Vitellogenin of *Blattella germanica* (L.) (Dictyoptera, Blattellidae): Nucleotide Sequence of the cDNA and Analysis of the Protein Primary Structure

Daniel Comas, Maria-Dolors Piulachs, and Xavier Bellés*

Department of Physiology and Molecular Biodiversity, Institut de Biologia Molecular de Barcelona (CID, CSIC), Barcelona, Spain

The cloning and sequencing of a cDNA of the vitellogenin gene from the cockroach *Blattella germanica* is reported. It is 5,749 nucleotides long and encodes an amino acid sequence of 1,862 residues (including a putative signal peptide of 17 residues). The vitellogenin sequence includes a long serine-rich stretch between amino acids 322 and 349, and two other stretches between amino acids 1691 and 1740. The vitellogenin of *B. germanica* shows a notable similarity (between 32 and 42%) to those described in other insects, and its alignment shows a high number of motifs conserved in all species, especially in the subdomains I–V. Non-parsimony methods (Neighbor Joining) of phylogenetic analysis of the insect vitellogenin sequences gave a tree showing a topology that is, in general, congruent with the currently accepted insect phylogenetic schemes. Arch. Insect Biochem. Physiol. 45:1–11, 2000.   © 2000 Wiley-Liss, Inc.

Key words: insect; Dictyoptera; cockroach; *Blattella germanica*; vitellogenin

INTRODUCTION

Vitellogenins are proteins that serve as precursors of vitellins, the main egg yolk protein in many oviparous vertebrates and invertebrates. In a great majority of insect species, vitellogenin is synthesized in the fat body, released into the hemolymph, and incorporated into developing oocytes through receptor-mediated endocytosis (Hagedorn et al., 1998; Sappington and Raikhel, 1998). In insects, the endocrine regulation of vitellogenesis has been the object of intensive research, especially in species where the process is ecdysteroid-dependent, like mosquitoes (Hagedorn et al., 1998; Sappington and Raikhel, 1998). Conversely, studies based in less modified species, like cockroaches in which vitellogenesis is juvenile hormone-dependent, are scarcer (see Wyatt and Davey, 1996).

In a previous paper, we reported the isolation and partial sequence of a vitellogenin cDNA on the German cockroach, *Blattella germanica*, and its use as a probe in Northern blot analysis to assess the developmental, tissue and sex specificity of the vitellogenin gene expression in this cockroach (Martín et al., 1998). The same probe was used to study the action of juvenile hormone

upon vitellogenin synthesis in vivo (Comas et al., 1999), in the frame of our investigations on the functional links between the cycle of such a hormone (Bellés et al., 1987; Maestro et al., 1994) and that of vitellogenin (Martín et al., 1995a,b, 1996; see also Bellés et al., 1999).

Besides these functional studies, we have also been interested in the structural and phylogenetical information that may be derived from vitellogenin primary sequences. Therefore, we have completed the study of the vitellogenin cDNA of *B. germanica*, which, in addition to leading to the characterization of the complete deduced amino acid sequence, has allowed comparison with other species in which the orthologous protein has been reported. Up to now, vitellogenin-based phylogenetical analyses in insects have included species of Hemiptera, Hymenoptera, Coleoptera, Lepidoptera, and Diptera (see Lee et al., 2000a, and references therein). The present work incorporates the "primitive" order Dictyoptera to the analysis, including the complete sequence of vitellogenin of *B. germanica* described herein, and that of *Periplaneta americana* sequenced by Tufail et al., (2000). The comparative analysis will show that the insect vitellogenin molecule seems to contain phylogenetic information.

MATERIALS AND METHODS
Insect Rearing and Tissue Collection

Specimens of *B. germanica* were taken from a colony reared in the dark at 30 ± 1°C and 60–70% relative humidity. Dissection of fat bodies and extraction of RNA were carried out as described by Martín et al. (1998), although in this case RNA was purified using a Rneasy mini kit (QIAGEN).

cDNA Cloning and Sequencing

In a previous study (Martín et al., 1998) we obtained a 2,645 nucleotide partial sequence of the vitellogenin cDNA of *B. germanica* corresponding to the 3′ region. In order to complete the 5′ end of this cDNA we followed the "Rapid amplification of cDNA ends" (RACE) method (Frohman et al., 1988) using the Gibco, BRL kit. An amount of 350 ng of total RNA from fat bodies of females between days 3 and 5 of adult life, were used as starting material. Oligonucleotides to be used as primers were designed on the basis of the 5′ end

of the partial sequence and RACE products (see Table 1). PCR conditions were: 94°C for 2 min; 94°C for 30 sec, between 60 and 65°C (depending on the primers) for 30 sec, 72°C for 2 min, for a total of 30–35 cycles; final extension at 72°C for 20 min. PCR products were subjected to electrophoresis on 1.2% agarose gel, purified using a QIAquick PCR Purification kit (QIAGEN) and subcloned in Pgem-T Easy vector (Promega). Sequence analysis was directly performed by the dideoxynucleotide chain termination method. Clones were sequenced on both strands using the SP6 and T7 sequencing primers and internal, specific primers, in an automated fluorescence sequencing system ABI (Perkin Elmer, Norwalk, CT).

Sequence Comparisons and Analysis

Vitellogenin amino acid sequences used for comparison were obtained from literature sources and the GenBank database as follows: silkworm moth, *Bombyx mori* (Yano et al., 1994a,b; D13160); gypsy moth, *Lymantria dispar* (Hiremath and Lehtoma, 1997a,b; V60186); boll weevil, *Anthonomus grandis* (Trewitt et al., 1992; M72980); yellow fever mosquito, *Aedes aegypti* (Chen et al., 1994; U02548); turnip sawfly, *Athalia rosae* (Nose et al., 1997; AB007850); the parasitoid wasp *Pimpla nipponica* (Nose et al., 1997; AF026789); the bean bugs, *Riptortus clavatus* (Hirai et al., 1998; U97277) and *Plautia stali* (Lee et al., 2000a; isoform 1: AB033498, isoform 2: AB033499, isoform 3: AB033500); the cicada *Graptopsaltria nigrofuscata* (Lee et al., 2000b; AB026848), and the American cockroach, *P. americana* (Tufail et al., 2000, AB034804). In the case of *P. stali*, only one isoform (Vg-1) was used, given the high percentage of identity between the three vitellogenin

TABLE 1. Oligonucleotides Used as Primers in 5′ RACE to Obtain the Complete cDNA of *Blattella germanica* Vitellogenin*

Name	Sequence (5′→3′)	Position
GSP11	GAAATTCCTGATCTGT	3277–3262
GSP12	GCCCGTGTTTTCTCCCATTT	3237–3218
GSP21	CCTGTCAAGACCTGAAATGTAT	2991–2970
GSP22	GTTGTCGAAAGCTGTGCTGAA	2964–2944
GSP31	CCTTGGCTTTCCTTCGCATTT	2281–2261
GSP32	CCACATTTGCAGCACTCTTA	2242–2223

*The positions of oligonucleotides are indicated by nucleotide numbers (see Fig. 1).

isoforms known for this species, and to avoid redundancy effects. The software package of the Genetics Computer Group (GCG, version 9.1) of the University of Wisconsin (Devereux et al., 1984) was used for sequence alignments, which were carried out with PILEUP and were not further hand-refined. Percentage similarity and percentage identity between sequences were estimated using the corresponding application of the Multiple Sequence Alignment Editor GeneDoc version 2.5.000 (Nicholas and Nicholas, personal communication, 1997). Phylogenetic analyses were carried out using amino acid sequences, with the Phylogeny Inference Package (Phylip, version 3.57c, Felsenstein, personal communication, 1995). For non-parsimony analyses, we followed the method of Neighbor Joining, and the distances between different vitellogenins were estimated with Kimura's formula and the application of Protdist. Parsimony analyses were carried out with the same Package using the Protpars method. Bootstrap analyses were carried out with the application Seqboot in the Phylip package, and the procedure was repeated 100 times.

RESULTS AND DISCUSSION

Vitellogenin cDNA and Deduced Amino Acid Sequence

The completion of the partial sequence previously reported (Martín et al., 1998), following the 5′ RACE method, yielded a sequence that is 5,749 nucleotides long (Fig. 1). The 5′ region begins with a short (24 bp) untranslated region, which is rather constant in size compared to different insect vitellogenin cDNAs (see, for example, Hirai et al., 1998). The translated region begins with the ATG codon corresponding to the initial methionine, and finishes with the stop codon located 5,611 bp downstream. After the stop codon there is an untranslated region of 139 bp, which contains two fused consensus polyadenylation signals (AATAAATAAA), reminiscent of the two complete and successive polyadenylation signals reported by Hiremath and Lehtoma (1997a) in *L. dispar*.

The deduced amino acid sequence has 1,862 amino acids (Fig. 1), and the calculated molecular weight is approximately 212 kDa, which is in fair agreement with the molecular weight of the pre-pro-vitellogenin estimated from electrophoretic data (Martín et al., 1995a). The first 17 amino acids would correspond to a signal peptide, as suggested by the hydrophobicity analysis (not shown) and by the predictions of the program SignalP V1.1 (Nielsen et al., 1997), and given that they fit the -3, -1 rule of von Heijne (1986). In addition, these 17 first amino acids align well with the signal peptides of other insect vitellogenins (Fig. 2), and show a relatively high percentage of similarity when compared (the highest similarity, 94%, was found between *A. rosae* and *R. clavatus*, and the lowest, 27%, between *B. germanica* and *B. mori*). The similarities in primary structure of these signal peptides (see also Romans et al., 1995) suggest that they are orthologous, and point to a common ancestor signal peptide for insect vitellogenins.

Other remarkable motifs of the amino acid sequence are the polyserine domains. There is a long serine-rich stretch between amino acids 322 and 349, which is 41 amino acids before a consensus RXRR (RPRR) cleavage sequence. There are two shorter serine-rich stretches between amino acids 1691 and 1740. The occurrence of such polyserine domains towards the 3′ region of the sequence is only paralleled in the vitellogenins of *A. aegypti* (Chen et al., 1994; Romans et al., 1995) and of *P. americana* (Tufail et al., 2000). Conversely, serine-rich stretches are quite frequent towards the 3′ end, around the cleavage site (see, for example, Lee et al., 2000b). The most frequent variant is the occurrence of two stretches flanking the cleavage site, as occurs in *B. mori* (Yano et al., 1994b), *A. aegypti* (Chen et al., 1994; Romans et al., 1995), *A. rosae* (Nose et al., 1997), *R. clavatus* (Hirai et al., 1998), *P. stali* (Lee et al., 2000a), and *P. americana* (Tufail et al., 2000). *G. nigrofuscata* has only one serine-rich stretch following the cleavage site (Lee et al., 2000b), and *B. germanica*, in this 3′ region, also has only one stretch, but preceding the putative cleavage site. All these variants in distribution of polyserine domains have been schematically represented in Figure 3. *A. grandis* (which is cleaved in the 5′ region, as usual) (Trewitt et al., 1992), *L. dispar* (which is cleaved in the 3′ region) (Hiremath and Lehtoma, 1997a), and *P. nipponica* (which is not cleaved) (Nose et al., 1997) have no clusters of serine residues at all.

```
   1  AGCAGTTCCACTCCGTGCTCCAACATGACTTGGAACGCTCTACTCTGCTGCCTCCTTGTGAGTGCTGCCTCC
   1                          M  T  W  N  A  L  L  C  C  L  L  V  S  A  A  S ---
                            ‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾

  73  GCAATTACACCAGGATGGCTTCCAATCAACAGCCAGTTAGACTATCATGTACATGGACGAACCTTTTCCTCG
  17  -- A  I  T  P  G  W  L  P  I  N  S  Q  L  D  Y  H  V  H  G  R  T  F  S  S
      ‾‾‾‾‾

 145  CTATTCCAAGTCGCCAATCAATACACGGGAATACTGTACAAGGCAAGACTCTCACTGGACAGGAATGAAGAT
  41  L  F  Q  V  A  N  Q  Y  T  G  I  L  Y  K  A  R  L  S  L  D  R  N  E  D

 217  CAACTCATTACGGGCAAAGTGACAGAGGCACAATTTTCACCAGTGAACACACAATTTAGCTCCGGTTGGGAT
  65  Q  L  I  T  G  K  V  T  E  A  Q  F  S  P  V  N  T  Q  F  S  S  G  W  D

 289  GAATCCGTTCCCGACGAGAAGCTCCACTGGGACGTGGTACCAATGTCACAACAACCTTTCCAAATTGAGCTC
  89  E  S  V  P  D  E  K  L  H  W  D  V  V  P  M  S  Q  Q  P  F  Q  I  E  L

 361  AACAGCAGAGGAGAGGTACGGAAGCTGAGAGTTAACAAATTTGTCGAACTCTGGGAAATCAACATGATAAAG
 113  N  S  R  G  E  V  R  K  L  R  V  N  K  F  V  E  L  W  E  I  N  M  I  K

 433  GCCATTATTAGTCAGCTGCAAGTAGTTGTTGACGAAGACAAAAAAGTTTACAGAGTATTTGAGAGCACAGTA
 137  A  I  I  S  Q  L  Q  V  V  V  D  E  D  K  K  V  Y  R  V  F  E  S  T  V

 505  ACTGGTAGATGTGAAGCTTTGTATGAAGTCGACCATCTTTATCCAACAACATATCTAAATCCATGGCAGTGG
 161  T  G  R  C  E  A  L  Y  E  V  D  H  L  Y  P  T  T  Y  L  N  P  W  Q  W

 577  ACTCAGCAGCATGACACCAAATTGAGAATTATGAAGACCCATCAATTCACAAACTGCCGTCACAACTCAGCT
 185  T  Q  Q  H  D  T  K  L  R  I  M  K  T  H  Q  F  T  N  C  R  H  N  S  A

 649  TACAAACTGCACTTCAATGCTTTCGAATATTTCCATTTGAAACAACACAAACCAGAAACTTTTCTAAGTAAT
 209  Y  K  L  H  F  N  A  F  E  Y  F  H  L  K  Q  H  K  P  E  T  F  L  S  N

 721  TCCGCTGTGAGCAGGGTAATTGCTGATGGTGACAACCTGAAGAACTTCACATTCTATTCTGGAGAAACCATC
 233  S  A  V  S  R  V  I  A  D  G  D  N  L  K  **N**  **F**  **T**  **F**  Y  S  G  E  T  I

 793  CACAAAATAGTCCTGAACCCAGAAATTTACAACAAGCAGAAAGGCATGCTGGTCAGCCACATTAATGTAACG
 257  H  K  I  V  L  N  P  E  I  Y  N  K  Q  K  G  M  L  V  S  H  I  **N**  **V**  **T**

 865  GTAGAAAGGAAGGGAAGAGAACTCACTGTCATCGATTATGAACTAAGAAATGTTGGTGATCTCTCTTACTCA
 281  **V**  E  R  K  G  R  E  L  T  V  I  D  Y  E  L  R  N  V  G  D  L  S  Y  S

 937  ACTAGTTTAGTTAAAGCACATTCGATGAGAAATAGCGCCTCCATGGATCTTAGCAGCAGTAGTATGAGCAGT
 305  T  S  L  V  K  A  H  S  M  R  N  S  A  S  M  D  L  S  S  S  S  M  S  S
                                                          ‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾

1009  AGCAGTAGTAGCAGTAGCAGTAGCAGCAGTAGTAGCAGCAGCAGCAGTAGCAGTAGCAGTGAGGAGCAC
 329  S  S  S  S  S  S  S  S  S  S  S  S  S  S  S  S  S  S  S  S  S  S  E  E  H
      ‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾‾

1081  CATAGCCACAATCAGAAGTTATCCAAGAAGAGACAGGTGCCACTGCCACGGCCATTGTTTGAAGCAAATTTT
 353  H  S  H  N  Q  K  L  S  K  K  R  Q  V  P  L  P  R  P  L  F  E  A  N  F

1153  GACGCCAGTTCTGGATTAACCACAGAACAACCTGTCACATTCCGTCCTCGACGTCAATTGTTCCAAGGACAA
 377  D  A  S  S  G  L  T  T  E  Q  P  V  T  F  R  P  R  R  Q  L  F  Q  G  Q

1225  GACATGAGCGAAGAAGAAACTGAGCAGAACCCTGAGATTATTCCAGCCAATTTGCTTCCAACTTATAATCTG
 401  D  M  S  E  E  E  T  E  Q  N  P  E  I  I  P  A  N  L  L  P  T  Y  N  L

1297  ATTCACAACACTAAACAAGTAGATGTAGATCCTGTCGGTGTTGCTGTTAGGCTGTCCAAAGACATTGCTGCT
 425  I  H  N  T  K  Q  V  D  V  D  P  V  G  V  A  V  R  L  S  K  D  I  A  A

1369  GACTTACAAGGTGAACCTAGGGTAGGAGAAGATAGACATATTCTTCCCAGATTCACTATCCTAGTTCGTCTA
 449  D  L  Q  G  E  P  R  V  G  E  D  R  H  I  L  P  R  F  T  I  L  V  R  L

1441  CTGAAACAACTCAAAGTATCACAAATTATGGAAGCAGCAAGAAAATTGTATAAACTTGAAAACGATCATCCT
 473  L  K  Q  L  K  V  S  Q  I  M  E  A  A  R  K  L  Y  K  L  E  N  D  H  P
```

Figure 1.

```
1513  AACTATATGAATTGGGATACATGGAGAGTCTACCGCGATGCTGTGTCACAAGCAGGTACTTGGTCTGCTCTA
 497   N  Y  M  N  W  D  T  W  R  V  Y  R  D  A  V  S  Q  A  G  T  W  S  A  L

1585  AATTCAATTCAGCAATTCATCTCTTCTGAAATGGTTGAACCAAAAGAAGCCTCACATCTCATTACCGTCTTA
 521   N  S  I  Q  Q  F  I  S  S  E  M  V  E  P  K  E  A  S  H  L  I  T  V  L

1657  CCAGCTGCCGTCTCAGATAAGAACAAGGCGTATCTGCATTTCTTGTTTGAAATGACAAAGGACCCTGTATTT
 545   P  A  A  V  S  D  K  N  K  A  Y  L  H  F  L  F  E  M  T  K  D  P  V  F

1729  AAGAACATGACATATGTTAACACCAGCCTCGTTCTAGCATTCTCAGAGGTTATACATCAAGTAGAAATGCAT
 569   K  N  M  T  Y  V  N  T  S  L  V  L  A  F  S  E  V  I  H  Q  V  E  M  H

1801  CAGGTCAGAGATTTGAAAATAAAGTCGGTATATATTCCATACCTGGTCCAAGAATTCGATGATGCAGTCAAA
 593   Q  V  R  D  L  K  I  K  S  V  Y  I  P  Y  L  V  Q  E  F  D  D  A  V  K

1873  GAGAATAACAGCATTAAAAATTCAGTTATACACCCATGCCCTTGGCGTTACCGGAAACACACACATTCTCCAT
 617   E  N  N  S  I  K  I  Q  L  Y  T  H  A  L  G  V  T  G  N  T  H  I  L  H

1945  TATCTCCGCCCCTATATTATTCAATTGAAGACCATTACTCACCATCAGCGACTCTTCATGGTCCAGTCCTTG
 641   Y  L  R  P  Y  I  I  Q  L  K  T  I  T  H  H  Q  R  L  F  M  V  Q  S  L

2017  GAAAGAGTGGTAGAACACAACCCTCGAAAGGTTATCGATCTCCTCTTGAGTCTTTATCTGGACCAAAACGAA
 665   E  R  V  V  E  H  N  P  R  K  V  I  D  L  L  L  S  L  Y  L  D  Q  N  E

2089  CATGCAGATATCAGGGTCGAAGCATTATTCTTGCTAATGAAGGCTGATCCCTCCATCCACGTCCTCAAGATG
 689   H  A  D  I  R  V  E  A  L  F  L  L  M  K  A  D  P  S  I  H  V  L  K  M

2161  GTTGCAGAACTAACCCATACGGAAAGCAACAATCAAGTTCTCTCAGCTTCGCAGTCTGCCATTAAGAGTGCT
 713   V  A  E  L  T  H  T  E  S  N  N  Q  V  L  S  A  S  Q  S  A  I  K  S  A

2233  GCAAATGTGGAAGGCGATATATACTCTGAAATGCGAAGGAAAGCCAAGGCCGTGGAACACCTGCTGAGCACA
 737   A  N  V  E  G  D  I  Y  S  E  M  R  R  K  A  K  A  V  E  H  L  L  S  T

2305  AGGAATATGGATGTATCATACTCCAAGAGTTACTTGTACGGCTATAAGAGCAAAAAAATCAATTATGATTCC
 761   R  N  M  D  V  S  Y  S  K  S  Y  L  Y  G  Y  K  S  K  K  I  N  Y  D  S

2377  CTCTATAATCTAAACTACATTGGAAGTGAGGACAGCATCTATCCTAAGTCCATGCTCTTAAATATCTTTACT
 785   L  Y  N  L  N  Y  I  G  S  E  D  S  I  Y  P  K  S  M  L  L  N  I  F  T

2449  AACAATCTTGGTAGAATCAACACCCATGTTCAGAAGGGATATATGGTATCAAGTATGACTGATCTATGGGAA
 809   N  N  L  G  R  I  N  T  H  V  Q  K  G  Y  M  V  S  S  M  T  D  L  W  E

2521  GCTTTCCACACAATATATAAAAAAGATAATGGATCACCTACTGATCCTAAAACTTTGGTAAAATTTGTTGAA
 833   A  F  H  T  I  Y  K  K  D  N  G  S  P  T  D  P  K  T  L  V  K  F  V  E

2593  GGCAATCTGAAATACTTTAACATGGGTGTCCAGAAATTCTGGGCATTTGACAACACAACATTCAGTAATGCA
 857   G  N  L  K  Y  F  N  M  G  V  Q  K  F  W  A  F  D  N  T  T  F  S  N  A

2665  TCTGCAGTAATACAGGAGTTCCTTAAAACTTACAAAAAACCTACAAATTTCAACCACACCAAACTCTCCAGC
 881   S  A  V  I  Q  E  F  L  K  T  Y  K  K  P  T  N  F  N  H  T  K  L  S  S

2737  AGCTCTTCAATTACTCTAACCCTGCCCTGTGCAATGGGTTTGCCTGCATATTTCAAGATGAACTCACCAAGT
 905   S  S  S  I  T  L  T  L  P  C  A  M  G  L  P  A  Y  F  K  M  N  S  P  S

2809  TTATGGAAGTATAATGGTGAATTTAGTATTCAGACTGATGCTAAAACTGATGTACCAATGTCTCTCGAGAAT
 929   L  W  K  Y  N  G  E  F  S  I  Q  T  D  A  K  T  D  V  P  M  S  L  E  N

2881  TTTATGAACATCACTGGTTCAATTAATCTTATGTTCAGCCAGATGTACCATGCCCAGCTTGCTTTCAGCACA
 953   F  M  N  I  T  G  S  I  N  L  M  F  S  Q  M  Y  H  A  Q  L  A  F  S  T

2953  GCTTTCGACAACAAGGAATACATTTCAGGTCTTGACAGGAAAGTGGAAGTTCATGTTCCAGTAAAATTCCAA
 977   A  F  D  N  K  E  Y  I  S  G  L  D  R  K  V  E  V  H  V  P  V  K  F  Q

3025  ATAAATCTGGATTTCAAGAACCACAACGGGTTCATACGAATCATCCCATTATTCACTGATAGAGATTACGAT
1001   I  N  L  D  F  K  N  H  N  G  F  I  R  I  I  P  L  F  T  D  R  D  Y  D
```

Figure 1. (*continued*).

```
3097    GTTTTGCAATGGCAGACTATTCCATACACAACCATCCATAACGTACCAGACTTTGAAACAGTCTACATGGAT
1025    V  L  Q  W  Q  T  I  P  Y  T  T  I  H  N  V  P  D  F  E  T  V  Y  M  D

3169    CAACTCTTCAAATTGATCCACGTGCGAAAAACAGCACACTTCGAGAAGAAAATGGGAGAAAACACGGGCATT
1049    Q  L  F  K  L  I  H  V  R  K  T  A  H  F  E  K  K  M  G  E  N  T  G  I

3241    GTATTTAAAGTAAAATATGATACAGATCAGGAATTTCTGGATACAAAGTGGTTTCTAGATGAATTCAAGGTT
1073    V  F  K  V  K  Y  D  T  D  Q  E  F  L  D  T  K  W  F  L  D  E  F  K  V

3313    CTCCAACTTTTCACTGGTCTTAATTATGACGTGCCCACAAAGGACATCTTCTACAACAACCTGACTGTATAT
1097    L  Q  L  F  T  G  L  N  Y  D  V  P  T  K  D  I  F  Y  N  N  L  T  V  Y

3385    TATGATCATGAAGACACCAAAAACCATGCCGTATCTTTTACTGTGACAAAAGAACAATCTAAGTTCTACGAA
1121    Y  D  H  E  D  T  K  N  H  A  V  S  F  T  V  T  K  E  Q  S  K  F  Y  E

3457    ACATTAAACCCAGTAGTTCAACAAAACCTAAAACTTTCATCTGGAAAGAAGCAAAAACATCGTAATGTAAAG
1145    T  L  N  P  V  V  Q  Q  N  L  K  L  S  S  G  K  K  Q  K  H  R  N  V  K

3529    AGCCATCGTATAAGGAGAGAATATACCGAGGATGAAAATCCAGCTATTCCAAAAGACAAACAACCAAACAGC
1169    S  H  R  I  R  R  E  Y  T  E  D  E  N  P  A  I  P  K  D  K  Q  P  N  S

3601    CATCCCAGACGTCAAGAATACCTAAGCAAATCAATGGCACTAACTGGAGATGCAACTGCAGTTGTACTTGAT
1193    H  P  R  R  Q  E  Y  L  S  K  S  M  A  L  T  G  D  A  T  A  V  V  L  D

3673    ATGACTTTGAAGTTTGAGGGGCCTGCTGAATCCTACTTTACAACCACCGTATCTCACGCTACATCTCTTGTT
1217    M  T  L  K  F  E  G  P  A  E  S  Y  F  T  T  T  V  S  H  A  T  S  L  V

3745    AATGGTAGCTCTAACTACCTACTCTTCTACGATCAACACTACTATGAGGAAAAGAAGAGAAATCAGTTCTGT
1241    N  G  S  S  N  Y  L  L  F  Y  D  Q  H  Y  Y  E  E  K  K  R  N  Q  F  C

3817    CTTTCATGGAGTGTCTACAAACCACAAGTTCCAATCATGAATATTTATAGTGCTTTTGAATTTGATCCAAAT
1265    L  S  W  S  V  Y  K  P  Q  V  P  I  M  N  I  Y  S  A  F  E  F  D  P  N

3889    TCCAAAGTGCATGCCATAATGAACATAGGAAAAGAATGTGAGAATGGAGGAAGTGCTGTAGCAAATATTGAT
1289    S  K  V  H  A  I  M  N  I  G  K  E  C  E  N  G  G  S  A  V  A  N  I  D

3961    ATGCTTAGATTGTCAGAACACCTAGATTACGTAAAGAATTTAACAGTCAGCAAGCTATGCGACCATGAAATG
1313    M  L  R  L  S  E  H  L  D  Y  V  K  N  L  T  V  S  K  L  C  D  H  E  M

4033    CGTACAAAACGTGATCACGTTCTGCCAGCTTGCAGGAATTCTACTGAGAGAGCCAGTGATCTCAACAGAGTT
1337    R  T  K  R  D  H  V  L  P  A  C  R  N  S  T  E  R  A  S  D  L  N  R  V

4105    CATGTTGATATCAATTATAACTTAAAGCAGCATGAAACATTCAAACGCAGAGTATACAAAGTATATGATTTT
1361    H  V  D  I  N  Y  N  L  K  Q  H  E  T  F  K  R  R  V  Y  K  V  Y  D  F

4177    GTACGTACCCACCTCTATCCACACGTATCAGAGGATGTGATCGTTGATAATCCAGCACAATTCATATCGGCA
1385    V  R  T  H  L  Y  P  H  V  S  E  D  V  I  V  D  N  P  A  Q  F  I  S  A

4249    AATTTCACATTGAAAGACAATACAAGAGCCTTCAACGTGAGCATTGAAACTCCTGTACTTAGTGTTAATGCA
1409    N  F  T  L  K  D  N  T  R  A  F  N  V  S  I  E  T  P  V  L  S  V  N  A

4321    ACTTCGGTTCGACTTCAATCATGGCAAAGTGAAATGCTCAGAATGAATCCAAGAACTTCATTTGCAAAACGT
1433    T  S  V  R  L  Q  S  W  Q  S  E  M  L  R  M  N  P  R  T  S  F  A  K  R

4393    TTTGCCAAATGGGCTCTACCCCTCTACTATAAACCAACGTGTGTTGTGGATTCTTCCTACATCAATACATTC
1457    F  A  K  W  A  L  P  L  Y  Y  K  P  T  C  V  V  D  S  S  Y  I  N  T  F

4465    GACAATTTTACATACAGTGCTCATCATATTGTACAGAACGATGCTTTCTACACTATATTGGATATTCCTCAA
1481    D  N  F  T  Y  S  A  H  H  I  V  Q  N  D  A  F  Y  T  I  L  D  I  P  Q

4537    AAGTTTAATATGGAATACTTTAAAGTTGCATTCAAACCAACATCCCCAGTTCCAAATATGCAGAGGGAAGTC
1505    K  F  N  M  E  Y  F  K  V  A  F  K  P  T  S  P  V  P  N  M  Q  R  E  V
```

Figure 1. *(continued)*.

```
4609   TTAGTGTTCCTGAGAAATGCAAAAATAGAATTGAAACCTAATCAAGGAATGCCAGAAGTTTACGTAGAAGGT
1529   L   V   F   L   R   N   A   K   I   E   L   K   P   N   Q   G   M   P   E   V   Y   V   E   G

4681   AAAAGAGTGGATTATAACCACCACCATTCAACTGACTTGAATGTGAGTCAAGACCGGATTGGTTATGTCTAT
1553   K   R   V   D   Y   N   H   H   H   S   T   D   L   N   V   S   Q   D   R   I   G   Y   V   Y

4753   GCCCTTCCAACAAAGGCTGCTCATATTGTATTTCCTTCATACGAGATCGAAATGTTTTACGATGGCTCCAGG
1577   A   L   P   T   K   A   A   H   I   V   F   P   S   Y   E   I   E   M   F   Y   D   G   S   R

4825   ATTATGATACAAGCTTCAAATATGTACCGAAATTTCACCAAAGGTCTCTGTGGTAACATGGATGGAGAATTC
1601   I   M   I   Q   A   S   N   M   Y   R   N   F   T   K   G   L   C   G   N   M   D   G   E   F

4897   GTGAACGACGTGTTGACTCCATGGGGCTGCTACGCCAAAGACATGGCACTCTTTGTAGCATCTTATGCCGAC
1625   V   N   D   V   L   T   P   W   G   C   Y   A   K   D   M   A   L   F   V   A   S   Y   A   D

4969   AACAGCAACAGCGAAGTAAGGAAGATCAAGGCTACCCAAAACGAGCAGACATGCGTCCCACAGTTTCATCAA
1649   N   S   N   S   E   V   R   K   I   K   A   T   Q   N   E   Q   T   C   V   P   Q   F   H   Q

5041   CCTCTAGTCAGCCATCAAATGAGATTGTCTCAAGTCATTAAGTTAGCAGATACCTCTTCCTCCTCTGAATCT
1673   P   L   V   S   H   Q   M   R   L   S   Q   V   I   K   L   A   D   T   S   S   S   E   S

5113   TCCTCTTCATCAGAATCTCATGAAAACAATTCATCACCATCTTCTGAGTCTCAAGTCAACAAGTCTAAGCGT
1697   S   S   S   E   S   H   E   N   N   S   S   P   S   S   E   S   Q   V   N   K   S   K   R

5185   CAACCAAATTCCAGACCAAGATCCAGTTCCAGTTCCAGCTCCAGCTCCAGTTCTGAATCCAATGAATCTGTC
1721   Q   P   N   S   R   P   R   S   S   S   S   S   S   S   S   S   S   E   S   N   E   S   V

5257   CTTGCCAAGAAAATCATTAACAACCAAATTGGACCAAAGCCTACTCTAATACCATCACAGTCTCCTATGACG
1745   L   A   K   K   I   I   N   N   Q   I   G   P   K   P   T   L   I   P   S   Q   S   P   M   T

5329   TCCGATGACAAATGCATGACACAACAACCAAGACACACTTATTATGAAAACCAATTCTGTGTCAGTGAAAAG
1769   S   D   D   K   C   M   T   Q   Q   P   R   H   T   Y   Y   E   N   Q   F   C   V   S   E   K

5401   CCCCTGGACACATGCATGCCCCTTATTTGCCATGCAACTGAATCCTATACAATAGATGTAAATTTTTACTGT
1793   P   L   D   T   C   M   P   L   I   C   H   A   T   E   S   Y   T   I   D   V   N   F   Y   C

5473   GTTCCACTAGGTCCTGCAGCTAATCATTACATGAAACTTGTCAAGAAAGGCATTCTCCCAGATTTAAGTCGC
1817   V   P   L   G   P   A   A   N   H   Y   M   K   L   V   K   K   G   I   L   P   D   L   S   R

5545   AACAGGAATGGTAAGAGGGTAGTTCTTCCTGTTGAAATCCCAATACAATGCGAACCTGTTTTAAACTGAAAT
1841   N   R   N   G   K   R   V   V   L   P   V   E   I   P   I   Q   C   E   P   V   L   N   *

5617   CCGATTCTTGTTAATGGACTTAATTTAAAAAATCTGACTTATGAAGAAACTGGAATATTTTCAAAAATTGA
5689   ACTATTTATGTATTT**AATAAATAAA**GAATGATAGCAACTTGAAAAAAAAAAAAAAAAAAAAAA
```

Fig. 1. Nucleotide sequence and deduced amino acid sequence of vitellogenin cDNA of *B. germanica* (GenBank accession number: AJ005115). The putative signal peptide is indicated by a single underline. Clusters of serine residues are indicated by double underline. Boldfaced letters indicate putative asparagine-linked glycosylation sites. Putative cleavage sites (having the consensus sequence RXXR) for the proteases of the subtilisin family are framed. Asterisk means stop codon. Italic boldfaced letters show the double polyadenylation signal.

In addition, the analysis of the amino acid sequence reveals six putative cleavage sites (having the consensus sequence RXXR) for the proteases of the subtilisin family (Barr, 1991) at positions 119–122, 283–286, 391–394, 505–508, 1171–1174, and 1337–1340. Finally, putative asparagine-linked glycosylation sites are present at positions 247, 278, 570, 575, 618, 874, 879, 898, 955, 1116, 1241, 1325, 1349, 1409, 1420, 1431, 1482, 1566, 1611, 1705, 1716, and 1741.

Comparison With Vitellogenin Sequences of Other Insect Species

Comparison of the amino acid sequence of *B. germanica* (Fig. 1) with those of other insect species revealed that the percentage of similar-

```
Blattella        :  MTWNALWCCLLVSAASA----
Periplaneta      :  -MWKGFLCCLLVAGVTS----
Graptopsaltria   :  -MWAPLYMCLLVAGAIA----
Plautia1         :  MLWSSAL--LLAFACLAAA--
Plautia2         :  MNWTLVA--LLTFVGLAAA--
Plautia3         :  -MWAPFT--LFVVAFLTLASA
Riptortus        :  -MWSPVIICLLVGLASA----
Anthonomus       :  -MWSTVALCLLVGLSYVSSS-
Pimpla           :  -MWCPLFLVLLAGAATA----
Athalia          :  -MWSPLLLCLLVGIASA----
Aedes            :  -MLAKLLLLALAGLTAAYQY-
Anopheles        :  -MIAKLLLLTLVGLCT-----
Lymantria        :  -MRLLLISAFIAVVSS-----
Bombyx           :  -MKLFVLAAIIAAVSS-----
Locusta-A        :  -MWAVIVLGLL??????????
Locusta-B        :  -MWALILSGLL??????????
```

Fig. 2.   Alignment of the signal peptide of vitellogenins from the following insects: *Blattella germanica, Periplaneta americana, Graptopsaltria nigrofuscata, Plautia stali* (isoforms 1, 2, and 3), *Riptortus clavatus, Anthonomus grandis, Pimpla nipponica, Athalia rosae, Aedes aegypti, Anopheles gambiae, Lymantria dispar, Bombyx mori* and *Locusta migratoria* (isoforms A and B). Database accession numbers for most of the species are indicated in Materials and Methods. Those of the partial sequences corresponding to *L. migratoria* are: M17333 (isoform A) and M17334 (isoforms B) (Locke et al., 1987). The signal peptide sequence of *A. gambiae* is reported in Romans et al. (1995).

ity decreases approximately in parallel to the decrease of phylogenetic relationships (*P. americana*: 42%, *G. nigrofuscata*: 39%, *R. clavatus*: 39%, *P. stali* (isoform 1): 37%, *P. nipponica*: 37%, *A. grandis*: 37%, *A. aegypti*: 36%, *B. mori*: 34%, and *L. dispar*: 32%). The exception is the hymenop-

teran *A. rosae*, which shows a remarkably high (41%) percentage of similarity.

An alignment of the vitellogenin sequences (excluding the signal peptide) of the 11 insect species mentioned above was carried out with the Pileup program from the GCG package (not shown), and revealed numerous conserved motifs (like the tetrapeptide GLCG in position 1615–1618) typical of vitellogenin sequences. The subdomains I–V, described as regions of relatively high amino acid conservation (Chen et al., 1997; Sappington and Raikhel, 1998), were also readily identifiable in the vitellogenin of *B. germanica* within the alignment (subdomain I: between amino acid 22 and 303; II: 442–826; III: 855–1034; IV: 1470–1646; V: 1788–1862).

Finally, we carried out a phylogenetic analysis of the vitellogenin sequences of the 11 available insect species following parsimony and non-parsimony approaches. Parsimony analyses were performed with the Protpars program from the Phylip package, and using the above-mentioned Pileup alignment. The topology of the resulting tree (not shown) presented serious incongruencies with respect to the currently accepted insect phylogenetic schemes (see, for example, Kukalová-Peck, 1991; Friedrich and Tautz, 1995, 1997). For example, the lepidopterans *L. dispar* and *B. mori* appeared as the sister group of all other orders, whereas the dipteran *A. aegypti* appeared in a very internal node. This could mean that the evolution of insect vitellogenins has not been strictly parsimonious.

Non-parsimony analyses were carried out fol-
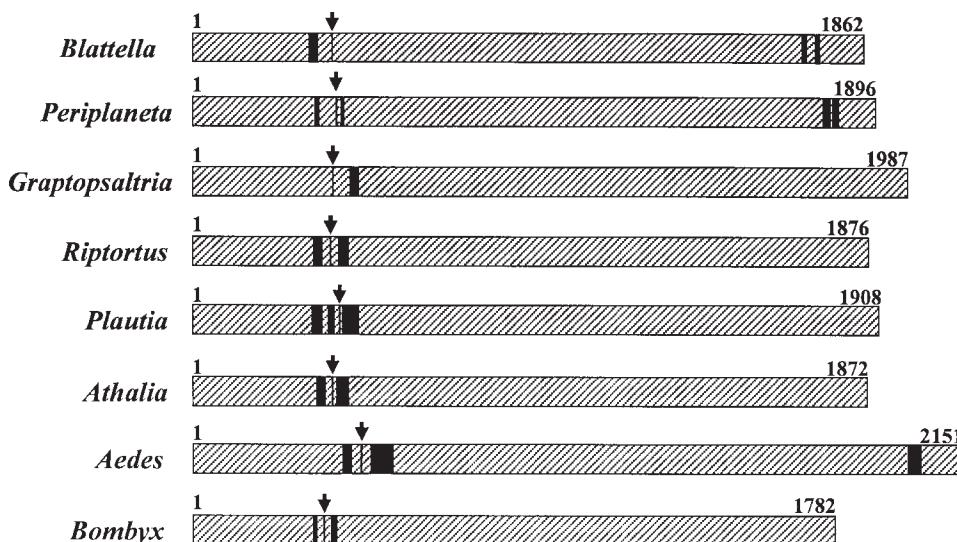
Fig. 3.   Schematic representation of the variants in distribution of polyserine domains in the vitellogenins of different insects. See the binomial nomenclature of all species in the legend of Figure 2. In the case of *Plautia stali*, isoform 1 has been represented. The serine-rich stretches are indicated as black segments, their length being proportional to the length of the stretch. The arrows and the narrow lines indicate the putative or demonstrated cleavage site.
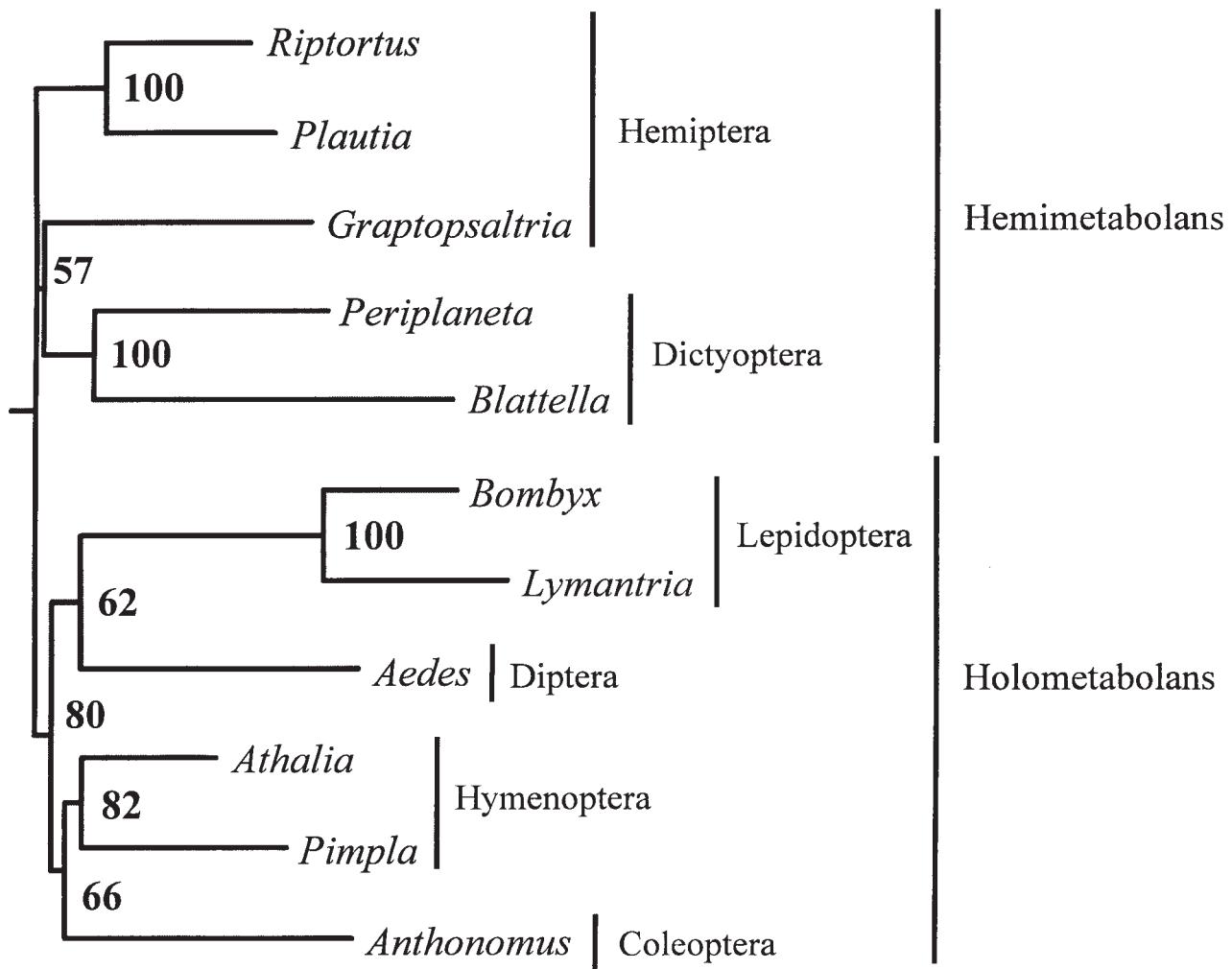
Fig. 4. Phylogenetic tree for the entire sequence of vitellogenin (less the signal peptide) in insects. It was obtained with the method of Neighbor Joining, using the distance of Kimura. Numbers in the nodes correspond to bootstrap values in 100 replicates. See the binomial nomenclature of all species in the legend of Figure 2.

lowing the method of Neighbor Joining, from the Phylip package, using the distance of Kimura and the entire sequences of vitellogenin (less the signal peptide). The most external nodes of the tree obtained (Fig. 4) separate holometabolan from hemimetabolan groups. The more internal ones group in a quite coherent manner the different orders (for example dipterans + lepidopterans), and the different species of the same order (for example *B. mori + L. dispar*). The sole readily apparent incongruence is the position of the cicada *G. nigrofuscata* clustering with the two dictyopterans (although this node was the one having the lowest bootstrap index) and not with the other hemipterans. The same analysis was carried with each one of the five subdomains of

Chen et al. (1997) (see also Sappington and Raikhel, 1998) but gave poorer results (not shown), the orders being incongruently distributed in most cases. Only the analysis based on subdomain II resulted in a relatively coherent topology, although in this case *G. nigrofuscata* clustered with the dipterans and lepidopterans. Again, this node was the one showing the lowest bootstrap index (47).

The results of the phylogenetic analysis suggest that the entire vitellogenin sequence should be used when searching phylogenetic inferences. In addition, the incongruent results emerging from parsimony methods of analysis suggest that the molecular evolution of insect vitellogenins has not been strictly parsimonious. However, this would not be exceedingly surprising taking into

account the extreme diversity of reproductive strategies occurring in insects (Engelmann, 1970; Bellés, 1998), which may have resulted in a great diversity of selective pressures acting upon such a molecule.

LITERATURE CITED

Bellés X. 1998. Endocrine effectors in insect vitellogenesis. In Coast GM, Webster SG, editors. Recent advances in arthropod endocrinology. Cambridge: Cambridge University Press, p 71–90.

Bellés X, Piulachs MD, Pascual N, Maestro JL, Martín D, Vilaplana Ll. 1999. Regulation of vitellogenesis in the German cockroach and the problem of cyclicity. In: Roubos E, Wendelaar Bonga SE, Vaudry H, De Loof A, editors. Recent developments in comparative endocrinology and neurobiology. Maastricht: Shaker, p 223–227.

Bellés X, Casas J, Messeguer A, Piulachs MD. 1987. In vitro biosynthesis of JH III by the corpora allata of adult females of *B. germanica* (L.). Insect Biochem 17:1007–1010.

Barr PJ. 1991. Mammalian subtilisins: the long-sought dibasic processing endoproteases. Cell 66:1–3.

Chen J-S, Cho W-L, Raikhel AS. 1994. Analysis of mosquito vitellogenin cDNA, similarity with vertebrate phosvitins and arthropod serum proteins. J Mol Biol 237:641–647.

Chen J-S, Sappington TW, Raikhel AS. 1997. Extensive sequence conservation among insect, nematode, and vertebrate vitellogenins reveals ancient common ancestry. J Molec Evol 44:440–451.

Comas D, Piulachs MD, Bellés X. 1999. Fast induction of vitellogenin gene expression by juvenile hormone III in the cockroach *Blattella germanica* (L.) (Dictyoptera, Blattellidae). Insect Biochem Mol Biol 29:821–827.

Devereux J, Haeberli P, Smithies O. 1984. A comprehensive set of sequences analysis programs for the VAX. Nucleic Acids Res 12: 387–395.

Engelmann F. 1970. The physiology of insect reproduction. Oxford: Pergamon Press.

Friedrich M, Tautz D. 1995. Ribosomal DNA phylogeny of the major extant arthropod classes and the evolution of myriapods. Nature 376:165–167.

Friedrich M, Tautz D. 1997. An episodic change of rDNA nucleotide substitution rate has occurred during the emergence of the insect order Diptera. Mol Biol Evol 14:644–653.

Frohman MA, Dush MK, Marton GR. 1988. Rapid production of full-length cDNAs from rare transcripts: am-

plification using a single gene-specific oligonucleotide primer. Proc Nat Acad Sci USA 85:8998–9002.

Hagedorn HH, Maddison DR, Tu Z. 1998. The evolution of vitellogenins, cyclorrhaphan yolk proteins and related molecules. Adv Insect Physiol 27:335–384.

Hirai H, Watanabe D, Kiyota A, Chinzei Y. 1998. Nucleotide sequence of vitellogenin mRNA in the bean bug, *Riptortus clavatus*: analysis of processing in the fat body and ovary. Insect Biochem Mol Biol 28: 537–547.

Hiremath S, Lehtoma K. 1997a. Complete nucleotide sequence of the vitellogenin mRNA from the gypsy moth: novel arrangement of the subunit encoding regions. Insect Biochem Mol Biol 27: 27–35.

Hiremath S, Lehtoma K. 1997b. Structure of the gypsy moth vitellogenin gene. Arch. Insect Biochem Physiol 36: 151–164.

Kukalová-Peck J. 1991. Fossil history and evolution of hexapod structures. In: Naumann ID, editor. The insects of Australia, 2nd ed. Vol. 1. Carlton: Melbourne University Press, p 141–179.

Lee JM, Hatakeyama M, Oishi K. 2000a. A simple and rapid method for cloning insect vitellogenin cDNAs. Insect Biochem Mol Biol 30:189–194.

Lee JM, Nishimori Y, Hatakeyama M, Bae T-W, Oishi K. 2000b. Vitellogenin of the cicada *Graptopsaltria nigrofuscata* (Homoptera): analysis of its primary structure. Insect Biochem Mol Biol 30:1–7.

Locke J, White BN, Wyatt GR. 1987. Cloning and 5′ end nucleotide sequences of two juvenile hormone-inducible vitellogenin genes of the African migratory locust. DNA 6:331–342.

Maestro J-L, Danés M-D, Piulachs M-D, Cassier P, Bellés X. 1994. Juvenile hormone inhibition in corpora allata from ovariectomized *Blattella germanica*. Physiol Entomol 19:342–348.

Martín D, Piulachs MD, Bellés X. 1995a. Patterns of haemolymph vitellogenin and ovarian vitellin in the German cockroach, and the role of juvenile hormone. Physiol Entomol 20:59–65.

Martín D, Piulachs MD, Bellés X. 1995b. Production of vitellogenin in vitro by the periovaric fat body of *Blattella germanica* (L.) (Dictyoptera, Blattellidae). Invert Reprod Dev 28:171–176.

Martín D, Piulachs MD, Bellés X. 1996. Production and extraovarian processing of vitellogenin in ovariectomized *Blattella germanica* (L.) (Dictyoptera, Blattellidae). J Insect Physiol 42:101–105.

Martín D, Piulachs, MD, Comas D, Bellés X. 1998. Isolation and sequence of a partial vitellogenin cDNA from the cockroach *Blattella germanica* (L.) (Dic-

tyoptera, Blattellidae), and characterization of the vitellogenin gene expression. Arch Insect Biochem Physiol 38:137–146.

Nielsen H, Engelbrecht J, Brunak S, von Heijne G. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Prot Eng 10:1–6.

Nose Y, Lee J M, Ueno T, Hatekayama M, Oishi K. 1997. Cloning of cDNA for vitellogenin of the parasitoid wasp *Pimpla nipponica* (Hymenoptera: Apocrita: Ichneumonidae): vitellogenin primary structure and evolutionary considerations. Insect Biochem Mol Biol 27:1047–1056.

Romans P, Tu Z, Ke Z, Hagedorn H. 1995. Analysis of a vitellogenin gene of the mosquito, *Aedes aegypti* and comparisons to vitellogenins from other organisms. Insect Biochem Mol Biol 25:939–958.

Sappington TW, Raikhel AS. 1998. Molecular characteristics of insect vitellogenins and vitellogenin receptors. Insect Biochem Mol Biol 28:277–300.

Trewitt PM, Heilmann LJ, Degrugillier SS, Kumaran AK.

1992. The boll weevil vitellogenin gene: nucleotide sequence, structure, and evolutionary relationship to nematode and vertebrate vitellogenin genes. J Mol Biol 34:478–492.

Tufail M, Lee JM, Hatakeyama M, Oishi K, Takeda M. 2000. Cloning of vitellogenin cDNA of the American cockroach, *Periplaneta americana* (Dictyoptera), and its structural and expression analyses. Arch Insect Biochem Physiol 45:37–46 (this issue).

von Heijne G. 1986. A new method for predicting signal sequence cleavage sites. Nucleic Acids Res 14:4683–4690.

Wyatt GR, Davey KG. 1996. Cellular and molecular actions of juvenile hormone. II. Roles of juvenile hormone in adult insects. Adv Insect Physiol 26:1–155.

Yano K, Sakurai MT, Izumi S, Tomino S. 1994a. Vitellogenin gene of the silkworm, *Bombyx mori*: structure and sex-dependent expression. FEBS Lett 356:207–211.

Yano K, Sakurai MT, Watabe S, Izumi S, Tomino S. 1994b. Structure and expression of mRNA for vitellogenin in *Bombyx mori*. Biochim Biophys Acta 1218:1–10.