldg., 700 Sumter St., Columbia, SC 29208, USA. Fax: +1 803 777 4002.
E-mail address: austin@biol.sc.edu

http://dx.doi.org/10.1016/j.ympev.2014.03.014
1055-7903/

toxins (Nakanishi et al., 2002). This difference might be attributed to differences in expression, toxin-binding, or both among *B. mori* APNs (Nakanishi et al., 2002). Consistent with an important role for APN1 in the toxicity of Bt Cry, resistance to Cry toxins is associated in several lepidopteran species with lack of expression of the APN1 ortholog (Herrero et al., 2005; Zhang et al., 2009).

A number of authors have published phylograms of lepidopteran APNs (Nakanishi et al., 2002; Rajagopal et al., 2003; Angelucci et al., 2008; Crava et al., 2010). Initially, these analyses were somewhat inconsistent in the identification and nomenclature of APN paralogs in Lepidoptera, but much of this confusion was resolved by the thorough analysis of Crava et al. (2010), which identified eight classes of lepidopteran APNs (APN1-8). However, published analyses have not used model-based phylogenetic methods, instead relying on pairwise alignment similarity; and therefore evolutionary relationships among the APN classes remain unclear. In addition, published analyses have often not included APN sequences from non-lepidopteran insect orders which might be used to date APN gene duplications relative to the origin of the Lepidoptera.

Here I conduct a maximum likelihood phylogenetic analysis of APN sequences from Lepidoptera and from Diptera (the flies), whose most recent common ancestor (MRCA) with Lepidoptera lived in the Permian Period 250–300 million years ago (Wiegmann et al., 2009). In order to examine patterns of functional differentiation among paralogs, I analyze data on gene expression in *B. mori* (Xia et al., 2007), examine patterns of sequence conservation, and reconstruct amino acid sequences changes on the branches leading to major clades of APNs. The goal of reconstructing ancestral amino acids was to identify amino acid residues that are both derived and conserved in each class of lepidopteran APNs. Derived residues (i.e., those that originated in the MRCA of a clade of sequences) that are conserved within that clade are candidates for functional specialization of clade members at the amino acid sequence level (Hughes, 2012a,b, 2013).

Gene duplication, followed by functional divergence of duplicated gene copies, is believed to be an important mechanism in the origin of new adaptive phenotypes (Nei, 1969; Hughes, 1994). In the genomes of multi-cellular eukaryotes, many genes belong to multi-gene families that have originated as a result of numerous gene duplication events over evolutionary history (Friedman and Hughes, 2001). Phylogenetic methods can be used to reconstruct the history of gene duplication within multi-gene families; and by comparing the phylogenies of genes with organismal phylogenies, it is possible to determine the timing of gene duplication relative to major cladogenetic events, thereby shedding light on the evolution of novel adaptive strategies (e.g., Friedman and Hughes, 2002; Roelofs and Rooney, 2003; Rewitz et al., 2007; Hughes 2012a,b, 2013). Here I apply these methods to gain insight into the functional diversification of lepidopteran APNs.

2. Methods

2.1. Phylogenetic analysis

The phylogenetic analysis was based on 81 selected APN-family amino acid sequences from Lepidoptera and Diptera. Lepidopteran species included representatives of seven families belonging to four superfamilies; the species used and their classification are listed in Table 1. Diptera were represented by two species of mosquitoes (family Culicidae) and two *Drosophila* species (Table 1). All sequences were downloaded from the NCBI database except for one (BGIBMGA008061), which was obtained from the Silkworm Genome Database (http://silkworm.genomics.org.cn/silkdb/). Biochemical assays for APN activity have been applied to certain sequences, particularly from Lepidoptera (Denolf et al., 1997; Rajagopal et al., 2003; Yang et al., 2010), but the majority of sequences have been assigned to the APN family by sequence homology (e.g., Crava et al., 2010).

The Cry toxin-binding domain was identified by homology following Nakanishi et al. (2002). Only sequences including the conserved DEP amino acid sequence motif at the C-terminus of the Cry toxin-binding domain were used in the analysis. The sequences used represented all of the Lepidopteran APN classes identified by Crava et al. (2010) except APN7, which lacks the DEP motif. Since no crystal structure of an insect APN is available, potentially functionally important motifs were identified by aligning insect APNs with human APN (gi157266300), for which a crytal structure is available (Wong et al., 2012). The human APN had a 33.6% mean amino acid sequence identity with the insect APN sequences used in this study.

Amino acid sequences were aligned by the CLUSTAL algorithm in MEGA 5.2 (Tamura et al., 2011); see Supplementary Fig. S1. In evolutionary analyses of a set of sequences, any site at which the alignment postulated a gap in any sequence was excluded from the analyses. The maximum likelihood (ML) analysis was based on the WAG + G + I + F model, which was chosen using the Bayes Information Criterion in MEGA 5.2. The reliability of the clustering patterns in the ML tree was tested by bootstrapping; 1000 bootstrap pseudo-samples were used.

The number of synonymous substitutions per synonymous site (d_S) and the number of synonymous substitutions per synonymous site (d_N) were estimated by the modified Nei-Gojobori method (Nei and Kumar 2000). This method takes into account transitional bias, which influences estimates of d_S and d_N, particularly at twofold degenerate sites (Li, 1993; Nei and Kumar, 2000; Vipan Kumar et al., 2012). We estimated the transition:transversion ratio from the sequences to be compared by the MCL method in MEGA.

Ancestral amino acid sequences (most probable ancestors) were reconstructed by ML in MEGA 5.2. For a given clade in the phylogeny, *conserved derived amino acid residues* were defined as those which were reconstructed to have arisen as amino acid replacements in the branch ancestral to that clade and which were 100% conserved in all members of that clade used in the analysis. Conserved derived amino acid residues are candidates for playing a role in clade-specific functions. Of course not all such replacements will be functionally significant, since some may result from selectively neutral substitutions that are conserved by chance. Obviously, the number of residues that are conserved within a given clade will be in part a function of the number of clade members (i.e., the number of evolutionary lineages) available for analysis, since more variants at functionally unimportant sites are likely to be seen in a larger sample of lineages than in a smaller sample. For this reason, the percentage of conserved sites which were derived was used as a measure of the extent of amino acid sequence specialization of each APN clade, since this percentage is expected to be independent of the number of lineages available for analysis.

2.2. Gene expression data

Normalized microarray expression scores of *Bombyx mori* (silkworm) probes corresponding to the *APN1*, *APN2*, *APN3*, *APN4*, *APN6*, and *APN8* genes were downloaded from the NCBI GEO database (accession GSE17571). The data were taken from a study (Xia et al. 2007) that used a custom genome-wide microarray with 22,987 70-mer oligonucleotide probes covering known and predicted *B. mori* genes. For each biological replicate, RNA extracted from 100 silkworms was pooled (Xia et al., 2007). The raw intensity data were normalized by a linear normalization method using four confirmed housekeeping genes as a standard (Xia et al., 2007). For purposes of the present analysis, when the original data set