MOLECULAR AND CELLULAR BIOLOGY, May 1982, p. 554–563
0270-7306/82/050554-10$02.00/0

Vol. 2, No. 5

Evolution of a Multigene Family of Chorion Proteins in Silkmoths

GEORGE C. RODAKIS, NIKOS K. MOSCHONAS, AND FOTIS C. KAFATOS*

Department of Cellular and Developmental Biology, The Biological Laboratories, Harvard University, Cambridge, Massachusetts 02138[1]; and Department of Biology, University of Athens, Panepistimiopolis, Athens 621, Greece[2]*

Received 25 November 1981/Accepted 6 January 1982

The evolution of the A family of chorion genes was examined by comparing new protein and DNA sequences from the silkmoths *Antheraea pernyi* and *Bombyx mori* with previously known sequences from *Antheraea polyphemus*. The comparisons indicated that the A family and its major subfamilies are ancient and revealed how parts of the genes corresponding to distinct regions of the protein structure have evolved, both by base substitutions and by segmental reduplications and deletions.

It is now clear that many genes in eucaryotes are members of multigene families. Many such families are developmentally regulated; genes which arose by reduplication but have become structurally distinct are also differently expressed at different developmental periods or in different tissues and may serve subtly different but important aspects of differentiated phenotypes (9). It is important to study how structural genes evolve in the context of how they become developmentally specified for stage- or tissue-specific expression.

We used the silkmoth chorion multigene families as a model system for studying these questions. Previous work from this laboratory has shown that most of the structural proteins which make up the elaborate eggshell (chorion) of the saturniid silkmoth *Antheraea polyphemus* are homologous and are encoded by two multigene families, named A and B (5, 6, 14, 15; G. Rodakis, Ph.D. thesis, University of Athens, Greece, 1978; C. W. Jones, Ph.D. thesis, Harvard University, Cambridge, Mass., 1980; C. W. Jones and F. C. Kafatos, submitted for publication). Different members of each family are expressed at various stages during the 52-h period of choriogenesis: early, middle, late, and very late (7, 17). For example, the characterized members of one subfamily (AI) of the major A family are expressed in the late period of choriogenesis, whereas members of another subfamily (AII) are expressed in the middle period of choriogenesis (17; J. C. Regier and F. C. Kafatos, manuscript in preparation).

Extensive sequencing studies have revealed the main structural features of *A. polyphemus* A proteins from both subfamilies (3a, 6, 14, 15, 19; Rodakis, Ph.D. thesis; Jones, Ph.D. thesis; Jones and Kafatos, submitted for publication). Here we report sequencing studies in two other silkmoths, the saturniid *Antheraea pernyi* and the bombycid *Bombyx mori*, which significantly enhance our understanding of the evolutionary history of the A family as well as of the constant and variable structural features of these proteins.

MATERIALS AND METHODS

Protein purification, characterization, and sequencing. *A. pernyi* chorions were dissolved at 5 mg/ml in sample buffer (6 M guanidine hydrochloride, 0.36 M Tris-hydrochloride [pH 8.4], 0.75 mM disodium EDTA, 2 mM dithiothreitol) under nitrogen for 6 to 8 h, with gentle shaking. The proteins were labeled with limiting amounts of [^{14}C]iodoacetamide (usually equivalent to 0.1 times the calculated molarity of cysteine; 2.5 μCi/mg of chorion, 50 mCi/mmol, Amersham Corp.) in the dark for 15 min. Excess dithiothreitol was then added to ensure complete reduction of the proteins, followed by a molar excess of unlabeled iodoacetamide to ensure complete carboxamidomethylation, and finally a large excess of 2-mercaptoethanol was added to destroy the unreacted iodoacetamide (2). Labeling to the extent of approximately 2 μCi/mg of chorion could be achieved by this procedure. For preparative experiments, the labeled protein was diluted as desired with unlabeled carboxamidomethylated chorion.

Chorion proteins were fractionated by chromatography on preparative columns of Bio-Gel P-150 (Bio-Rad Laboratories). In one experiment, a column (83 by 5 cm) (Pharmacia Fine Chemicals, Inc.) was loaded with 250 mg of protein, and in a second experiment a column (70 by 8 cm) was loaded with 750 mg of protein; samples were 30 mg/ml in 7 M guanidine hydrochloride–15 mM Tris-hydrochloride (pH 8.5). The column was eluted with the same buffer for 6 days at 3 to 5 drops per min. Under these conditions, the resolution of A and B proteins was good. Fractions

were assayed for radioactivity by liquid scintillation, combined (see Fig. 2), exhaustively dialyzed against water, and lyophilized.

Protein fractions were redissolved in 6 M urea, 50 mM Tris-hydrochloride (pH 8.4), 0.75 mM disodium EDTA, 1.5 mM lysine and fractionated as described before (3) on isoelectric focusing slab gels (26 by 15 by 0.2 cm) containing 5% acrylamide, 0.3% *N,N'*-methylene-bis-acrylamide, 6 M deionized urea (Schwarz-Mann, Ultrapure), and 1.5% pH 4 to 6 Ampholines (LKB Instruments Co.). Each sample contained 25 mg of protein in 0.25 ml and was loaded on a strip of Whatman 3MM paper (13 by 1 cm). Current not exceeding 20 mA was applied across thle long dimension of the gel. After 14 h at 1,000 V, the gel was submerged in 40% saturated ammonium sulfate solution. The desired bands were excised with a razor blade, fragmented, thoroughly washed in the same solution, rinsed with distilled water, and eluted in the 6 M guanidine hydrochloride sample buffer (Rodakis, Ph.D. thesis). The eluate was membrane filtered, dialyzed against distilled water, and lyophilized. The protein was suspended in ice-cold 70% ethanol, and after 12 h at −20°C, it was collected by centrifugation (8,500 rpm). Aliquots were characterized by electrophoresis in a sodium dodecyl sulfate (SDS)-polyacrylamide slab gel (15).

Amino acid analysis and sequencing by Edman degradation (using an updated 890B Beckman sequencer) were performed as described before (15). A modified 0.1 M Quadrol program (1, 18) was used. Residues were identified as described before (15) after conversion in 1.0 M HCl at 80°C for 5 to 12 min. Methods of identification included gas chromatography (with and without silylation), thin-layer chromatography, liquid scintillation (for detection of [^{14}C]carboxamidomethylated cysteine) and amino acid analysis after back conversion. Duplicate sequence determinations were performed on two independently prepared samples of each protein as follows: A4–c5, 200 and 220 nmol, 43 and 44 residues, respectively; A3–d1, 220 and 450 nmol, 24 and 34 residues, respectively; A2–b3, 230 and 330 nmol, 25 and 27 residues, respectively.

DNA sequencing. The chorion DNA insert (353 base pairs) of cDNA clone m2274 from *B. mori* (4) was sequenced by the method of Maxam and Gilbert (10). All relevant methods used in our laboratory have been described previously (5, 6). The sequencing strategy was based on 5'-end labeling at the following four restriction endonuclease sites: *Hin*fI (beginning at nucleotide 28), *Hin*fI (nucleotide 133), *Ava*II (nucleotide 213) and *Bgl*II (nucleotide 306). All parts of the sequence were determined from both strands, except for nucleotides 1 to 85 and 311 to 353, which were only determined on the antimessage and message strands, respectively. The sequence GCCCGCGCCGCTGGCGCC (nucleotides 219 to 936) was determined with some difficulty, presumably because it can assume a stable hairpin structure; nevertheless, even the most troublesome nucleotides at the end of the hairpin stem, 220 to 222 and 235 to 236, were determined unambiguously from the message and antimessage strands, respectively. The sequence was analyzed and plotted by a computer program written by J. Pustell. The *A. polyphemus* sequences, except for pc292 (19), were obtained by C. W. Jones (Ph.D. thesis) and are discussed in detail elsewhere (3a; Jones and Kafatos, submitted for publication).

RESULTS

General structure of A proteins in *A. polyphemus*. Through extensive sequencing studies on proteins (6, 14, 15; Rodakis, Ph.D. thesis) and DNA clones (19; Jones, Ph.D. thesis; Jones and Kafatos, submitted for publication; Hamodrakas et al., in press), the general primary structural features of A proteins in *A. polyphemus* are now known (Fig. 1). In addition to the amino-terminal signal peptide, three major regions or domains exist in each protein (3a; Regier and Kafatos, manuscript in preparation).

(i) The central region (42 to 48% of the total length) is very conservative, showing a minimum of 77% sequence identity in seven proteins, and is predicted to be highly structured chiefly into β-pleated sheets (3a); it is enriched in valine and alanine relative to the rest of the sequence and does not contain obvious peptide repeats.

(ii) The amino-terminal region or left arm (35 to 42% of the total) is much more variable, less structured, enriched in glycine, tyrosine, and leucine, and contains a prominent array of tandem repeats of the pentapeptide Leu·Gly·Tyr·Gly·Gly (LGYGG in the one-letter code) or variants thereof. Tandem repeats of Cys·Gly (CG) are also present; in one subfamily (AI, typified by 18c), the repeats are preceded by the amino-terminal dipeptide Tyr·Gly (YG), whereas in a second subfamily (AII, typified by pc609 and pc292) the repeats are preceded by a longer, more variable peptide.

(iii) The carboxy-terminal region or right arm (15 to 16% of the total) is moderately conservative, not highly structured, enriched in glycine and cysteine, and contains three to four tandem CG repeats.

Purification, characterization, and sequence analysis of A proteins from *A. pernyi*. To obtain information on the A family in another species of the genus *Antheraea*, we purified and partially sequenced three A proteins from *A. pernyi*.

As summarized in Fig. 2, purification involved protein fractionation by size on a Bio-Gel P-150 column, resolution by charge on a preparative isoelectric focusing gel, and verification of purity by electrophoresis on an SDS-polyacrylamide gel. The two major Bio-Gel P-150 fractions (III and II) were highly enriched in A and B proteins, respectively. Of several A proteins recovered from fraction III after isoelectric focusing, three were substantially homogeneous by SDS electrophoresis. By convention (2), they are referred to by a code representing their coordinates in both isoelectric focusing and SDS electrophore-

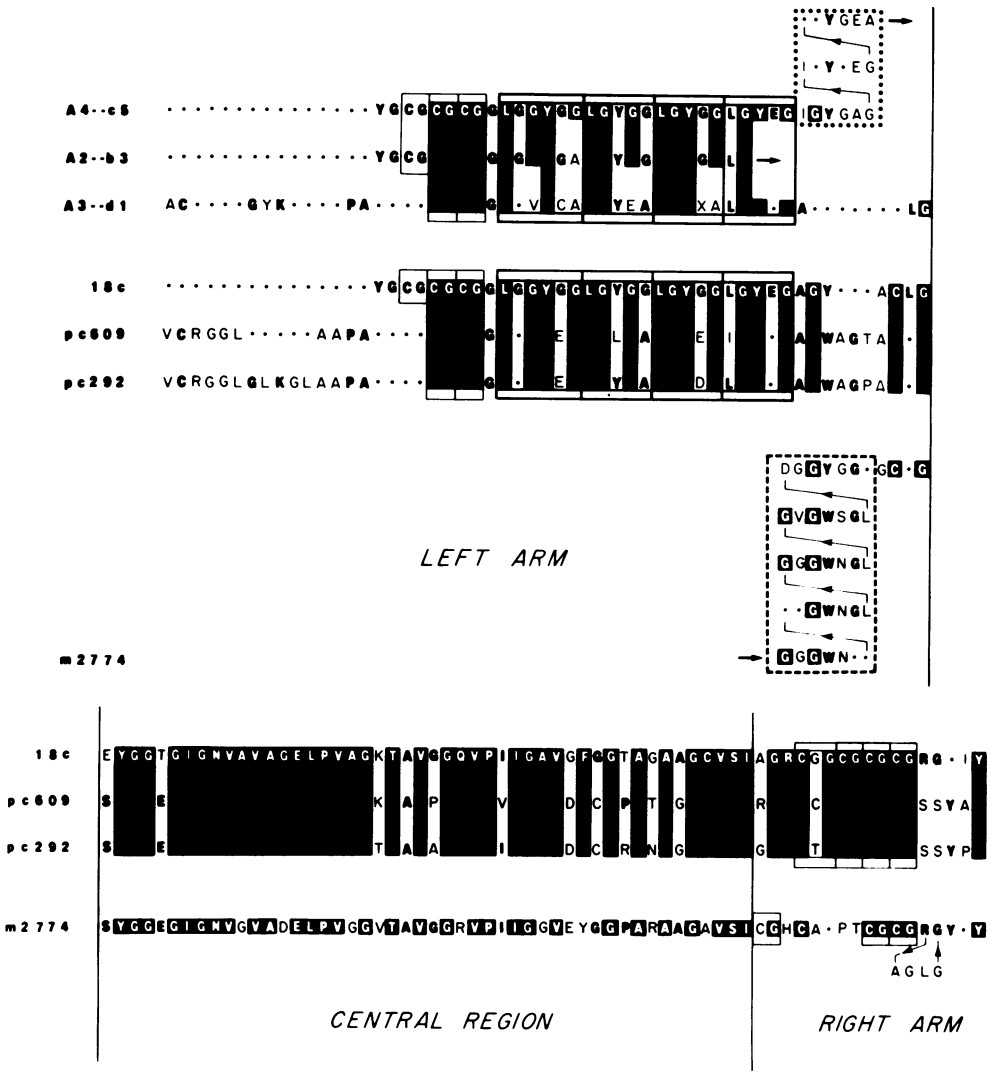


FIG. 1. Comparison of A family amino acid sequences in the three regions of the mature protein: left arm, central region, and right arm. In each region, the top set of sequences (A4--ca5, A2--b3, and A3--d1) is from *A. pernyi*, the middle set (18c, pc609, and pc292) is from *A. polyphemus*, and the bottom sequence (m2774) is from *B. mori*. Residues which are identical in all seven A sequences known in *A. polyphemus* (8, 12) are symbolized by a black background and are shown in white letters in sequence 18c; those which are also present in A family sequences from *A. pernyi* or *B. mori* are also shown by white letters and a black background in sequences A4--c5 and m2774, respectively. Other residues which are shared between *A. pernyi* or *B. mori* sequences and one or more *A. polyphemus* sequences are indicated in black boldface letters. Gaps necessary for sequence alignment are shown by dots. Repetitive peptides are boxed in thin lines (consensus CG), thick solid lines (consensus LGYGG), thick dashed lines (consensus GGGWNGL), or thick dotted lines (consensus IGYGEG). Expansion of a repetitive sequence array relative to the A proteins of *A. polyphemus* is indicated by stacked peptides connected by arrows. An insert of unknown origin in the right arm of m2774 is also indicated by arrows. All arrows point in the direction from the amino to the carboxyl terminus. The amino-terminal sequence of m2774 is not known.

sis in order of increasing molecular weight: A2–b3, A3--d1, and A4--c5.

Amino acid analyses (Table 1) revealed that the purified A proteins are reasonably representative of the total A fraction (III). Although some variations between components exist (e.g., in the content of glycine and alanine), similar variations are also observed among A proteins in *A. polyphemus* (Table 1).

All three proteins had free amino termini and

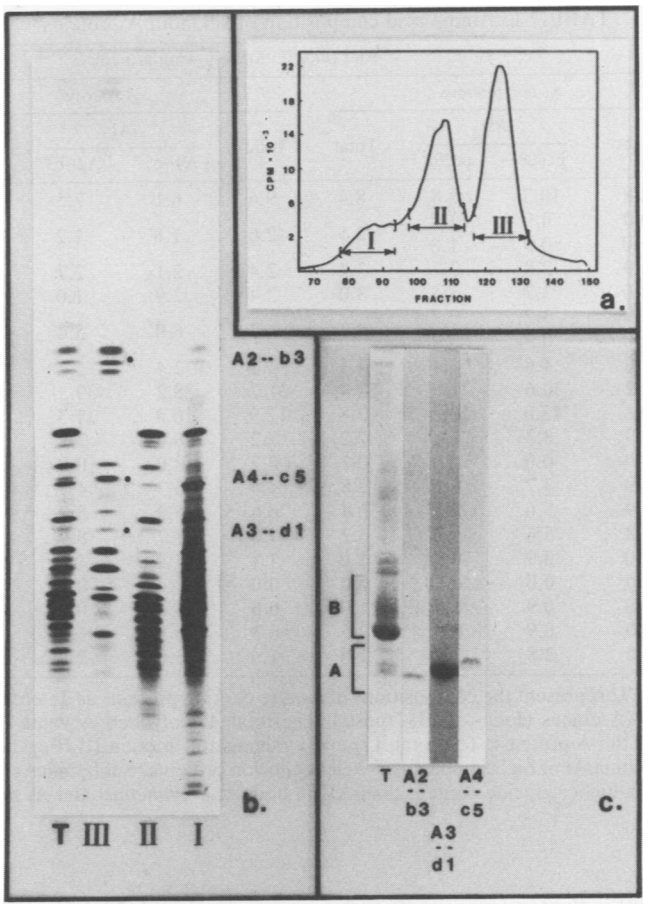


FIG. 2. Purification of three A family proteins from *A. pernyi*. (a) Total *A. pernyi* chorion proteins were fractionated on a Bio-Gel P-150 column (see the text). (b) Pooled fractions (I, II, and III) were analyzed by isoelectric focusing together with total chorion protein (T); dots and codes indicate the three proteins which were obtained from fraction III and sequenced. (c) Documentation of the purity of these three proteins and comparison of their electrophoretic mobilities with those of the A and B components in total chorion proteins (T).

were directly subjected to partial sequence analysis in duplicate (24 to 44 residues) by Edman degradation, using a Beckman sequencer. Figure 1 compares the derived sequences with those observed in *A. polyphemus*. Although the new sequences do not extend beyond the variable left arm, it is clear that they belong to the same two subfamilies as the *A. polyphemus* sequences; two are members of the AI (18c) subfamily, and the third is a member of the AII (pc292/pc609) subfamily. All three sequences show the tandem dipeptide and pentapeptide repeats characteristic of the left arm in *A. polyphemus*.

In particular, the features shared between A3--d1 and the pc292/pc609 subfamily include two (rather than three) CG repeats, a relatively long amino-terminal peptide (rather than YG), and two deletions (of glycine and glutamic acid) within the pentapeptide repeats. Relative to pc292, the amino-terminal peptide of A3--d1 is foreshortened by two deletions and modified by two amino acid replacements; foreshortening (through a different deletion) also occurs in pc609. In fact, the amino terminal peptide of A3--d1 is more reminiscent of another *A. polyphemus* AII protein, which has been sequenced partially and begins with ACVGYKG (Rodakis, Ph.D. thesis). Other differences between A3--d1 and pc609/pc292 are a deletion near the distal end of the arm (larger than a similar deletion found in 18c), and at least five amino acid replacements within the pentapeptide repeat array. In summary, A3--d1 is clearly a member of the AII subfamily, but has diverged considerably from the pc292/pc609 prototypes by amino

TABLE 1. Amino acid composition of silkmoth A proteins[a]

Amino acid	Amt (mol %) for following protein:								
	A. polyphemus				*A. pernyi*				*B. mori*
	AI 18c	AII		Total	Total	AI		AII	m2774
		pc609	pc292			A4--c5	A2--b3	A3--d1	
Cysteine	8.8	10.2	8.8	8.4	9.6	6.1	7.5	7.4	5
Aspartic acid	0.0	0.9	1.8	2.5	2.6	1.8	1.2	2.9	2
Asparagine	1.0	0.9	1.8						4
Threonine	2.9	2.8	2.7	3.4	2.4	2.1	2.7	2.7	2
Serine	1.0	3.7	3.5	3.0	2.4	2.9	1.0	2.8	3
Glutamic acid	2.9	3.7	2.7	4.2	4.2	6.0	5.4	4.1	3
Glutamine	1.0	0.9	0.9						0
Proline	2.0	4.6	4.4	4.1	3.1	2.4	2.2	3.9	4
Glycine	40.2	30.6	31.0	32.4	31.7	38.2	39.3	31.8	38
Alanine	10.8	13.0	13.3	13.8	12.9	10.4	11.3	13.7	8
Valine	6.9	8.3	7.1	7.2	7.2	6.0	7.1	7.4	9
Methionine	0.0	0.0	0.0	0.1	0.2	0.0	0.0	0.6	0
Isoleucine	4.9	3.7	3.5	3.8	4.5	5.3	4.5	4.8	4
Leucine	5.9	5.6	7.1	6.4	6.6	5.9	5.2	6.1	5
Tyrosine	7.8	5.6	6.2	6.2	9.0	7.7	8.9	7.9	5
Phenylalanine	1.0	0.9	0.9	1.0	1.1	1.2	1.1	1.1	0
Histidine	0.0	0.0	0.0	0.0	0.0	0.4	0.1	0.0	1
Lysine	1.0	0.9	0.9	1.0	0.6	0.6	0.0	0.9	0
Tryptophan	0.0	0.9	0.9	0.5	0.3	0.0	0.0	0.0	4
Arginine	2.0	2.8	2.7	1.8	1.6	3.0	2.5	1.9	3

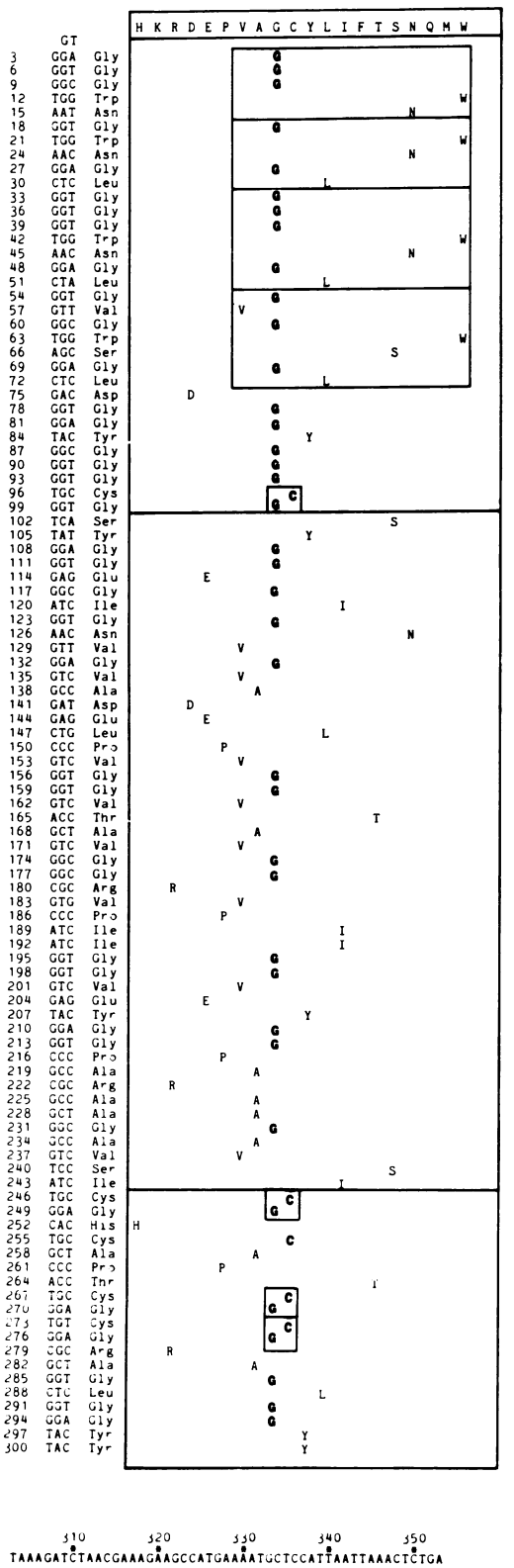
[a] 18c, pc609, and pc292 represent the compositions of mature chorion proteins as determined by sequencing of the corresponding cDNA clones (Jones, Ph.D. thesis). The total *A. polyphemus* value is for fraction s/s (15) which contains 90% of the A proteins. The total *A. pernyi* value is for fraction III (Fig. 1). Both total fractions have very minor contaminants of higher-molecular-weight chorion proteins. Subfamilies are indicated as AI and AII. Note that AI has a higher glycine content than AII in both *A. polyphemus* and *A. pernyi*.

acid replacements within the tandem repeats and by replacements plus deletions in the rest of the arm.

A2--b3 is nearly identical to 18c in the portion sequenced; it only differs by one replacement out of 27 amino acid residues. A4--c5 is identical to 18c for the first 30 residues, but then deviates substantially. Although several alignments are possible, we prefer the one shown in Fig. 1; it suggests that in A4--c5 a hexapeptide has reduplicated twice and has been modified by small deletions and replacements, yielding three tandem but variable repeats of the consensus sequence IGYGEG. We favor that interpretation because we consider unlikely the major alternative, i.e., that this portion of the arm evolved solely by replacements; if no reduplication had taken place, the very conservative beginning of the central region should have been encountered in the portion sequenced. Furthermore, we favor the reduplication hypothesis because of a similar event in the evolution of an A protein in *B. mori* (see below). The IGYGEG repeat array is aligned in Fig. 1 with a rather variable portion of the *A. polyphemus* sequence, but it should also be noted that its consensus sequence is related to the preceding LGYGG array. The origin of the IGYGEG repeat will be further considered below.

Determination of a *B. mori* A protein sequence. For convenient determination of a *B. mori* sequence, we used a cDNA clone m2774, which has been shown by hybrid-selected translation to encode an A-size chorion protein designated as A4 (4). The chorion DNA insert of m2774 was sequenced in its entirety by the method of Maxam and Gilbert (10) and conceptually translated; only one of the six reading frames was open. The m2774 DNA and protein sequences are presented in Fig. 3.

The clone was incomplete at the end corresponding to the 5′ end of the mRNA; no 5′-untranslated, initiator ATG, or recognizable signal peptide (5) sequences are present. At the opposite end, a 52-base-pair 3′-untranslated sequence beginning with a TAA termination codon was present; it does not contain the sequence AATAAA (13), and thus is presumably incomplete. The coding sequence of m2774 corresponded to 100 amino acid residues, and therefore represents nearly the entire mature protein sequence; by comparison with sequenced *A. polyphemus* A proteins, the *B. mori* A4 protein migrated on SDS-polyacrylamide gels with a



mobility suggesting a length of 106 ± 4 amino acids (data not shown). The m2774 protein sequence was clearly homologous with the A family of *A. polyphemus* (Fig. 1) and could similarly be divided into a central region flanked by two arms (Fig. 1 and 3).

The central region had a high degree of sequence identity (68.7 to 72.9%) relative to the A protein sequences of *A. polyphemus*. Of the 48 amino acid residues in this region, 30 were shared between m2774 and all 7 known *A. polyphemus* sequences; an additional 8 residues of m2774 were shared with some but not all *A. polyphemus* sequences. The interspecies homology was equally apparent at the DNA level (70.8 to 74.3% identity, as compared with a minimum of 77.1% within *A. polyphemus*; Fig. 4). The sequences of the two species were completely colinear; all differences were base substitutions as opposed to deletions, duplications, or insertions.

By contrast with the central region, both arms of m2774 were considerably more variable relative to the *Antheraea* sequences. In the right (carboxy-terminal) arm (Fig. 5), only 7 out of 19 amino acid residues could be matched with all 7 *A. polyphemus* sequences, and 3 residues could be matched with some but not all *A. polyphemus* sequences. Furthermore, several small deletions or insertions must be postulated, and the alignment is not unambiguous. Of the three to four tandemly repetitive CG dipeptides characteristic of this arm in *A. polyphemus*, only two were preserved in m2774; one additional (but nontandem) CG repeat was generated in m2774 by a single base substitution.

In the left (amino-terminal) arm, the sequence homologies were even more limited and ambiguous. The alignments shown in Fig. 1 are based on the following observations.

The available m2774 sequence begins with four tandem repeats of the concensus heptapeptide GGGWNGL. Only the third and fourth

FIG. 3. DNA and protein sequence of the *B. mori* m2774 chorion component. The 3'-untranslated part of the nucleotide sequence is shown at the bottom (nucleotides 303 to 353). The coding sequence (nucleotides 1 to 302) is presented vertically as codons and is conceptually translated. Within the outlined blocks on the right, amino acid residues are plotted in vertical columns (in the order indicated at the top, according to the one-letter code), to emphasize the patterns of the protein sequence. The three blocks correspond to different regions (domains) of the protein from top to bottom: left (amino-terminal) arm, central region, and right (carboxy-terminal) arm. The repetitive dipeptide CG is shown in small boxes; larger rectangular boxes outline the tandem repeats (consensus GGGWNGL) in the left arm. See also Fig. 1.

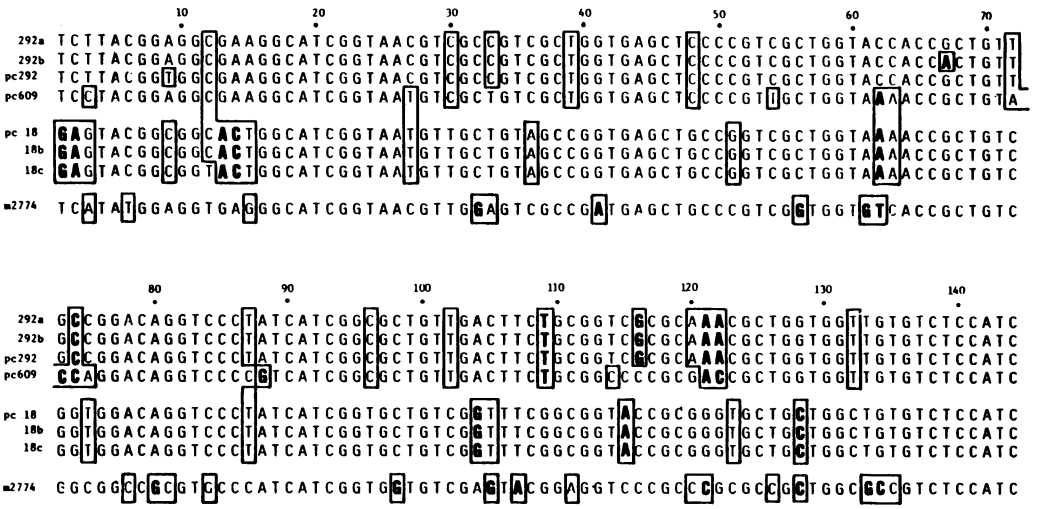


FIG. 4. Comparison of nucleotide sequences of the A family in the central region. Nucleotides are numbered arbitrarily from the 5′ end of the central region. All known A sequences are presented. Substitutions are boxed and are shown in boldface if they lead to amino acid replacements.

residues (GW) of that peptide are invariant in all four repeats; the second residue (G) varies in two repeats, and the first (G) and fifth through seventh (NGL) vary in one repeat each. These differences between repeats are caused both by amino acid replacements and by deletions. Despite these differences and the prevalence of glycine residues which could lead to matches of no particular significance, we are confident that these postulated repeats are real; they contained all four of the W residues present in m2774, three of the four N residues, and three of the five L residues, all in the same order. In location and variability, if not in sequence, this repeat array is reminiscent of the IGYGEG repeats of the A4--c5 sequence of *A. pernyi* (see below).

Immediately after the heptapeptide repeats of m2774 was the sequence DGGYGG. Conceivably, this sequence may also be a variant of the GGGWNGL repeat (see below). A CG dipeptide followed, as in pc292 and pc609 immediately preceding the central region.

Since the m2774 sequence is incomplete at the amino-terminal end, we do not know whether any CG or LGYGG sequences are present there. As already noted, however, the missing sequence cannot be longer than 10 amino acid residues.

DISCUSSION

Age of the A family. The evidence presented clearly establishes that the A family of chorion proteins is at least as old as the last common ancestor of *B. mori* and the genus *Antheraea*.

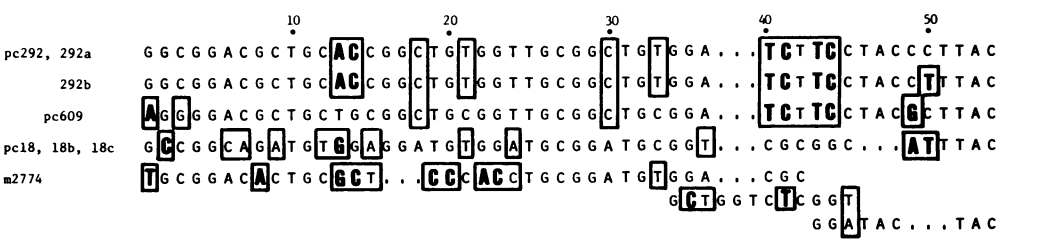


FIG. 5. Comparison of nucleotide sequences of the A family in the right arm. Nucleotides are numbered arbitrarily from the 5′ end of the arm. All known A sequences are presented. Deletions and insertions are indicated by dots, and reduplications are indicated by vertical stacking of subsequences which should be read from left to right, top to bottom. Substitutions are boxed and are shown in boldface if they lead to amino acid replacements. The alignment of m2774 with the remaining sequences is not unambiguous. As shown, it suggests that the apparent insertion of four codons in m2774 (see Fig. 1) is the result of two reduplications of ancestral sequences (corresponding to positions 34 to 42 and 42 to 44, respectively); it is also possible that these four codons arose as duplications of positions 29 to 42 and 42 to 44 plus a deletion corresponding to positions 33 to 34 or as an outright insertion.

These moths represent two families of the superfamily Bombycoidea (Bombycidae and Saturniidae, respectively). Although the absence of a pertinent fossil record precludes exact estimates, it is very probable that these moth families are at least 50,000,000 years old. This estimate is based on the following facts (2, 8, 16). The earliest lepidoptera (micropterygids) are known from the lower Cretaceous; many modern lepidopteran families were extant in the Eocene, and, in fact, most families may have originated in an earlier adaptive radiation associated with that of the angiosperms; finally, a distinctly saturniid fossil is known from the lower Oligocene, i.e., approximately 40,000,000 years ago (12).

It is also notable that all three of the sequenced *A. pernyi* proteins, which were selected simply by purity and which are representative of the A family in *A. pernyi* according to size and amino acid composition have features characteristic of either the AI or the AII subfamily of *A. polyphemus*. Clearly, these A protein subfamilies, AI and AII, are at least as old as the two *Antheraea* species. Again, in the absence of a fossil record, it is not possible to estimate that age exactly. It is known, however, that *A. pernyi* and *A. polyphemus* are distant branches of the genus *Antheraea*; the former is a Palearctic (Chinese) species and the latter a Nearctic (North American) species formerly classified in a distinct genus, *Telea* (11). Determination of overall single-copy DNA sequence divergence through melting analysis of genomic DNA hybrids will probably be the best means of estimating the ages of the various moth species and families. An independent indication of the age of the AI and AII protein subfamilies can be obtained by comparing their sequence divergence from each other and from the m2774 sequence. In the central region (144 base pairs), where comparisons can be made most confidently (Fig. 4), the three *A. polyphemus* AI sequences differ from the four AII sequences by 31 to 33 base substitutions, of which 18 to 20 are silent; both subfamilies differ from m2774 by 37 to 42 base substitutions, of which 19 to 24 are silent. We conclude that, to the extent that the available sequences are representative, the AI and AII subfamilies separated from each other not too long after their last common ancestor with m2774, which, of course, must have existed before the separation of the bombycid and saturniid lineages. Clearly, both the A chorion family as a whole and its AI and AII subfamilies are reasonably ancient.

In *A. polyphemus*, the AII subfamily appears to be developmentally "middle," and the AI subfamily appears to be developmentally "late" (17; Regier and Kafatos, manuscript in preparation). It will be interesting to determine whether that is also true in *A. pernyi*, or whether specific A genes or groups of genes have changed their developmental specificity in different saturniid lineages.

Domain structure of the A proteins. The central region and flanking arms of A proteins were initially defined by a combination of secondary structure predictions and compositional and sequence comparisons within *A. polyphemus* (3a, 19). The central region is thought to function as a "core" that builds up chorion fibers; it was identified as a sequence segment which is highly conserved, highly structured in the β-pleated sheet conformation, devoid of internal peptide repeats, and enriched in V and A residues. The remaining left and right arms were characterized as more variable, less structured segments, enriched in G plus C, L, or Y residues, and marked by internal tandemly repetitive peptides. The interspecies comparisons confirm and highlight these features. In particular, the strict conservation of the central region even in m2774 is in marked contrast to the relatively prominent sequence differences (including deletions, insertions, and duplications) in the arms; the importance of tandem arrays within the left arm is underscored by the observation of novel types of repeats in A4--c5 and m2774.

The sequence variations within the left arm are of special interest. The proximal (to the amino terminus) AI and AII subfamily-specific features of *A. polyphemus* are maintained in the *A. pernyi* sequences in the face of considerable sequence diversification; presumably, these features serve subfamily-specific functions important in both *Antheraea* species. Conversely, distal sequences at a site near the border with the central region, which are quite variable in *A. polyphemus*, are even more variable in *A. pernyi* and *B. mori* (Fig. 1). We do not know the function of this apparent "hypervariable" site in the protein structure. It might serve as a variable hinge between the highly structured central region and the proximal, tandemly repetitive segment of the arm. In any case, it appears that it is at this hypervariable site that new tandem repeat arrays have appeared during the evolution of the A family.

Mechanisms of evolutionary diversification in the A family. From comparisons between various *A. polyphemus* sequences, it has been concluded that, in addition to base substitutions, chorion sequences often accept "segmental mutations" of a special type: deletions and duplications which are usually associated with small direct DNA repeats, both tandem and nontandem (6; Jones, Ph.D. thesis; Jones and Kafatos, submitted for publication). This conclusion is considerably strengthened by the interspecies

comparisons reported here. Both substitutions and segmental mutations have been accepted in the arms, whereas the central region has evolved only by substitutions, presumably because it is strongly constrained in terms of protein structure.

A much larger data base is necessary before we can be certain of how the new repeat arrays of the left arm originated. However, we favor the following model, which is implicit in the alignments shown in Fig. 1.

We note that the only W residue anywhere within the AII subfamily of *A. polyphemus* is found in the hypervariable site, embedded within a GXGWXGX sequence. Thus, it is not implausible that the GGGWNGL repeat array of m2774 arose from such a sequence by modification, tandem reduplication, and further modification. On the other hand, the hypervariable site in 18c includes a Y rather than a W residue. This replacement suggests that the DGGYGG sequence of m2774 may also be a modified GGGWNGL repeat (especially since both Y and W are aromatic residues and since replacements of G by the acidic amino acid E or D are not uncommon in chorion; Fig. 1). Of course, if we accept this interpretation, we cannot specify whether the common ancestor of m2774 and the *A. polyphemus* sequences contained a W or a Y at this site: a Y/W replacement would have occurred independently at least twice during evolution.

The observation of a Y residue at this site in 18c is also important in suggesting that, in an independent but similar process, the same hypervariable sequence also gave rise to the IGYGEG repeat array of the *A. pernyi* A4--c5 protein. Furthermore, it is interesting how the hypervariable site differs even within the *A. polyphemus* AI subfamily. Although in 18c it includes the tripeptide AGY, in 18b and pc18 only T is found (codon ACT), which differs from I by a single base substitution (possible codon ATT, ATC, or ATA). Finally, IGYGEG is also reminiscent of the preceding LGYGG tandem repeats. Thus, we can suggest the following admittedly tentative picture; the hypervariable site was originally derived from the preceding LGYGG array and gave rise to new repeat arrays (IGYGEG or GGGWNGL) by reduplication and sequence divergence, and to various other sequences (in *A. polyphemus* and *A. pernyi*) by point mutations and small deletions.

Throughout the sequences considered here, generation of differences does not involve introns; none of the chromosomal chorion genes studied have an intron anywhere other than in the signal peptide-encoding region (5). The major differences in the hypervariable site apparently originated as segmental mutations, i.e., as deletions and duplications such as those which occur with reasonably high frequency during evolution even in other parts of the chorion genes (Jones and Kafatos, submitted for publication). We do not know whether the hypervariable site merely accepts such segmental mutations at a higher than average rate or somehow even enhances their occurrence.

ACKNOWLEDGMENTS

We thank F. M. Carpenter and C. D. Michener for discussions on phylogeny, J. C. Regier for advice on protein purification and sequencing, and C. W. Jones for advice on DNA sequencing and for permission to present the unpublished *A. polyphemus* sequences. We thank A. Georgi for technical assistance, C. Phillips for graphics, B. Klumpar for photography, and S. Foy for secretarial work.

The work was supported by grants from the National Institutes of Health, the National Science Foundation, and the University of Athens (to F.C.K.), from the National Institutes of Health and NATO (to G.C.R.), and from the Greek Service for Scientific Research and Technology (to N.K.M.).

LITERATURE CITED

1. **Brauer, A. W., M. N. Margolies, and E. Haber.** 1975. The application of 0.1 M quadrol to the microsequence of proteins and the sequence of tryptic peptides. Biochemistry **14**:3029–3035.
2. **Common, I. F. B.** 1975. Evolution and classification of the lepidoptera. Annu. Rev. Entomol. **20**:183–203.
3. **Efstratiadis, A., and F. C. Kafatos.** 1976. The chorion of insects: techniques and perspectives. Methods Mol. Biol. **8**:1–124.

3a. **Hamodrakas, S. J., C. W. Jones, and F. C. Kafatos.** 1982. Secondary structure predictions for silkmoth chorion proteins. Biochim. Biophys. Acta **700**:42–51.

4. **Iatrou, K., S. G. Tsitilou, M. R. Goldsmith, and F. C. Kafatos.** 1980. Molecular analysis of the Gr^B mutation in *Bombyx mori* through the use of a chorion cDNA library. Cell **20**:659–669.
5. **Jones, C. W., and F. C. Kafatos.** 1980. Structure, organization and evolution of developmentally regulated chorion genes in a silkmoth. Cell **22**:855–867.
6. **Jones, C. W., N. Rosenthal, G. C. Rodakis, and F. C. Kafatos.** 1979. Evolution of two major chorion multigene families are inferred from cloned cDNA and protein sequences. Cell **18**:1317–1332.
7. **Kafatos, F. C., J. C. Regier, G. D. Mazur, M. R. Nadel, H. M. Blau, W. H. Petri, A. R. Wyman, R. E. Gelinas, P. B. Moore, M. Paul, A. Efstratiadis, J. N. Vournakis, M. R. Goldsmith, J. R. Hunsley, B. Baker, J. Nardi, and M. Koehler.** 1977. The eggshell of insects: differentiation-specific proteins and the control of their synthesis and accumulation during development, p. 45–145. *In* W. Beermann (ed.), Results and problems in cell differentiation, vol. 8. Springer-Verlag, Berlin.
8. **Laurentiaux, D.** 1953. Classe des insectes, p. 397–527. *In* J. Piveteau (ed.), Traite de paleontologie, vol. 3. Masson, Paris.
9. **Long, E. O., and I. B. David.** 1980. Repeated genes in eukaryotes. Annu. Rev. Biochem. **49**:727–764.
10. **Maxam, A., and W. Gilbert.** 1977. A new method for sequencing DNA. Proc. Natl. Acad. Sci. U.S.A. **74**:560–564.
11. **Michener, C. D.** 1952. The Saturniidae (Lepidoptera) of the western hemisphere. Bull. Am. Mus. Natl. Hist. **98**:339–501.
12. **Packard, A. S.** 1974. Monograph of the bombycine moths of North America. III. Memoirs Natl. Acad. Sci. U.S.A. **12**:271.

13. **Proudfoot, N. J., and G. G. Brownlee.** 1976. 3′ noncoding region sequences in eukaryotic messenger RNA. Nature (London) **263**:211–214.
14. **Regier, J. C., F. C. Kafatos, R. Goodfliesh, and L. Hood.** 1978. Silkmoth chorion proteins: sequence analysis of the products of a multigene family. Proc. Natl. Acad. Sci. U.S.A. **75**:390–394.
15. **Regier, J. C., F. C. Kafatos, K. J. Kramer, R. L. Heinrikson, and P. S. Keim.** 1978. Silkmoth chorion proteins: their diversity, amino acid composition, and the amino terminal sequence of one component. J. Biol. Chem. **253**:1305–1314.
16. **Riek, E. F.** 1970. Fossil history, p. 168–186 *In* CSIRO (ed.), The insects of Australia. Melbourne University Press, Carlton, Australia.
17. **Sim, G. K., F. C. Kafatos, C. W. Jones, M. D. Koehler, A. Efstratiadis, and T. Maniatis.** 1979. Use of a cDNA library for studies on evolution and developmental expression of the chorion multigene families. Cell **18**:1303–1316.
18. **Terhorst, C., P. Parham, D. L. Mann, and J. L. Strominger.** 1976. Structure of HLA antigens: amino-acid and carbohydrate compositions and NH_2-terminal sequences of four antigen preparations. Proc. Natl. Acad. Sci. U.S.A. **73**:910–914.
19. **Tsitilou, S. G., J. C. Regier, and F. C. Kafatos.** 1980. Selection and sequence analysis of a cDNA clone encoding a known chorion protein of the A family. Nucleic Acids Res. **8**:1987–1997.