Gene 467 (2010) 35–40

Contents lists available at ScienceDirect

Gene

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

Evolutionary conservation of amino acid composition in paralogous insect vitellogenins

Austin L. Hughes *

Department of Biological Sciences, University of South Carolina, Columbia SC 29205, USA

ARTICLE INFO

Article history:
Accepted 12 July 2010
Available online 23 July 2010

Received by S. Yokohama

Keywords:
Amino acid composition
Protein evolution
Purifying selection
Vitellogenin

ABSTRACT

Comparison of paralogous vitellogenins in 10 insect species representing six orders showed a remarkable degree of conservation of amino acid composition in spite of sequence differences. For example, the correlation between the percentages of the 20 amino acids in two vitellogenins from the beetle *Tribolium castaneum* was 0.975, even though the two amino acid sequences differed from each other at 49.4% of sites. There was a positive correlation between the frequency of occurrence of reciprocal pairs of amino acids in more distantly related paralogs, and this correlation was generally strongest when both of the amino acids in the pair were nutritionally essential. These results imply that conservation of amino acid composition occurs through amino acid replacements that result in a balanced loss and gain of each amino acid residue. Thus insect vitellogenins seem to be subject to an unusual kind of purifying selection, where the amino acid content is conserved rather than the sequence *per se*, selection apparently arising from the nutritional needs of the developing embryo appears to be responsible for maintaining the balance of amino acids.

1. Introduction

Proteins play a wide variety of roles in organisms, from structural components and enzymes to signaling molecules and receptors. Rather atypical among proteins are those whose functions involve the storage of amino acids for use in development of offspring, including the seed storage proteins of higher plants (Shewry and Halford 2002; Shutov et al., 2003) and the yolk storage proteins of animals (Tufail and Takeda, 2008). In insects, there are two distinct families of yolk proteins. One of these families, confined to the order Diptera, includes the yolk proteins of *Drosophila melanogaster*, whose expression in the egg has provided an important model system for understanding hormonal regulation of gene expression (Bownes, 1994). The other family of insect yolk proteins, known as vitellins, are processed from precursors known as vitellogenins, which are synthesized in the fat body and certain other tissues (Tufail and Takeda, 2008). The latter family is known to occur not only in certain Diptera (though not in *Drosophila*) but also several other insect orders and in other arthropod classes (Hwang et al., 2010).

Alignments of insect vitellogenins have indicated a small number of primary sequence features conserved in most but not all members of the family: (1) a motif GLCG or GICG in the C-terminal region, conserved in most insect vitellogenins; (2) nine conserved cysteine residues C-terminal to the latter conserved motif; and (3) an RXXR motif in the N-terminal region of most vitellogenins, where the protein is cleaved by proteases to form vitellins (Tufail and Takeda, 2008). The functional role of the GLCG/GICG motif and the conserved cysteines is not known. In Hemimetabola (insects with incomplete metamorphosis), the vitellogenin is cleaved into several polypeptides, whereas in most Holometabola (insects with complete metamorphosis), vitellogenin is cleaved into just two polypeptides. In the wasps, bees, and ants (Hymenoptera: Apocrita), cleavage is absent (Tufail and Takeda, 2008).

In spite of these apparent functional constraints, it might be predicted that most of vitellogenin protein will evolve in a very different fashion from typical proteins. Since the protein's primary function is to provide a reserve of amino acids for use in development, it might be predicted that many amino acid replacements will be selectively neutral or nearly so, as long as the overall protein maintains approximately the proportions of the various amino acids needed by the developing embryo. A mutation that causes the proportion of each amino acid residue to deviate from the balance that meets the nutritional requirements of the developing embryo would be predicted to be slightly deleterious, while a mutation that restores the balance of amino acids would be slightly advantageous. The effect of each individual mutation might be expected to be slight because of the substantial size of the vitellogenin protein. Over evolutionary time, we might expect that such slightly deleterious and slightly

Abbreviations: r_{rec}, the correlation between the frequency of each amino acid difference with that of its reciprocal amino acid difference; r_{use}, the correlation between the percentages of the 20 amino acids.

* Department of Biological Sciences, University of South Carolina, Coker Life Sciences Bldg., 700 Sumter St., Columbia, SC 29208, USA. Tel.: +1 803 777 9186; fax: +1 803 777 4002.

E-mail address: austin@biol.sc.edu.

0378-1119/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2010.07.007

advantageous mutations would balance each other, thereby preserving the nutritional value of the vitellogenin. The insect vitellogenins constitute a multi-gene family, with 2 or 3 members reported from a number of species. The present study takes advantage of comparisons between paralogous vitellogenins of 10 insect species, representing six orders, in order to test the hypothesis that amino acid composition is conserved by a balance between slightly deleterious and slightly advantageous mutations.

2. Methods

2.1. Phylogenetic analysis

Phylogenetic analyses employed 60 amino acid sequences of vitellogenins from 37 insect species and five species of non-insect arthropods (for accession numbers, see Supplementary Table S1). Since most of these sequences were derived from unmapped genomes, it was not always possible to determine whether two database accessions from a given species represented two distinct loci, allelic sequences from the same locus, or alternative transcripts from the same locus. Therefore, as an operational rule of thumb, I used in analyses only one of any two sequences from the same species that differed from each other by less than 1% at amino acid sites. Amino acid sequences were aligned using the CLUSTAL X program (Thompson et al., 1997); in phylogenetic analysis, any site at which the alignment postulated a gap in any sequence was excluded from the computation of pairwise distances so that a comparable set of amino acid positions was used for each comparison. A phylogenetic tree was constructed by the neighbor-joining method (Saitou and Nei, 1987) on the basis of the JTT distance (Jones et al., 1992), with the assumption that rate variation among sites followed a gamma distribution. The shape parameter of the gamma distribution ($a = 2.19$) was estimated by the TREE-PUZZLE program (Schmidt et al., 2002). The reliability of clustering patterns in phylogenetic trees was assessed by bootstrapping (Felsenstein, 1985); 1000 bootstrap samples were used.

2.2. Paralogous pair comparisons

From the phylogenetic tree, pairs of paralogous vitellogenin sequences were chosen for 10 insect species; pairs of paralogous genes were chosen so that each pair was determined by the phylogenetic analysis to be phylogenetically (and thus statistically) independent of all other pairs. Note that not all of these pairs were reciprocally monophyletic; nonetheless, they were phylogenetically independent in that each sequence difference between the members of a given pair occurred independently of differences between the members of other pairs. These comparisons involved the following pairs (with numbers of aligned amino acid sites): (1) *Rhyparobia maderae 1* and *2* (1910 sites); (2) *Periplaneta americana 1* and *2* (1776 sites); (3) *Plautia stali 1* and *2*; (1146 sites); (4) *Aedes aegypti B1* and *B2* (1164 sites); (5) *Culex quinquefaciatus A1* and *A2* (1165 sites); (6) *Ochlerotatus atropalpus B1* and *B2* (1167 sites); (7) *Pediculus humanus 1* and *2* (1124 sites); (8) *Tribolium castaneum 1* and *2* (161 sites); (9) *Nasonia vitripennis 1* and *2* (1469 sites); and (10) *Solenopsis invicta 2* and *3* (1500 sites).

In analysis of the amino acid composition of these 10 pairs of sequences, I excluded any site at which the alignment postulated a gap in one of the two sequences relative to the paralogous sequence with which it was compared. The nutritionally essential amino acids for insects (F, H, I, K, L, M, R, T, V, and W) were defined according to Nation (2008). Chemical distances between amino acids were from Miyata et al. (1979). I compared the frequencies of reciprocal amino acid differences between the two members of each pair as follows. Let *1* and *2* designate the two sequences in the pair. If at a given site, sequence *1* has a certain amino acid (amino acid *X*) and sequence *2* has another amino acid (amino acid *Z*), then the amino acid pair for that site is *XZ*. The reciprocal amino acid difference (*ZX*) would occur at a site where sequence *1* has *Z* and sequence *2* has *X*. For example, the amino acid pairs IL (Ile-Leu) and LI constitute reciprocal amino acid differences. In the comparisons between the 10 pairs of paralogs, 180 of the 190 theoretically possible amino acid differences occurred at least once. For each of the 10 pairs of paralogs, I computed the correlation (r_{rec}) between the frequency of each amino acid difference with that of its reciprocal amino acid difference.

3. Results

3.1. Phylogenetic analysis

The phylogenetic tree of insect vitellogenin amino acid sequences was rooted with sequences from the tick *Ixodes scapularis* and from five species of Crustacea (Fig. 1). Vitellogenin sequences from 7 insect orders were included, and in each case the sequences from a given order clustered together (Fig. 1). The clusters of sequences from Coleoptera, Pthiraptera, Lepidoptera, and Diptera each received 100% bootstrap support; and that of sequences from Hymenoptera received 85% bootstrap support (Fig. 1). In the case of Hymenoptera, a cluster of sequences from the suborder Apocrita received 100% bootstrap support, but a sequence from the coleseed sawfly *Athalia rosae* (suborder Symphyta) fell outside that cluster (Fig. 1). Overall, deep branches within the phylogenetic tree were not well resolved; but the topology did not correspond to the known relationships of the insect orders (Kjer et al., 2006; Whiting, 2002). In particular, the hemimetabolous orders Blattodea and Hemiptera did not cluster outside the holometabolous orders (Fig. 1).

When multiple sequences were available from a given species, those sequences clustered together or with sequences from closely related species. For example, three sequences from the imported fire ant *S. invicta* clustered together with 95% bootstrap support (Fig. 1). Likewise, three sequences from the brown-winded green bug *P. stali* clustered together with 99% bootstrap support (Fig. 1). Of two sequences from the American cockroach, one (*P. americana 2*) clustered with sequences from two other cockroach species with 98% bootstrap support, while the other (*P. americana 1*) fell outside that cluster (Fig. 1). The three cockroach species included belong to three different families: Blattidae (*P. americana*), Blattellidae (*Blatella germanica*), and Blaberidae (*R. maderae*). Thus, the topology supports the hypothesis that the two *P. americana* genes duplicated before Blattidae diverged from the latter two families.

Available sequences from the order Diptera represented three subfamilies from a single family, Culicidae (mosquitos): Anophelinae (*Anopheles*), Culicinae (*Culex*, *Ochlerotatus*, and *Aedes*) and Toxorhynchitinae (*Toxorhynchites*). The phylogenetic tree supported cases of both ancient and more recent duplication within this family. For example, *C. quinquefasciatus A1* clustered outside all other mosquito vitellogenins, including *C. quinquefasciatus A2*; and this pattern received 100% bootstrap support (Fig. 1). This topology supports the hypothesis that these two genes of *C. quinquefasciatus* duplicated before the Culicinae diverged from the other two subfamilies. On the other hand, the sequences from *Ae. aegypti* formed two clusters of two members each (designated, respectively, as *B1* and *B2* and *C1* and *C2*; Fig. 1). In the case of each of these two clusters, a sequence from *O. atropalpus* clustered outside the pair of *Ae. aegypti* genes; and in each case this topology received 100% bootstrap support (Fig. 1). This topology supported the hypothesis that the duplication of *Ae. aegypti B1* and *B2* and the duplication of *Ae. aegypti C1* and *C2* occurred after *Aedes* diverged from *Ochlerotatus*.

3.2. Amino acid and nucleotide usage

In comparisons between paralogous pairs of vitellogenins from 10 insect species, the frequencies of the 20 amino acids were highly