EVOLUTION & DEVELOPMENT 6:2, 67–77 (2004)

Evolution of a novel function: nutritive milk in the viviparous cockroach, *Diploptera punctata*

Anna Williford, Barbara Stay, and Debashish Bhattacharya*

Department of Biological Sciences, University of Iowa, Iowa City, Iowa 52242, USA

*Author for correspondence (e-mail: dbhattac@blue.weeg.uiowa.edu)

SUMMARY Cockroach species show different degrees of maternal contribution to the developing offspring. In this study, we identify a multigene family that encodes water-soluble proteins that are a major component of nutritive "Milk" in the cockroach, *Diploptera punctata*. This gene family is associated with the evolution of a new trait, viviparity, in which the offspring receive nutrition during the gestation period. Twenty-five distinct *Milk* complementary DNAs were cloned and partially characterized. These complementary DNAs encode 22 distinct Milk peptides, each of length 171 amino acids, including a 16-amino acid signal peptide sequence. Southern blot analysis confirms the presence of multiple copies of *Milk* genes in *D. punctata*. Northern analysis indicates tissue- and stage-specific *Milk* gene expression. Examination of the deduced amino acid sequences identifies the presence of structurally conserved regions diagnostic of the lipocalin protein family. The shared exon/intron structure of one of the *Milk* loci with lipocalin genes further supports a close evolutionary relationship between these sequences.

INTRODUCTION

A fundamental question in evolution is how novel complex traits arise in lineages. The most informative models to address this question are taxa that include species displaying incremental stages in the evolution of a complex trait (Reznick et al. 2002). This approach allows the identification of the sequence of established interactions that together result in a complex trait and potentially, the separate examination of each step. Central to the establishment of new functional interactions is the evolution of new gene functions. Novel gene functions are often accompanied by gene duplication. Eukaryotic genomes are rich in gene families (Li et al. 2001). It is estimated, for example, that the *Drosophila* genome contains 674 protein families, some containing up to 111 members, and that the *Caenorhabditis elegans* genome contains 1219 protein families, with up to 242 family members (Gu et al. 2002). Gene duplication may precede or follow the acquisition of a new function and is often necessary for the retention of a novel function. Because function is a result of an interaction, understanding the evolution of a new protein function requires not only the examination of changes in the protein itself, but changes in the interacting components as well. In this article, we study the evolution of a new protein function intimately involved with the emergence of a new life history trait in the cockroach *Diploptera punctata*.

The trait of interest is viviparity, a reproductive strategy that refers to the provision of nutrients to the developing embryos during the gestation period. This maternal contribution is supplemental to the yolk deposited in the egg during oocyte formation. The proteins in question are Milk proteins, the major component of the nutrient supply. Here we show that Milk proteins are encoded by a multigene family that putatively arose through multiple rounds of gene duplication. The cockroach is a promising model with which to study the contribution of gene duplication to the origin of new traits because (a) the evolution of Milk proteins is vital to the evolution of a new trait (viviparity), (b) cockroaches include species with different reproductive strategies and the direction of evolutionary change is known (i.e., viviparity is a derived trait), and (c) the evolution of Milk proteins can be examined in relation to the parallel changes in morphology and physiology that facilitated the evolution of viviparity.

Cockroaches, order Dictyoptera, suborder Blattaria, include about 4000 species only one of which, *D. punctata*, is known to be viviparous (Roth and Willis 1955). Most cockroaches are oviparous species that lay eggs, and the yolk deposited during oocyte maturation is the only maternal contribution to its offspring. Less numerous are ovoviviparous species of cockroaches that retain embryos within the female's body throughout embryogenesis but do not provide nourishment in addition to that present in the egg (Roth 1970). In *D. puncata*, and in ovoviviparous species of

cockroaches, embryos complete their development within a brood sac and are born as first instar larvae. The brood sac is an infolding of the ventral intersegmental membrane at the posterior end of the abdomen that provides water to the embryos and protection from desiccation and parasitism (Roth and Willis 1954; Stay and Coop 1974). However, in *D. punctata*, the brood sac epithelium also functions as the source of a nutritive secretion ingested by the embryos during 80% of the gestation time (Stay and Coop 1973, 1974). As the gestation time progresses, the yolk enclosed in the embryo gut is replaced by Milk. At about 43% of the gestation time, small crystals begin to accumulate in the embryo midgut; these are presumed to be a storage form of Milk proteins. These proteins are deposited in large amounts and are primarily responsible for the 60-fold increase in embryo protein content during the gestation period (Stay and Coop 1973).

In an attempt to understand the transition from ovoviviparity to viviparity, we focus on Milk proteins, a major component of the nutritive secretion. In this article, we report a partial characterization of Milk proteins, the identification of 25 distinct *Milk* complementary DNAs (cDNAs), the exon/intron structure of one *Milk* locus, the phylogenetic history of *D. punctata* Milk proteins, and the relationship between Milk and lipocalin proteins.

MATERIALS AND METHODS

Insects

Animals were maintained at 27°C with a 12:12-h light:dark cycle and fed Purina Lab Chow (St. Louis, MO, USA) and water. The age of the animals is given in days, with the day of adult emergence designated as day 0. A batch of 12 fertilized eggs is deposited in the brood sac at day 8 of adult female age. Embryos are retained within the brood sac during embryogenesis, which lasts about 62 days, and are born as first instar larvae on day 70 of adult female age.

Isolation of brood sac secretion and crystallized protein from the embryo midgut

Embryos were removed from the brood sac. Strips of filter paper (Whatman no. 40, Whatman, Kent, UK) were rolled into cylinders and substituted for the embryos within the brood sac for 24 h. The filter papers were then removed, and their content was eluted in water. Females at 59 days of age were used for the analysis of the protein of the brood sac secretion. Midgut crystallized protein was collected from the embryos of a 61-day-old female. Intact guts were isolated from the embryos and transferred to insect Ringer's solution. The gut wall was cut, the crystals were collected and washed six times in Ringer's, centrifuged, and resuspended in water.

Gel electrophoresis of proteins

Protein samples were separated on 15% SDS polyacrylamide gels as described by Hames (1990). Samples containing 5 μg of protein were loaded in each well. The proteins were visualized by staining with Coomassie Blue.

Protein sequencing

Proteins of brood sac secretion and crystallized protein of embryo midgut were separated on a 15% SDS polyacrylamide gel and transferred to a ProBlot membrane (Gooderham 1984). The four Milk bands of the brood sac secretion and one band of midgut crystal protein were excised and their partial N-terminal amino acid (aa) sequences determined by the Edman degradation method (Molecular Analysis Facility, University of Iowa).

Isolation of messenger RNA

Messenger RNA (mRNA) was isolated with a Quick-Prep® Micro-mRNA Purification Kit (Pharmacia Biotechnology, Piscataway, NJ, USA) from the brood sacs of 59-day-old females for identification of 3′ and 5′ cDNA ends. To avoid sequence variation that might be introduced by using mRNA pooled from many individuals, mRNA was isolated from the brood sac of a single 59-day-old female for the final determination of the Milk coding sequences from a single individual. For Northern analysis, mRNA was isolated from brood sacs of 2-day-old virgin females and 14-, 17- to 25-, and 59-day-old mated females and from the testes of 10-day-old males.

Identification of Milk cDNAs

Brood sac mRNA was used for the first-strand synthesis of cDNA with M-MLV reverse transcriptase and an oligo$(dT)_{20}$ primer. For the isolation of *Milk* 3′ cDNA end, a degenerate forward Milk 1 primer and reverse oligo$(dT)_{20}$ primer were used in polymerase chain reaction (PCR) (Table 1, Fig. 1). The first 8 aa obtained from protein sequencing of band 2 were used to design the Milk 1 primer. The following PCR conditions were used: [primer], 1 μM each; [$MgCl_2$], 1.5 mM; [dNTP], 0.2 mM each; final volume, 50 μl. Temperature profiles were as follows: 94°C for 3 min, followed by 34 cycles of 94°C for 1 min, 40°C for 1 min, and 72°C for 1 min with a final 10-min extension at 72°C.

The 5′ cDNA end was obtained with a 5′ RACE System (GibcoBRL, Grand Island, NY, USA) according to the manufacturer's instructions. The first-strand cDNA synthesis was performed with the gene-specific primer, GSP 1 (Table 1), designed using sequence information from cloned 3′ end PCR products. An Oligo(dC) anchor sequence was then attached to the 5′ end of the cDNA. Subsequent PCR was performed with the nested gene-specific primer, GSP 2, and the anchor primer provided with the 5′

Table 1. Primer sequences used for the identification of *Milk* cDNAs and genomic DNA fragment

Primer Name	Primer Sequence
Milk 1	5′-GARAARCCNTGYCCNCCNGARAA-3′
GSP 1	5′-GACGGAACAGATGTGT-3′
GSP 2	5′-GTCTGTAGTGTATGAGAGCAAG-3′
11A	5′-AGATGAAGGTGGTGTTGATCT-3′
12B	5′-ATACTTCTATGTAACTTCATSCC-3′
1A	5′-GTACTAAASACAAGATGAAGGTG-3′
13B	5′-AGAACTTGTTAGSGTGATCATG-3′
10A	5′-GATACCCATGAATAYGAYTCAGAA-3′
4B	5′-CAAAATTCTTGTAGTCCAAACAAAC-3′

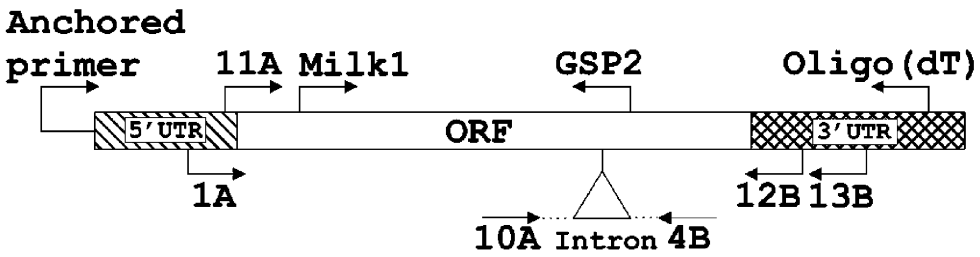

Fig. 1. Diagram of *Milk* cDNA. 3′ cDNA end was obtained by Milk 1 and Oligo(dT)$_{20}$ primers; 5′ cDNA end was obtained by anchored and gene-specific (GSP2) primers. Coding regions of Milk proteins from a single individual were obtained by Milk1/Oligo(dT)$_{20}$ and 11A/ Oligo(dT)$_{20}$ primer sets. Genomic fragment of *Milk* gene was obtained by 1A/13B and 11A/12B primer sets. Genomic *Milk* sequence used as a hybridization probe in Southern analysis was obtained by 10A and 4B primers.

RACE System (Table 1, Fig. 1). The amplification conditions were 94°C for 3 min, followed by 34 cycles of 94°C for 1 min, 43°C for 1 min, and 72°C for 1 min with a final 10-min extension at 72°C.

Milk coding sequences from a single individual were obtained by PCR performed with two sets of primers, Milk 1/ Oligo(dT)$_{22}$ and 11A/Oligo(dT)$_{22}$ (Table 1, Fig. 1). Primer 11A was designed using sequence information obtained by 5′ RACE. The amplification conditions were 94°C for 3 min, followed by 30 cycles of 94°C for 1 min, 42°C for 1 min, and 72°C for 1 min with a final 10-min extension at 72°C.

Northern analysis

Approximately 1 μg of mRNA isolated as described above was separated on a polyacrylamide gel and transferred to a nylon membrane (Sambrook et al. 1989). The RNA was cross-linked by ultraviolet irradiation and then prehybridized at 60°C for 1 h with 0.25 M Na_2HPO_4 (pH 7.4), 1 mM EDTA, 1% bovine serum albumin, and 0.7% SDS. Hybridization was carried out overnight at 60°C in the same buffer with [α-^{32}P]dCTP-labeled probe. The probe was prepared by random-primed labeling using a 700-bp Milk1/Oligo(dT)$_{20}$ PCR fragment (Fig. 1). After hybridization, the membrane was washed in 2× SSC, 0.1% SDS for 30 min at 60°C, followed by another 30-min wash in 0.1× SSC, 0.1% SDS at 60°C, and then used for autoradiography.

Phylogenetic analysis

The Milk protein data set was used for the phylogenetic analyses. Amino acid sequences were manually aligned using the SeqApp program (Gilbert 2002). A phylogenetic tree was inferred using unweighted maximum parsimony (MP) and 164 aligned positions (alignment available upon request from D. B.). Heuristic searches with the data set were done with PAUP* (Swofford 2002) using the tree bisection-reconnection branch-swapping algorithm to find the shortest trees. The number of random-addition replicates was set to 1000 for the tree search. The stability of monophyletic groups in the MP tree was assessed with 2000 bootstrap replicates using the settings described above except that 10 random addition replicates were used for each bootstrap data set. We also did Bayesian analysis of the Milk data (MrBayes V3.0b4; Huelsenbeck and Ronquist 2001) using the WAG+Γ model (Whelan and Goldman 2001). Metropolis-coupled Markov chain Monte Carlo from a random starting tree was initiated in the Bayesian inference and run for 2 million generations with trees sampled every 1000th generation. After discarding the first 1000 trees, a consensus tree was made with the remaining 1000 (i.e., post "burn-in") phylogenies sampled to determine the posterior probabilities (BPs) at the different nodes.

Identification of the exon/intron structure of one Milk locus

Genomic DNA was isolated from embryos of 59-day-old females using the GenomicPrep™ Cell and Tissue DNA Isolation Kit (Amersham Pharmacia Biotechnology, Piscataway, NJ, USA). The genomic sequence of one *Milk* locus was obtained by a PCR performed with Platinum®*Pfx* DNA polymerase (Invitrogen, Carlsbad, CA, USA) and two sets of nested gene-specific primers, 1A/13B and 11A/12B (Table 1, Fig. 1) using genomic DNA as the template. The amplification conditions were 94°C for 3 min, followed by 34 cycles of 94°C for 30 sec, 50°C for 30 sec, and 68°C for 4.5 min with a final 10-min extension at 68°C.

Southern analysis

Hybond-N$^+$ nylon membrane containing blotted genomic DNA that had been digested with *Eco*RI was kindly provided by Dr. William Bendena (Queen's University, Kingston, Ontario, Canada). Genomic DNA was isolated from whole adult males. A random-primed ^{32}P-labeled genomic PCR fragment containing an intron was obtained with the 10A/4B primers and used as the hybridization probe (Table 1, Fig. 1). After hybridization for 22 h at 60°C, the membrane was washed with 2× SSC, 0.1% SDS at 60°C, followed by two washes with 0.1× SSC, 0.1% SDS at 60°C and autoradiography.

Cloning procedures, DNA sequencing, and sequence analysis

PCR products were cloned using either the TA Cloning® Kit (Invitrogen) or the pGEM®-T Vector System (Promega, Madison, WI, USA) according to the manufacturer's instructions. The

sequences of cloned PCR products were determined using an ABM-3700 Sequencer at the University of Iowa DNA Facility. DNA sequences were manually aligned using the SeqApp computer program. ClustalW was used to machine-align the Milk and lipocalin proteins followed by manual optimization using SeqApp.

RESULTS

Milk proteins from the brood sac secretion and embryo midgut

Milk proteins from the brood sac secretion of a 59-day-old female were chosen for detailed analysis. At this stage Milk proteins are abundant and can be isolated on filter papers without damage to the brood sac and contamination of the filter paper with the hemolymph. Separation of brood sac proteins on 15% SDS-PAGE resolved four major bands of approximate sizes 27, 32, 35, and 37 kDa (Fig. 2A). N-terminal sequences of the four Milk bands shown in Fig. 2B indicate that the four peptides are related. There are 11 invariant and 4 ambiguous amino acid positions. The ambiguity in some positions is indicative of sample heterogeneity, suggesting the presence of multiple related peptides in each band.

Separation of crystallized protein from the embryo midgut by SDS-PAGE resulted in a diffuse single band of size 23 kDa (Fig. 2C). The N-terminal amino acid sequence of this band was identical to the amino acid sequence of the third or fourth band from the brood sac secretion (Fig. 2D). This result clearly shows that the crystallized protein found in the midgut of the embryos is one of the Milk proteins present in the brood sac secretion.

Milk cDNAs and deduced amino acid sequences

The sequences of the cloned 3′ cDNA ends each contained the coding region for a C-terminal peptide of length 155 aa and a 3′ untranslated region with a polyadenylation site. The sequences of the cloned 5′ cDNA ends each contained a 5′ untranslated region and a coding region for the signal peptide sequence that was 16 aa in length. To confirm that the multiple sequences of 3′ and 5′ cDNA ends are not the result of allelic variation due to sampling of multiple individuals, reverse transcriptase PCR reactions were repeated using mRNA isolated from a single individual. Sequencing of Milk

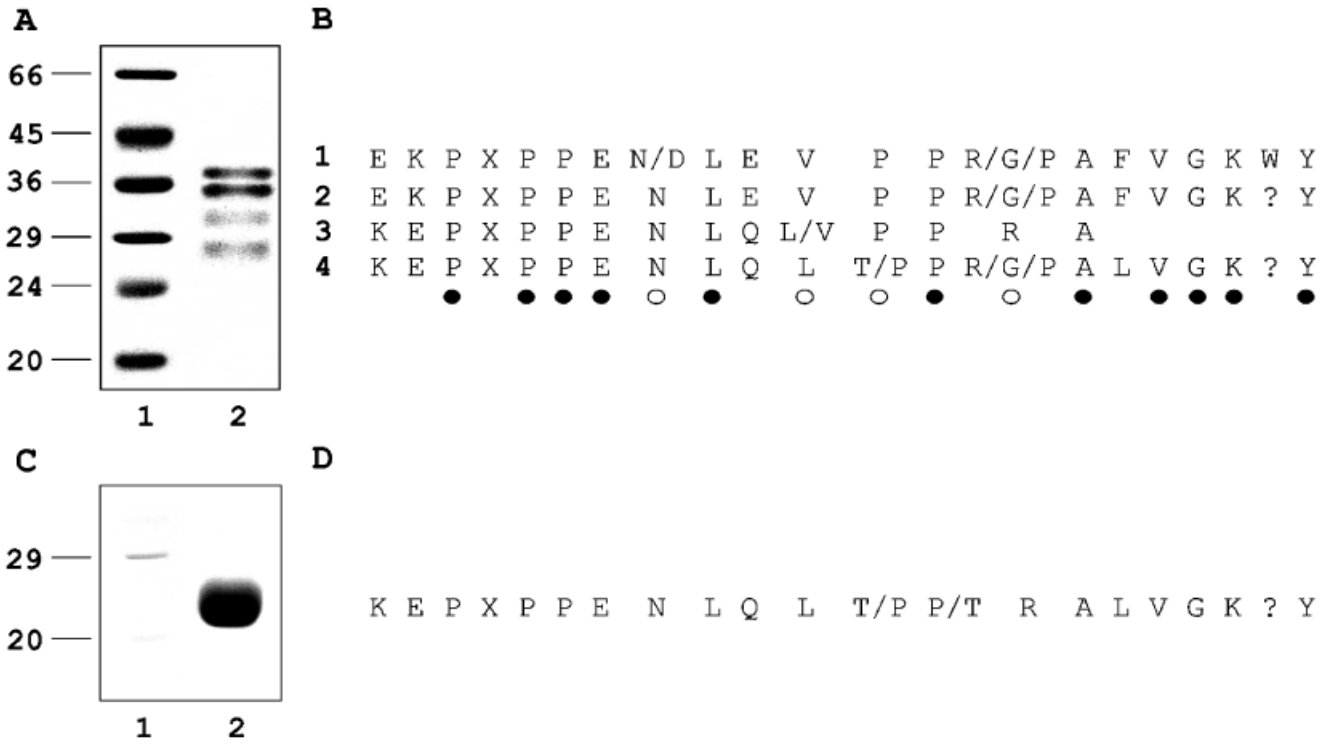

Fig. 2. Milk proteins of the brood sac secretion and embryo midgut. (A) Brood sac secretion proteins collected from a 59-day-old female separated by 15% SDS-PAGE and stained with Coomassie Blue. Lane 1: protein standards; lane 2: Milk proteins. (B) N-terminal amino acid sequence of the four bands (• , invariant positions; ○ , ambiguous positions; X, possibly cysteine; ?, undetermined). (C) Crystallized midgut protein. Lane 1: protein standards; lane 2: gut crystal protein separated by 15% SDS-PAGE and stained with Coomassie Blue. (D) N-terminal amino acid sequence of the crystal protein.

1/Oligo(dT)$_{22}$ cloned PCR product (designated as Z clones) resulted in seven distinct cDNAs encoding a C-terminal 155 aa peptide. Sequencing 11A/Oligo(dT)$_{22}$ cloned PCR products (designated as Y clones) resulted in 18 distinct cDNAs, encoding a 171-aa peptide that included a 16-aa-long signal peptide sequence. The extent of the nucleotide sequence divergence within the coding regions ranges from 0% to 18%. The 25 distinct cDNAs encode 22 distinct peptides (GenBank accession numbers AY447981–AY448005). Nucleotide and deduced amino acid sequences of one of the *Milk* cDNAs (clone 5Y) are shown in Fig. 3. This cDNA encodes a 155-aa mature protein with a signal peptide sequence. The amino acid sequence that matches the amino acid sequence obtained by Edman degradation is in bold text and underlined. This result shows that the cDNA encodes a peptide of the first or second band of the brood sac secretion proteins. Of the 25 cDNAs, 19 encode peptides with N-terminal sequences of the mature protein that match that of bands 1 and 2, 1 cDNA encodes a peptide with an N-terminal sequence that matches that of band 3, and 2 cDNAs encode peptides with N-terminal sequences that match that of band 4. The 3 remaining cDNAs encode peptides with an identical N-terminal amino acid sequence that is closely related to the sequences obtained by direct protein sequencing of the four bands. To demonstrate this relationship, the alignment of the five selected amino acid sequences deduced from cDNAs is shown in Fig. 4. This figure also shows eight conserved regions found in all 22 peptides despite high amino acid sequence divergence (as much as 35% between some peptides). Milk proteins are rich in leucine, valine, asparagine,

```
                                                                  ↓
ATG AAG GTG GTG TTG ATC TTT ATT GCT GCT ATT CTT CTA GCT AAC GCA GAA AAA CCT      57
 m   k   v   v   l   i   f   I   A   A   I   L   L   A   N   A  E   K   P       3

TGT CCT CCA GAA AAT TTA GAA GTT CCT CCT AGA GCG TTT GTA GGG AAG TGG TAC CTC     114
 C   P   P   E   N   L   E   V   P   P   R   A   F   V   G   K   W   Y   L      22

CTA TCT GCG AGT CCT GAC ATT TTC GAG AAT CTG ACC AAC ATC ACA GAA GTT TAT ATC     171
 L   S   A   S   P   D   I   F   E   N   L   T   N   I   T   E   V   Y   I      41
                               ---------   ---------

CCT CGA GGC GAT GAT TAC ATT GGA AAT ATA ACA GTA GTC TCT CCT CAA TAT GGA CAA     228
 P   R   G   D   D   Y   I   G   N   I   T   V   V   S   P   Q   Y   G   Q      60
                           ----------

GAG ACA CAT CGT GTT AAC TTG ACA TTT TCA GGA AAA ACT CTG AAA CTT AAA ATC AAT     285
E   T   H   R   V   N   L   T   F   S   G   K   T   L   K   L   K   I   N      79
                    ----------                                          ---

GAT ACC CAT GAA TAT GAT TCA GAA AGT CGG ATA CTT GCA GTT GAT AAG GAC TAT GTC     342
 D   T   H   E   Y   D   S   E   S   R   I   L   A   V   D   K   D   Y   V      98
 ------

ATC AGC TAT GTA TAT CCT GAA GCT GAC CCC AGA GGC ATT GCT CTC ATA CAC TAC AGA     399
I   S   Y   V   Y   P   E   A   D   P   R   G   I   A   L   I   H   Y   R     117

CAA CCA TTT CCT AAA GAA GAT GTT ATA AAG AGG GTG AAG AAA GCT CTC AAG GAT GTT     456
 Q   P   F   P   K   E   D   V   I   K   R   V   K   K   A   L   K   D   V     136

TGT TTG GAC TAC AAG TAT TTT AGT AAT GAT ACA TCT GTT CCT TCT GAT TAT GTG GAA     513
 C   L   D   Y   K   Y   F   S   N   D   T   S   V   P   S   D   Y   V   E     155
                               ----------

TGA AGTTACATAGAAGTATGGATTTATGAATGATTTTAGAATAATTCATGATCACTCTAACAAGTTCTTTTTAT     587
 *
GTAAAAATTATTGATTTGATGAAACCAAAAGAATGTGACAAATCATGTAACTAAATATAATAAAATTTAACGAAC     662
ATAAAAAAAAAAAAAAAAAA                                                         682
```

Fig. 3. Nucleotide and deduced amino acid sequence of one *Milk* cDNA (clone 5Y). The amino acid sequence is numbered from the start of the mature peptide. The 11A primer sequence is underlined, and corresponding amino acids are shown in small letters. The signal peptide cleavage site is indicated by an arrow. The N-terminal sequence of mature protein determined by protein sequencing is shown in bold and is underlined. Six potential glycosylation sites are underlined with dots. Stop codon is marked by an asterisk, polyadenylation sequence is shown in bold.

```
                    1              2     3
15Y(band 1)  EKPCPPEDLE VPPRAFVGKW YLRSASPDIF ENLTNITEVY IPRGDDYIGN      50
5Y (band 2)  EKPCPPENLE VPPRAFVGKW YLLSASPDIF ENLTNITEVY IPRGDDYIGN      50
23Y(band 3)  KEPCPPENLQ LPPRALVGKW YLRTTSPDIF KQVSNITEVY SAHGNDYYGN      50
19Y(band 4)  KEPCPPENLQ LTPRALVGKW YLRTTSPDIF KQVSNITEFY SAHGNDYYGT      50
3Y (new)     QKPCPPENVK VPPRALVGKW YLRSASPDIF EQVSNVTEVY SAHGNHYYGN      50

                          4                                5
15Y          ITVVSPQYGL ETHRVNLTLS GKTLKLKIND THEYDSESQI LAVDKDYVIS      100
5Y           ITVVSPQYGQ ETHRVNLTFS GKTLKLKIND THEYDSESRI LAVDKDYVIS      100
23Y          VTDYTPKHGL ETHQVNLTVS GSTLKFKMND THDYGTDYQI LAVDKDYFIF      100
19Y          VTDYSPEYGL EAHRVNLTVS GRTLKFYMND THEYDSEYEI LAVDKDYFIF      100
3Y           VTDVSHKHGP ETHRVNLTVS GSTLKFKMND THDYDTDYQI LAVDKDYFIF      100

                          6    7                   8
15Y          YGHPPAVPSG LALIHYRQPF PKEDVIKRVK KALKDVCLDY KYFGNDTSVP CDYMV 155
5Y           YVYPEADPRG IALIHYRQPF PKEDVIKRVK KALKDVCLDY KYFSNDTSVP SDYVE 155
23Y          YGYPKAIPSG LGLVHYRQPC PKEDVIKRVK KNLKNVCLDY KYFSKDTSVH CHSME 155
19Y          YGHPPAAPSG LALIHYRQSC PKEDIIKRVK KSLKNVCLDY KYFGNDTSVH CRYLE 155
3Y           YGYPKAIPSG LALIHYRQPC PKEDILKKVK KDLKNVCLDY KNFGNDTSVH CRYLE 155
```

Fig. 4. A subset of mature Milk proteins (155 aa) deduced from cDNA sequences. These sequences show all the N-terminal sequences found in deduced amino acid sequences. The first four peptides have N-terminal sequences that match the N-terminal sequences obtained by protein sequencing of the four bands. Eight regions that are conserved in all 22 peptides are shown in bold.

lysine, proline, and tyrosine residues, each accounting for 7–8% of the amino acids, but are low in methionine and tryptophan residues (approximately 0.5% each).

All identified *Milk* cDNAs encode a mature protein of length 155 aa with a predicted molecular weight of about 16 kDa. However, the apparent molecular weight of Milk proteins determined by SDS-PAGE ranges from 27 to 37 kDa. This discrepancy in size can be at least partially accounted for by the differential glycosylation of Milk proteins. As detected by periodic acid–Schiff stain, Milk proteins are glycosylated (data not shown) and *Milk* cDNAs contain four to six potential glycosylation sites (Fig. 3) that might be differentially glycosylated depending on protein folding.

Northern analysis

To determine the stage and tissue specificity of *Milk* gene expression, Northern analysis was performed with mRNA isolated from the brood sacs of mated females of different ages, virgin females, and from the testes of adult males (Fig. 5). *Milk* transcripts are first detected in the brood sac of a 20-day-old female, a day before embryos are able to drink. No transcripts are found in the brood sac of virgin females or in the testes (Fig. 5). It should be noted that transcripts of only one size (approximately 750 bp) are detected, indicating that *Milk* cDNAs are homogeneous in size. The slight variation in mRNA loading levels (see Fig. 5, top) does not alter these conclusions.

Phylogenetic analysis

MP trees were inferred from the Milk protein alignment. One of the 105 equally parsimonious trees (length 171 steps, consistency index = 0.69) found with the MP analysis is shown in Fig. 6. Milk proteins are resolved into two distinct clades (I and II, bootstrap support [BS] = 100%, BP = 1.00). Clade I is likely divisible into three subclades (Ia, Ib, and Ic). Subclades Ia and Ib form a monophyletic group that is strongly supported by the bootstrap (BS = 85%) and Bayesian analyses (BP = 1.00). The monophyly of Ib is strongly supported by the Bayesian analysis (BP = 0.99), whereas subclade Ia has moderate bootstrap support (BS = 71%). The monophyly of subclade Ic is only moderately supported by the bootstrap analysis (BS = 76%).

Exon/intron structure of one Milk locus

The presence of introns was detected by PCR reactions with various primers using genomic DNA and cDNA as templates. Based on the sizes of the PCR products, the length of the genomic *Milk* fragment was estimated to be about 4 kb. The 4-kb genomic fragment obtained with use of the 11A and 12B PCR primers (Fig. 1) was cloned and completely sequenced (W3 clone; GenBank accession number AY448006). The

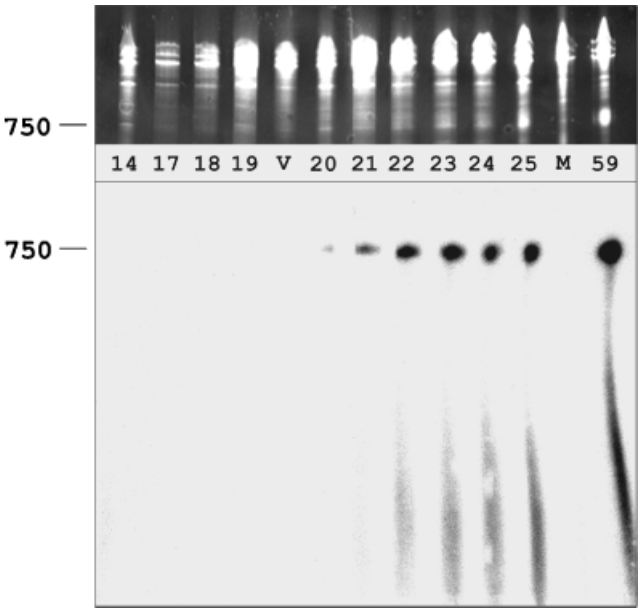

Fig. 5. Northern blot analysis. (Top) mRNA from brood sacs of mated and virgin females and of testes separated on polyacrylamide gel and stained with ethidium bromide. (Bottom) Autoradiographic detection of *Milk* transcripts by ^{32}P-labeled probe (700-bp cDNA fragment). Numbers indicate the age of mated females from which mRNA was isolated; V-mRNA from virgin females; M-mRNA from testes. RNA size (bp) is indicated on the left.

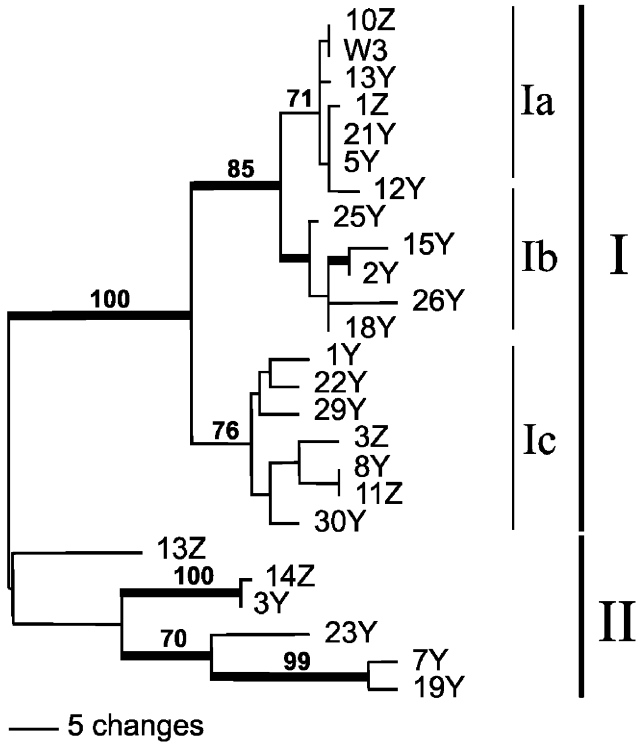

Fig. 6. Phylogeny of Milk proteins. One of 105 equally parsimonious MP trees is shown. The numbers above the branches indicate MP bootstrap values (2000 replicates). The thick branches denote more than 95% posterior probability determined by Bayesian analysis.

sequenced fragment of the *Milk* gene contains five exons and four introns (Fig. 7). The total length is 4168 bp, including 513 bp of the coding sequence distributed among five exons. The coding sequence of the W3 clone is identical to the 2Z clone obtained with the Milk1/Oligo(dT)$_{22}$ primers.

Southern analysis

Genomic DNA was isolated from whole adult males and digested with the *Eco*RI restriction endonuclease. The 10A/4B fragment (930 bp, and see Fig. 1), amplified from the 4-kb fragment, was used as the hybridization probe. This probe hybridizes to at least eight bands varying in size from 2 to 10 kb (Fig. 8).

DISCUSSION

Milk proteins

A series of previous studies, together with the data presented here, lead us to conclude that the proteins analyzed in this article are ingested by the embryos and used for growth. First, Stay and Coop (1973) showed that embryos begin to increase in dry weight on day 23 of the adult female age. Evans and Stay (1989) then demonstrated that Milk proteins first appear in the brood sac tissue of 20-day-old females, suggesting that the increase in dry weight of embryos is due to the uptake of proteins secreted by the brood sac wall. The study of *Milk* gene expression reported here is consistent with this result, demonstrating that *Milk* transcripts first appear in the brood sacs of 20-day-old females. At this time, embryos have reached the stage of dorsal closure and are capable of drinking; just 2 or 3 days later, the increase in dry weight becomes detectable (Stay and Coop 1973). This correlation between developmental stages of the embryos, their increase in dry weight, and the presence of Milk in the brood sac strongly supports the idea that the proteins analyzed here are the proteins responsible for embryo growth. Finally, direct evidence that the brood sac secretion is ingested by the embryos is provided by the finding that the N-terminal sequence of the fourth band of the brood sac secretion is identical to the N-terminal sequence of the protein found in crystallized form inside the embryo gut. The shift in size of the crystallized protein from the embryo midgut indicates that Milk proteins undergo modifications before crystallization either by degradation or partial carbohydrate removal.

The analysis of the amino acid composition of Milk proteins deduced from the cDNA sequences reveals that Milk proteins contain all the essential amino acids, an important characteristic for proteins that function as nutrients. How-

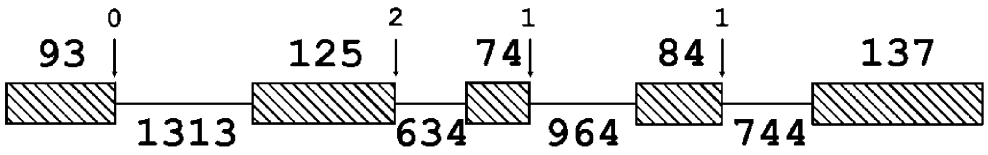

Fig. 7. Exon/intron structure of one *Milk* locus. Shaded boxes indicate exons; lines indicate introns. Numbers above boxes indicate exon length (nts). Numbers below lines indicate intron length (nts). Arrows with numbers above them indicate intron phases.

ever, tryptophan and methionine are present in very low amounts, suggesting that these amino acids are acquired by embryos from another source, possibly bacterial endosymbionts (Gilmour 1965). An unexpected feature of the nutritional proteins (if their nutritional value is evaluated solely on the basis of the presence of essential amino acids) is the presence of eight conserved regions in Milk proteins. These conserved regions point to the existence of structural constraints suggesting that Milk proteins might have a function in addition to nutrition.

Milk gene family

The first indication that Milk proteins comprise a gene family came from the finding that the N-terminal amino acid sequences of the four bands share 57% sequence identity. Conclusive evidence for the existence of a *Milk* gene family is provided by the determination of 25 distinct cDNAs encoding 22 distinct proteins. The Southern blot analysis confirms the presence of multiple loci, although we detect only eight to nine bands. Some bands show relatively stronger hybridization signals, suggesting the existence of multiple target sequences within these fragments. This is likely if *Milk* genes are organized in tandem repeats. In addition, not every gene locus is detected by Southern analysis because the hybridization probe we used recognizes an intron and this sequence is not expected to be conserved in all *Milk* loci. It is also possible that some cDNAs represent allelic variation rather than different loci. Although the exact number of copies cannot be established from the present data, it is clear that *Milk* genes exist in multiple copies. The phylogenetic tree shows that Milk proteins form two distinct clades, I and II (Fig. 6). Many Milk sequences within clade I (Ia) have very short branch lengths, indicating a recent origin through a series of gene duplication events. Milk proteins within clade II are relatively more divergent with long branches, suggesting an earlier origin or elevated mutation rates. Multiple copies of *Milk* genes likely ensure the supply of a large amount of this protein that is critical for embryo development. In fact, in the late stages of the gestation period, 400 μg of protein can be collected from the filter paper in a 24-h period (Stay and Coop 1974). This is an underestimate of the amount of protein that is actually synthesized by the brood sac because at late stages of gestation the embryo batch gains about 1 mg of protein in a 24-h period (Stay and Coop 1973).

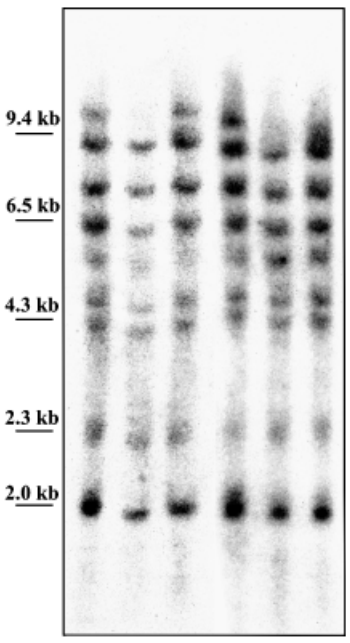

Fig. 8. Southern blot analysis. Genomic DNA from six adult males digested with *Eco*RI and hybridized to ^{32}P-labeled 10A/4B (930 bp) fragment.

Relationship between Milk and lipocalin proteins

A BLAST search (July 2003) against the nonredundant protein database with Milk proteins as a query sequence identified a protein of the aphrodisiac secretion from the cockroach *Leucophaea maderae* as a putative homologue. Although sequence identity and sequence similarity between Milk and the aphrodisiac proteins are low (20% and 45%, respectively), one feature of the *L. maderae* aphrodisiac protein is of interest. This protein is a component of a male tergal gland secretion and is ingested by the female during courtship behavior (Korchi et al. 1999). Both brood sac and tergal glands are modified epidermal cells of the insect integument. The brood sac is an infolding of the intersegmental membrane, and thus Milk proteins are secreted to the outside of the cuticle as are the proteins of the tergal gland. The protein of the aphrodisiac secretion was initially placed

	14	89	114
18Y	RAFVGKWYLLSASP	QILAVD-KDYVISY	IHY--RQPFPKE
Gallerin	TAYQGVWYEISKTP	NVIATDYQNYAIAY	ILS--RAKKLEG
Bilin-BP	SNYHGKWWEVAKYP	NVLSTDNKNYIIGY	VLS--RSKVLTG
Apolipoprotein D	NKYLGRWYEIEKIP	WILATDYENYALVY	ILA--RNPNLPP
Bla g 4	ERFRGSWIIAAGTS	SVLATDYENYAIVE	IRFSVRRFHPKL
Insecticyanin	SAFAGAWHEIAKLP	WVLATDYKNYAINY	ILS--KSKVLEG
	----SCR 1-----	-----SCR 2----	----SCR 3---

Fig. 9. Amino acid alignment of a Milk protein (18Y clone) and lipocalin proteins. The portion of the alignment corresponding to SCR1, SCR2, and SCR3 is shown. Gallerin, *Galleria mellonella* (greater wax moth), AAA85089; Bilin-BP, *Pieris brassicae* (cabbage butterfly), 1BBPA; Apolipoprotein D, *Homo sapiens* (human), AAB32200; Bla g 4, *Blattella germanica* (German cockroach), P54962; Insecticyanin, *Manduca sexta* (tobacco hornworm), Q00630.

into the lipocalin protein family and later in the broader calycin protein family (Korchi et al. 1999; Cornette et al. 2001). To determine whether Milk proteins might belong to the lipocalin family, we identified features common to both protein groups. Close examination of Milk amino acid sequences revealed the existence of the three structurally conserved regions (SCRs) diagnostic of the lipocalin protein family (Flower 1996). Lipocalins are small (150–200 aa) proteins that often function in the transport of small hydrophobic molecules, but various other functions have also been identified (Flower et al. 2000; Lögdberg and Wester 2000). Members of the lipocalin family show low sequence identity (approximately 20% in pairwise comparisons) but are structurally conserved. The lipocalin fold consists of a single eight-stranded antiparallel β-sheet that folds onto itself to form a β-barrel that contains a ligand-binding site (Flower 1996). Importantly, regions of Milk proteins corresponding to SCR1, SCR2, and SCR3 are conserved in all 22 peptides (regions 2, 5, and 6 in Fig. 4). The portion of the alignment of Milk proteins and some lipocalin proteins containing SCRs is shown in Fig. 9. The glycine and tryptophan residues are the most highly conserved residues in SCR1 among lipocalins. These amino acids contribute to the stability of the protein structure (Gasymov et al. 1999; Greene et al. 2001). The interaction of this tryptophan with a conserved arginine residue in SCR3 facilitates the formation of the β-barrel in the retinol-binding protein (Greene et al. 2001). The arginine residue from SCR3 is also conserved in Milk proteins (Fig. 9). Although sequence identity within structurally conserved regions is only moderate, conservation of the most important SCR1 and SCR3 residues as well as the relative positions of the three conserved motifs is unlikely to have arisen through convergence.

To further investigate the relationship of Milk proteins to lipocalins, the exon/intron structure was examined for one of the *Milk* loci. The comparison of exon/intron structures and the locations of SCRs of arthropod and vertebrate lipocalins are shown in Fig. 10 (Salier 2000). The arthropod lipocalin,

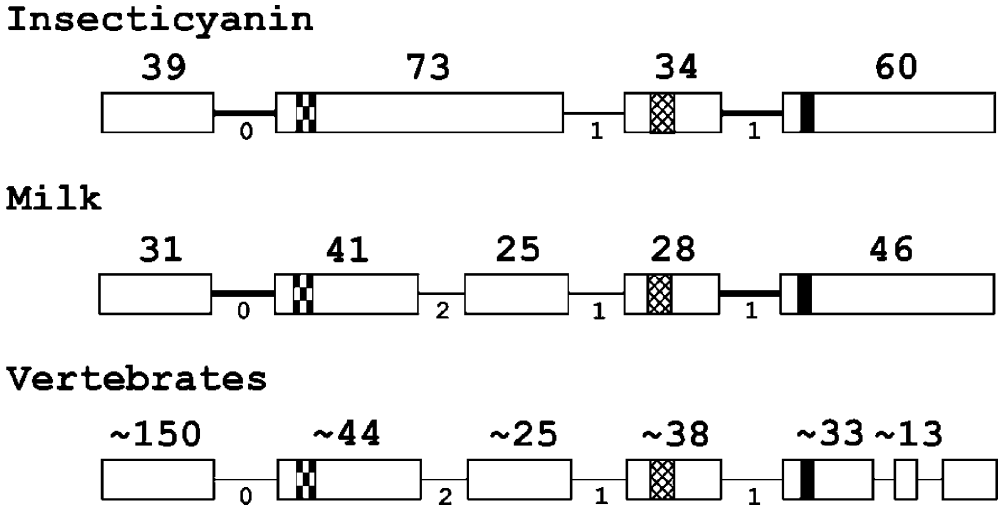

Fig. 10. Comparison of intron/exon structure of *Milk* genes with that of arthropod and vertebrate lipocalin genes (modified from Salier 2000). Conserved intron positions are shown with thick lines. Numbers below lines indicate intron phases. Numbers above exons indicate number of amino acids encoded by each exon. Checkered bar, SCR1; cross-hatched bar, SCR2; solid bar, SCR3.

insecticyanin, is a pigment protein found in the epidermis and the hemolymph of *Manduca* larvae that exists in two isoforms encoded by two genes (Li and Riddiford 1992). Insecticyanin genes contain four exons and three introns. The coding region of the *Milk* gene is interrupted by four introns. The positions and phases of the first and last introns of the two genes are conserved. Vertebrate lipocalin genes generally contain seven exons, but the SCRs are located within exons 2, 4, and 5 as in Milk proteins. The conservation of intron phases is apparent between Milk, insecticyanin, and vertebrate lipocalins. In summary, the following features shared by Milk and lipocalin proteins suggest that Milk proteins belong to the lipocalin protein family: (a) the size of the primary transcript of the *Milk* gene is about 4.2 kb, falling within the range found for lipocalin genes (3–6 kb); (b) the size of the Milk proteins is 171 aa, falling within the range found for lipocalin proteins (150–200 aa); (c) there are three SCRs; and (d) there is conservation of intron positions and intron phases. Another important piece of evidence supporting the evolutionary relationship between lipocalins and Milk proteins will come from the determination of the crystal structure of Milk proteins, which is currently underway.

Evolution of Milk proteins

Roth (1970) suggested that viviparous species might have evolved from ovoviviparous species of cockroaches that retain embryos within the brood sac but provide no nutrients to the developing embryos. This hypothesis is supported by the phylogenetic studies based on the molecular data that place *D. punctata* within the ovoviviparous clade (Kambhampti 1995). For viviparity to evolve, changes in brood sac morphology must occur before the evolution of nutrient secretion. Epithelial cells of the intersegmental membrane must accommodate secretory and duct cells to become functional glands. Importantly, the morphology of the brood sacs of the two ovoviviparous species *Byrsotria fumigata* and *Gromphadorhina portentosa* have been shown to be similar to that of *D. punctata* with epithelial cells equipped with a secretory apparatus and ducts leading to the cuticular surface of the brood sac (Snart et al. 1984a). Moreover, the presence of some secretory material on the brood sac surface has been demonstrated by scanning electron microscopy (Snart et al. 1984b). The presence of the secretory material does not imply that this material has nutritive function, but this observation clearly indicates that at least one of the requirements for the evolution of viviparity, functional morphology of the brood sac, has been established in ovoviviparous cockroach species. Neither the composition nor the function of the ovoviviparous brood sac secretion is currently known but deserves thorough investigation for a more complete understanding of the evolution of the nutritive function of Milk proteins.

The relationship between Milk proteins, proteins of the aphrodisiac secretion of the ovoviviparous *L. maderae*, and lipocalins suggests that the ancestor of Milk proteins might have been a ubiquitously expressed cuticular surface lipid carrier protein that underwent functional divergence associated with the ability to bind different ligands in different expression domains (tergal gland and the brood sac). The possibility that Milk proteins might function as carriers of hydrophobic molecules is supported by the observation that the brood sac secretion contains large amounts of cholesterol (Ingram et al. 1977). Lipocalin proteins are known to bind a variety of hydrophobic molecules, including cholesterol (Glasgow et al. 1995; Rassart et al. 2000). Moreover, cholesterol cannot be synthesized by insects de novo and must be acquired from the diet or in the form of a sterol precursor (Rees 1985). Thus, we suggest that Milk proteins may also function in cholesterol transport to the embryos (Stay, unpublished data).

We speculate that the initial brood sac secretions contained proteins capable of binding hydrophobic molecules but were limited in quantity and functioned mainly in embryo protection. Once embryos evolved the ability to drink, they began to ingest the materials secreted by the brood sac wall (there is no known mechanism for selective drinking). These secretions must have in some way benefited embryos, most likely as additional nutrients supplemented with cholesterol. Continuous removal of the secretion by embryos could maintain increased production of protein provided by the increased number or expression levels of *Milk* genes.

We believe that *Milk* genes arose through the modification of lipocalin-like genes that already existed in ovoviviparous species. This situation represents an example of co-option of pre-existing genes for a novel function (True and Carroll 2002). Milk proteins are produced in large amounts by the brood sac to serve a nutritive function that is unrelated to their original function in ovoviviparous species. However, such co-option would be impossible without the embryo's ability to drink. Thus, the developmental/physiological changes in the embryos direct the functional shift from nonnutritive to nutritive proteins and constitute a selective pressure to maintain duplicated *Milk* genes. Under this scenario, gene co-option is expected to precede most duplication events of *Milk* genes. We also note that acquisition of a novel function does not need to be associated with the change in the coding sequence or the protein expression pattern provided that a new interacting component appears (be it another molecule or, in our case, drinking embryos). These observations are consistent with the models based on multifunctionality of proteins where acquisition of a novel function precedes gene duplication (Piatigorsky and Wistow 1991; Hughes 1994; Ganfornina and Sanchez 1999).

In summary, we present here the case of a multigene (*Milk*) family that encodes proteins with a recently evolved nutritive

function. This function is associated with the appearance of a novel trait, viviparity, in the cockroach *D. punctata*. Additional comparative analyses using closely related cockroach species with different reproductive strategies will be required to elucidate further the mechanism underlying this evolutionary novelty.

Acknowledgments

We thank Drs. William Bendena, Hwan Su Yoon, Mei-Yeh Lu, Gery Hehman, Josep Comeron, and Ms. Dawn Simon for their helpful discussions and assistance with experiments. We also thank two anonymous reviewers for their valuable suggestions. This work was supported by grants from the National Institutes of Health (grant AI 15320) and Sigma Xi.

REFERENCES

Cornette, R., Farine, J.-P., Quennedey, B., and Brossut, R. 2001. Molecular characterization of a new adult male putative calycin specific to tergal aphrodisiac secretion in the cockroach *Leucophaea maderae*. *FEBS Lett*. 507: 313–317.

Evans, L. D., and Stay, B. 1989. Humoral induction of milk synthesis in the viviparous cockroach *Diploptera punctata*. *Invertebr. Reprod. Dev*. 15: 171–176.

Flower, D. R. 1996. The lipocalin protein family: structure and function. *Biochem. J*. 318: 1–14.

Flower, D. R., North, A. C., and Sansom, C. E. 2000. The lipocalin protein family: structural and sequence overview. *Biochim. Biophys. Acta* 1482: 9–24.

Ganfornina, M. D., and Sanchez, D. 1999. Generation of evolutionary novelty by functional shift. *BioEssays* 21: 432–439.

Gasymov, O. K., Abduragimov, A. R., Yusifov, T. N., and Glasgow, B. J. 1999. Binding studies of tear lipocalin: the role of the conserved tryptophan in maintaining structure, stability and ligand affinity. *Biochim. Biophys. Acta* 1433: 307–320.

Gilbert, D. G. 2002. SeqApp. A Macintosh biosequence editor, analyzer, and network handyman. Available from ftp://iubio.bio.indiana.edu/molbio/seqapp/.

Gilmour, D. 1965. *The Metabolism of Insects*. Oliver and Boyd Ltd., Edinburgh, UK, pp. 115–147.

Glasgow, B. J., Abduragimov, A. R., Farahbakhsh, Z., Faull, K. F., and Hubbell, W. L. 1995. Tear lipocalins bind a broad array of lipid ligands. *Curr. Eye Res*. 14: 363–372.

Gooderham, K. 1984. Transfer techniques in protein blotting. In J. Walker (ed.). *Methods in Molecular Biology*. Humana Press Inc., Clifton, NJ, pp. 165–178.

Greene, L. H., Chrysina, E. D., Irons, L. I., Papageorgiou, A. C., Acharya, K. R., and Brew, K. 2001. Role of conserved residues in structure and stability: tryptophans of human serum retinol-binding protein, a model for the lipocalin superfamily. *Protein Sci*. 10: 2301–2316.

Gu, Z., Cavalcanti, A., Chen, F.-C., Bouman, P., and Li, W.-H. 2002. Extent of gene duplication in the genomes of *Drosophila*, nematode, and yeast. *Mol. Biol. Evol*. 19: 256–262.

Hames, B. D. 1990. *Gel Electrophoresis of Proteins: A Practical Approach*. 2nd ed. Oxford University Press, New York.

Huelsenbeck, J. P., and Ronquist, F. 2001. MrBayes: Bayesian inference of phylogeny. *Bioinformatics* 17: 754–755.

Hughes, A. L. 1994. The evolution of functionally novel proteins after gene duplication. *Proc. R. Soc. Lond. B*. 256: 119–124.

Ingram, M. J., Stay, B., and Cain., G. D. 1977. Composition of milk from the viviparous cockroach, *Diploptera punctata*. *Insect Biochem*. 7: 257–267.

Kambhampti, S. 1995. A phylogeny of cockroaches and related insects based on DNA sequence of mitochondrial ribosomal RNA genes. *Proc. Natl. Acad. Sci. USA* 92: 2017–2020.

Korchi, A., Brossut, R., Bouhin, H., and Delachambre, J. 1999. cDNA cloning of an adult male putative lipocalin specific to tergal gland aphrodisiac secretion in an insect (*Leucophaea maderae*). *FEBS Lett*. 449: 125–128.

Li, W., and Riddiford, L. M. 1992. Two distinct genes encode two major isoelectric forms of insecticyanin in the tobacco hornworm, *Manduca sexta*. *Eur. J. Biochem*. 205: 491–499.

Li, W.-H., Gu, Z., Wang, H., and Nekrutenko, A. 2001. Evolutionary analysis of the human genome. *Nature* 409: 847–849.

Lögdberg, L., and Wester, L. 2000. Immunocalins: a lipocalin subfamily that modulates immune and inflammatory responses. *Biochim. Biophys. Acta* 1482: 284–297.

Piatigorsky, J., and Wistow, G. 1991. The recruitment of crystallins: new functions precede gene duplication. *Science* 252: 1078–1079.

Rassart, E., Bedirian, A., Do Carmo, S., Guinard, O., Sirois, J., Terrisse, L., and Milne, R. 2000. Apolipoprotein D. *Biochim. Biophys. Acta* 1482: 185–198.

Rees, H. H. 1985. Biosynthesis of ecdysone. In G. A. Kerkut and L. I. Gilbert (eds.). *Comprehensive Insect Physiology, Biochemistry and Pharmacology*. Vol. 7. Pergamon, New York, pp. 249–293.

Reznick, D. N., Mateos, M., and Springer, M. S. 2002. Independent origins and rapid evolution of placenta in the fish genus *Poeciliopsis*. *Science* 298: 1018–1020.

Roth, L. M. 1970. Evolution and taxonomic significance of reproduction in Blattaria. *Annu. Rev. Ent*. 15: 75–96.

Roth, L. M., and Willis, E. R. 1954. The reproduction of cockroaches. *Smithson. Misc. Coll*. 122: 1–49.

Roth, L. M., and Willis, E. R. 1955. Intra-uterine nutrition of the "beetle-roach" *Diploptera dytiscoides* (Serv.) during embryogenesis with notes on its biology in the laboratory. *Psyche* 62: 55–68.

Salier, J.-P. 2000. Chromosomal location, exon/intron organization and evolution of lipocalin genes. *Biochim. Biophys. Acta* 1482: 25–34.

Sambrook, J., Fritsch, E. F., and Maniatis, T. 1989. *Molecular Cloning: A Laboratory Manual* 2nd Ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Snart, J. O. H., Greenwood, M., Beck, R., and Highnam, K. C. 1984a. The functional morphology of the brood sac in two species of ovoviviparous cockroaches, *Byrsotria fumigata* (Guerin) and *Gromphadorhina portentosa* (Schaum). 1. Scanning and light microscopy. *Intl. J. Invertebr. Reprod. Dev*. 7: 345–355.

Snart, J. O. H., Greenwood, M., Beck, R., and Highnam, K. C. 1984b. The functional morphology of the brood sac in two species of ovoviviparous cockroaches, *Byrsotria fumigata* (Guerin) and *Gromphadorhina portentosa* (Schaum). 2. Transmission electron microscopy. *Intl. J. Invertebr. Reprod. Dev*. 7: 357–367.

Stay, B., and Coop, A. 1973. Developmental stages and chemical composition in embryos of the cockroach *Diploptera punctata*, with observations on the effect of diet. *J. Insect Physiol*. 19: 147–171.

Stay, B., and Coop, A. C. 1974. "Milk" secretion for embryogenesis in a viviparous cockroach. *Tissue Cell* 6: 669–693.

Swofford, D. L. 2002. *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods)*. 4.0b8. Sinauer, Sunderland, MA.

True, J. R., and Carroll, S. B. 2002. Gene co-option in physiological and morphological evolution. *Annu. Rev. Cell Dev. Biol*. 18: 53–80.

Whelan, S., and Goldman, N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol*. 18: 691–699.