Evolution and higher-order structure of architectural proteins in silkmoth chorion

Jerome C.Regier

Department of Biochemistry, Molecular Biology and Cell Biology,
Northwestern University, 2153 Sheridan Road, Evanston, IL 60201, USA

Communicated by F.C.Kafatos

Genomic and cDNA clones have been sequenced that encode the E2 silkmoth chorion protein. E2 assembles with E1 [Regier, J.C. and Pacholski, P. (1985) *Proc. Natl. Acad. Sci. USA*, 82, 6035–6039] to form the ‘filler’ that helps mold prominent chorion surface structures called aeropyle crowns. E2 has two distinct domains. The amino terminal domain consists of four alternating stretches of hydrophobic and hydrophilic residues, the first three of which are homologous in sequence to about half of the E1 protein. Comparison of predicted secondary structures provides further support for the localized homology of E2 and E1. The carboxy terminal domain of E2 is much longer, is hydrophilic and consists entirely of multiple tandem copies of a single, variant hexapeptide repeat sequence that is absent from E1. Numbers of hexapeptide repeat sequences differed dramatically in two animals. The types of events required for such variation are discussed. Finally, we have elaborated our earlier model for how E proteins may assemble *in vivo* to form filler.

Key words: protein assembly/repetitive polypeptide/secondary structure prediction/sequence analysis/structural protein

Introduction

A molecular understanding of eucaryotic morphogenesis requires structural information about the components and knowledge of how individual structural genes are regulated in time, space and amount. Studies of the silkmoth chorion serve as a paradigm for such an approach. Its morphogenesis entails the temporally regulated production of more than 100 follicle cell-specific proteins in widely varying amounts (Paul and Kafatos, 1975; Regier *et al.*, 1980). Chorion proteins assemble to form two predominant types of structures visible in the electron microscope: (i) lamellae that are highly ordered helicoidal arrays of fibrils, and (ii) filler that is a sponge-like network (Mazur *et al.*, 1982; Regier *et al.*, 1982). Lamellar sequences number more than 100 and have been grouped into six multigene families, based on extensive sequence and cross-hybridization analysis (Goldsmith and Kafatos, 1984; Eickbush *et al.*, 1985; Regier and Kafatos, 1985). Weak, but still convincing, homologies are detectable among different families, based largely on the highly conserved central domain of all lamellar sequences. At the protein level the central domain has a highly ordered secondary structure that is thought to form the fibrillar core (Hamodrakas *et al.*, 1985). The more divergent amino and carboxy terminal arm sequences may serve more specific functions, e.g. determining the pitch of the helicoidal array or the thickness of individual fibrils (Mazur *et al.*, 1982). Lamellar proteins account for the large majority of the chorion's mass. Overlapping subsets of lamellar proteins are synthesized at all stages of choriogenesis – during the early

period when the lamellar framework forms, during the middle and late periods when the framework expands and densifies, and during the very late period when a few new lamellae form in a spatially restricted portion of the follicle called the aeropyle crown region (Mazur *et al.*, 1980; Regier *et al.*, 1982).

Filler accounts for roughly 5% of the chorion's mass and consists of only two proteins, named E1 and E2 (Mazur *et al.*, 1980). E1 and E2 cross-link shortly after synthesis and are deposited specifically at sites underlying three-cell junctions, until the very end of choriogenesis (Regier and Kafatos, 1985). By contrast, lamellar proteins are deposited uniformly over the cell surface. In the *Antheraea polyphemus* silkmoth major E1 and E2 protein synthesis is restricted in time to the very late period and in space to the aeropyle crown region. Filler functions to hollow breathing channels through lamellar chorion and to mold a small number of outer surface lamellae into crown-shaped structures called aeropyle crowns. In the region lacking aeropyle crowns, called the flat region, the small amount of filler serves simply to form breathing channels.

Recently, E1 cDNA and genomic clones have been sequenced (Regier and Pacholski, 1985). No homology was detected with lamellar sequences. Interestingly, secondary structure predictions and hydropathy profiles revealed a periodicity that was postulated to have functional significance.

In many ways an understanding of filler morphogenesis can serve the same purposes mentioned at the outset as studying the chorion as a whole, with the advantage that it is much less complex and has the additional interesting feature of aeropyle crown region specificity. Recent studies have shown that E1 and E2 genes are expressed at the RNA and protein levels with similar temporal- and cell-specificities (Mazur *et al.*, 1980; Regier *et al.*, 1986). Furthermore, these two types of genes are present in low gene copy number and are closely linked on the chromosome (Hatzopoulos and Regier, 1986). What we describe here is the complete primary structure of the coding region of the E2 gene and its nearly full-length cDNA. We have compared their structure with that of E1 and lamellar sequences. This has provided insight into the evolution of filler sequences. With the completion of the E2 sequence, we now know the protein sequences for both components of filler and are able to make certain suggestions about how these proteins might interact and assemble.

Results

Sequencing strategy for the E2 gene and cDNA

The entire E2 gene plus flanking regions and a near full-length E2 cDNA (5295 and 1890 bp, respectively) have been sequenced. In many cases, both strands were sequenced; otherwise, the same strand was sequenced independently at least twice. There are no ambiguities in the sequence.

In this report we will present and discuss the coding region sequences of the gene and cDNA (1683 and 1855 bp, respectively; Figure 1). Non-coding sequences will be presented elsewhere (Regier *et al.*, 1986). E2 has a single intron within its

Fig. 1. Nucleotide sequences of E2 gene and cDNA coding regions plus their conceptualized translation products. The nucleotide sequence of the coding region of the E2 gene is presented along with its conceptualized translation product. Numbering of nucleotides (above the sequence) and of amino acids (below the sequence) begins from the initiating methionine codon and residue. The cDNA sequence is displayed above the gene sequence. The 5' end of the cDNA is incomplete and terminates within the signal-peptide-encoding region (see below). Identical nucleotides are indicated by a dot. Otherwise, differences are displayed as nucleotides interrupting the dotted line. Nucleotide differences that result in amino acid changes are shown within parentheses below the gene's translation product. The cDNA contains a single insertion of 216 bp which can be mapped onto the gene sequence anywhere after nucleotide 970 and before nucleotide 1001. This latitude results from the location of the insertion within a highly repeated region as discussed in the text. We have placed the insertion between nucleotides 972 and 973 of the gene. The cDNA insertion sequence is not numbered. The triangle identifies the location of an intron. The three most probable sites separating signal peptide from mature protein are each identified by arrows. Hexapeptide sequences that share four out of six residues with the sequence Lys Lys Asp (Gly/Asp) (Val/Glu) (Gly/Ser) are underlined. Thin vertical lines beginning between residues 127 and 128 demarcate the remainder of the sequence into more variable repeating subunits that include the above. All repeats contain 18 nucleotides/six residues except at the carboxy terminus and where indicated by double-headed arrows.

Table 1. Comparison of amino acid composition of purified E2 protein and that deduced from cloned gene and cDNA (residues per 100 residues)

	E2 protein ^a	Chromosomal clone ^b	cDNA clone ^b
Cys	0.8	1.1	1.1
Asx	22.3	22.3 ^c	22.6 ^c
Thr	4.7	5.2	4.4
Ser	7.4	9.2	9.0
Glx	5.7	5.2 ^d	5.2 ^d
Pro	1.9	1.8	1.6
Gly	19.1	18.5	18.1
Ala	4.6	4.1	3.9
Val	9.4	9.0	9.4
Met	0.6	0.7	0.8
Ile	4.5	4.6	4.1
Leu	3.9	3.7	3.3
Tyr	0.7	0.7	0.5
Phe	0.2	0.2	0.2
His	0.4	0.4	0.5
Lys	11.9	11.6	14.0
Trp	0.0	0.0	0.0
Arg	1.7	1.7	1.3

^aFrom Regier *et al.* (1980).^bChromosomal clone: includes residues 20–561 of conceptually translated sequence shown in Figure 1; cDNA clone: includes residues 20–561 and 72 residue insert of conceptually translated sequence shown in Figure 1.^cChromosomal clone: 11.2 Asn + 11.1 Asp; cDNA clone: 9.9 Asn + 12.7 Asp.^dChromosomal clone: 1.1 Gln + 4.1 Glu; cDNA clone: 1.0 Gln + 4.2 Glu.

signal peptide-coding region, like all other chorion genes analyzed to date (see below; Jones and Kafatos, 1980; Iatrou and Tsitilou, 1983).

Sequence of the E2 gene and cDNA coding regions

An open reading frame of 1638 nucleotides is found in the E2 gene (nucleotides 46 to 1683 in Figure 1), which, when coupled with 45 nucleotides in the 5' exon, encodes a polypeptide of 561 residues. Correct localization of exon splice sites is based on overlap by the cDNA clone (Figure 1), on the presence of appropriate consensus sequences at the borders of the intron (GTAAG...TTTATTTTCAG; Mount, 1982) and multiple translational termination signals therein. Identification of the initiating methionine residue is based on its presence at the start of a signal peptide-like sequence, on the absence of other upstream methionine codons not followed by potential translational termination signals, and on the presence of a likely TATA box 40 nucleotides upstream (data not shown). Three positions between residues 17 and 25 are identified as potential cleavage sites for signal peptidase based on their agreement with the (–3, –1) rule of von Heijne (1984). Other chorion sequences are known to be cleaved between the fourth and fifth residues following the intron, consistent with one of our three possibilities. The molecular weight of the mature E2 gene product is 54 542, assuming signal peptide cleavage occurs after residue 19.

The 1858-nucleotide-long coding portion of the E2 cDNA shares extensive homology with the gene; there are 10 point substitutions, five of which result in amino acid replacements, and one insertion (or deletion in the gene) of 216 nucleotides which occurs between nucleotides 973 and 974 of the gene sequence. The molecular weight of the mature E2 cDNA product is 62 631.

A likely explanation for the observed differences between E2

gene and cDNA is that they represent alleles, since the two clones were derived from separate animals and population polymorphism is common among chorion sequences (Regier *et al.*, 1978; Jones and Kafatos, 1981, 1982). Alternatively, they could represent distinct genes. Southern hybridization analysis of genomic DNA and of E2 genomic clones is most consistent with two copies per haploid genome. These copies are very similar in primary structure (Hatzopoulos and Regier, 1986).

The deduced amino acid compositions of the E2 gene and cDNA sequences closely match that of the purified protein (Table 1). E2 is unusual in having large amounts of aspartic acid, asparagine, glycine and lysine.

Internally repetitive regions

The E2 protein is more than four times larger than E1 and most other chorion proteins (Regier and Kafatos, 1985). The superabundance of short, internally repetitive sequences is the major explanation for this size difference (Figure 1). The region of repeats in the polypeptide encoded by the E2 gene includes 77% of the sequence, from about residue 128 to the carboxy terminus. Many of the repeats have diverged substantially and are not obvious without careful analysis, although it is clear from Figure 1 that most (66 out of 74) are hexapeptides and most (64 out of 74) begin either with asparagine or lysine. Repeats in the central portion are most conserved and will be discussed first (underlined in Figure 1).

The gene encodes 24 hexapeptides that are identical in at least four out of six residues with the sequence Lys Lys Asp (Gly/Asp) (Val/Glu) (Gly/Ser). Sixteen of these hexapeptides, from residues 277 to 372, are displayed in a fashion which illustrates that tandem triplets of hexapeptides form a larger repeating unit (Figure 2A). Hexapeptides are distinguished largely in their fourth, fifth and sixth residues.

The cDNA encodes the same set of tandem triplets of hexapeptides, with an additional 12 hexapeptides inserted between residues 324 and 325 (Figure 1). A careful inspection of the cDNA sequence in this region reveals two identical sets of 11 hexapeptides each (see brackets in Figure 2B).

A model for the evolution of the repetitive regions shown in Figure 2B is illustrated in Figure 2C. First, an ancestral sequence encoding the hexapeptide Lys Lys Asp (Gly/Asp) Val (Gly/Ser) triplicates. Substitutions occur such that all three hexapeptides become distinct. Following this, the 18-residue repeat, consisting of three hexapeptides, undergoes multiple (five are shown) rounds of complete replication and additional substitutions to form a sequence identical to the gene sequence. Finally, a small partial duplication followed by a large reduplication and one silent substitution that distinguishes the two large reduplicated portions occur, resulting in the cDNA sequence.

The observed length polymorphism in the repeated regions of the E2 gene and cDNA is not unexpected as they are derived from different animals and similar length polymorphisms are not uncommon in other genes (Kasarda *et al.*, 1984; Pustell *et al.*, 1984). Even so, we were concerned that it might reflect an artifact of cloning or subcloning. To test whether a deletion occurred in the gene during cloning, we restricted eight independently isolated E2 gene-containing chromosomal clones and two subclones used for sequence analysis (all from the same animal) in such a way as to yield single fragments that contained most of the repeat sequences and little else. After gel electrophoresis, we probed these samples with a similar end-labeled cDNA fragment that contains mostly repeat sequences to look for differences in lengths of repeated regions (Figure 3). In all

		LYSINE REPEATS																	
		(position within repeat)																	
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
A		22	24	1	23	21	5	1	24	0	2	9	0	5	0	1	5	1	1
G		1	0	19	1	0	19	23	0	0	21	14	1	18	0	0	18	23	0
C		0	0	2	0	0	0	0	2	1	0	6	0	1	1	0	0	0	1
T		1	0	2	0	3	0	0	0	22	0	1	17	1	23	22	1	0	22
consensus→		A	A	G	A	A	G	G	A	T	G	G	T	G	T	T	G	G	T
		A																	
		*	*		*			*		*	*	*	*	*	*	*	*	*	*

		ASPARAGINE REPEATS																	
		(position within repeat)																	
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
A		45	43	8	16	9	24	19	12	14	13	10	10	16	10	20	13	22	11
G		3	6	1	26	8	5	26	25	2	28	28	5	23	5	5	31	14	0
C		2	1	5	3	19	5	3	7	2	4	4	8	1	10	9	0	5	8
T		0	0	36	5	14	16	1	5	31	2	5	24	5	20	11	1	4	26
consensus→		A	A	T	G		A	G	G	T	G	G	T	G	T	A	G	A	T
					A		T	A						A		T		G	

Fig. 4. Comparison of base compositions of lysine and asparagine repeats. The cumulative base compositions of lysine and asparagine repeat sequences (see text for definitions) are shown for each of the 18 positions in the upper and lower panels, respectively. Consensus sequences are shown at the bottom of each panel. Consensus bases represent at least 44% of the total at any given position. Additional consensus bases are included below if they occur more than 50% as often as the major consensus bases. Asterisks identify those positions where consensus bases match for lysine and asparagine repeats.

of 24 lysine repeats in the gene (or 31 out of 36 in the cDNA) satisfy the minimal condition, and 14 are perfectly duplicated in the gene (or 28 in the cDNA).

The lysine and asparagine repeats are clearly homologous, as is most evident by comparing their composite nucleotide compositions at each of the 18 nucleotide positions within the repeat sequence (Figure 4). Consensus sequences for the two types of repeats match in 14 of the 18 positions.

Hydrophobicity and secondary structure prediction

Kyte and Doolittle (1982) described a method which progressively evaluates the hydrophobicity and hydrophilicity of a protein along its amino acid sequence as a set of hydropathy indices. Over the first 127 residues of the mature E2 sequence, the hydropathy index reveals four roughly equally spaced peaks of hydrophobicity (peak-to-peak distances about 28 residues in all cases), separated by valleys of hydrophilicity (Figure 5A). The remainder of the E2 sequence is generally hydrophilic, with only a few short stretches of residues with positive hydropathic indices.

We have also predicted the secondary structure of the E2 polypeptide using two distinct probabilistic methods (Figure 5A). Both methods predict hydrophobic peak 1 to be organized as β -sheet and peak 2 as α -helix. Peak 3 is predicted to contain a short stretch of both α -helix and β -sheet, and peak 4 to be largely β -turn with a short stretch of β -sheet. The regions between hydrophobic peaks 1 and 2, between peaks 3 and 4, and immediately after peak 4 are predicted to contain β -turns.

Secondary structure predictions of the repetitive region (after residue 127) are more problematic. For one, predictions by the two methods for β -sheet are not in close agreement. For another,

long stretches of uninterrupted β -turn secondary structure predicted by both methods may not be stable.

Comparison of E2 with other chorion sequences

The best-fit alignment of E2 with E1, the other component of filler in the silkmoth chorion, reveals convincing localized homologies at their amino terminal ends (Figure 6). Three stretches, 12–21 residues in length, are each 62% identical, or more than 85% if similar residues are included. By contrast, only 20% of residues match outside these stretches for the sequences shown. The best-fit alignment between E1 and the carboxy terminal portion of E2 has less than 10% matching.

To evaluate the statistical significance of the similarities between E2 and E1, we have compared the first 98 residues of E2 with 1000 randomly permuted versions of the first 95 residues of E1, using a recently described algorithm (Lipman and Pearson, 1985). The similarity score for a comparison of the actual E1 sequence with E2 (score = 145) was about 18 SD above the mean for similarity scores of the permuted versions (range = 14–58). The significance of the similarity is expressed as a z value, with values greater than 10 judged 'significant'. For comparison of E2 (residues 1–98) and E1 (residues 1–95), the z value is 19.6. The region of homology extends to residue 85 (for E1) and 88 (for E2). Beyond this there are no significant similarities.

The three stretches of greatest sequence homology between E2 and E1 overlap extensively with portions of the first three hydrophobic peak sequences and their hydrophobic spacers in both E2 (Figure 5A) and in E1 (Figure 5B; Regier and Pacholski, 1985). Likewise, the secondary structure predictions of these

Fig. 5. Higher-order structure of the E2 and E1 proteins. (A) The average hydrophobicity/hydrophilicity of overlapping nonapeptides (called the hydropathic index; Kyte and Doolittle, 1982) of the E2 cDNA translation product is displayed. Values for each nonapeptide are positioned over each fifth residue beginning at residue 20, the presumed amino terminal end of the mature protein. Bars with cross-hatching identify four relatively hydrophobic segments. Solid bars identify three stretches of residues that are relatively conserved in the E1 protein (see Figure 6). Brackets identify tandem 66 residue repeats. Secondary structures (α , α -helix; β , β -sheet; T, β -turn) are derived from prediction schemes of Garnier *et al.* (1978; see thin lines) and Chou and Fasman (1978, 1979; see thick lines). In the scheme of Garnier *et al.*, decision constants for α -helix and β -sheet were set at 158 and 50, respectively. In the Chou and Fasman scheme, β -turns were predicted when the conformational parameters $P_i > 1.00$, P_α and P_β and when the products of positional frequencies for adjacent tetrapeptides (P_i) $> 1.30 \times 10^{-4}$. The Chou and Fasman parameters are the same as those used for prediction of the E1 protein secondary structure (Regier and Pacholski, 1985). Vertical arrows identify cysteine residues. Numbering of the sequence is the same as in Figure 1, including the unnumbered insertion between positions 324 and 325. (B) The hydropathy plot for E1 (residues 20–169) is displayed, with four small gaps introduced to improve alignment of the sequence with E2 (see Figure 6). Bars with cross-hatching identify five relatively hydrophobic segments. Brackets identify tandem tetradecapeptide repeats. Modified from Regier and Pacholski, 1985.

regions are very similar for E2 and E1 when both predictive methods are in agreement — α -helix or β -sheet for hydrophobic peaks 1–3 separated by hydrophilic valleys that contain β -turn (between peaks 1 and 2) or no predicted secondary structure (between peaks 2 and 3). Finally, both proteins have cysteine residues in valleys 1 and 3, as well as in peak 4 (Figure 5; Regier and Pacholski, 1985). It seems reasonable that E2 and E1 have maintained similar higher-order structures in the regions of sequence homology.

We searched for homology between the nonrepetitive region of E2 (residues 1–127) and all other sequences found in three

large databases (The European Molecular Biology Laboratory Nucleotide Sequence Data Library, The National Institutes of Health DNA Sequence Library, the National Biomedical Research Foundation's Protein Sequence Database), using the IFIND software program available through the Bionet Resource (c/o IntelliGenetics, 1975 Camino Real West, Mountain View, CA 94040-2216, USA). E1 gave the highest similarity score for sequences in all three databases and the only one whose score did not form part of the continuum of other scores. For the nucleotide search, the E1 score was 82 (word size = 2, gap penalty = 2, fast = NO).

Fig. 6. Homology between E1 and E2 protein sequences. The best alignment between E1 (upper line) and the amino terminal domain of E2, including signal peptide (lower line) is presented. The carboxy terminal portion of E1 (after residue 123) does not align with the amino terminal domain of E2 and its sequence is not shown. A vertical line is drawn between two residues when they match. Brackets identify three stretches of residues conserved between E1 and E2. The 1-letter amino acid code is used. Numbering begins from the initiating methionine residue.

Fig. 7. Comparisons that suggest the presence of repeated subunits within the amino terminal domain. Residues found in and around hydrophobic peaks 2 (first and third lines) and 3 (second and fourth lines) of E1 (Regier and Pacholski, 1985) and E2 are aligned (see Figure 5). Positions where there are identical residues in hydrophobic peaks 2 and 3 are boxed. Gaps are indicated by -. Numbering is the same as in Figure 6.

We searched for similar lysine- and asparagine-like hexapeptide repeats in other sequences, using the QUEST software program from Bionet. For the protein searches, we looked for at least three tandem repeats of Lys Lys Asp ... or four of Asn ..., where . could be any amino acid. None were found. For the nucleotide searches, we looked for coding regions that contained at least three tandem repeats of AA·AA·GA· ... or four of AA[T or C] ... For the first octadecanucleotide, no sequences were found. For the second, four groups of closely related sequences were found that could be translated in the appropriate open reading frame. One (plasmodium surface antigen) encoded multiple tetrapeptide repeats of Asn Ala Asn Pro. Another (trypanosome ribosomal protein) contained long stretches of polyasparagine. The last two (drosophila GAG-like protein and yeast positive regulatory protein) contained no evidence of being within a repeat region other than for the presence of the four asparagine codons. We conclude that the lysine and asparagine repeats of E2 have not been found in other sequences.

Finally, we observe that the region of homology for E2 and E1 contains an obvious repetitive pattern in its hydrophathy plot as well in certain features of its secondary structure, i.e. predictions of α -helix and β -sheet by both methods are confined to the three hydrophobic peaks (Figure 5; Regier and Pacholski, 1985). This could reflect an underlying repetitive pattern of sequence homology, consistent with the idea that this region formed by a series of partial gene duplications. To test this hypothesis, we carefully searched for sequence homologies around different hydrophobic peaks from both E1 and E2. To our surprise we observed a striking degree of similarity around peaks 2 and 3 from both E1 and E2 (Figure 7). Eight out of 17 positions show matching, and differences in other positions are largely conservative. There are three pairwise sequence comparisons (E2, peak 2 $\times$ E2, peak 3; E2, peak 2 $\times$ E1, peak 3; E1, peak 2 $\times$ E1, peak 3) whose statistical significance can be judged by comparing one sequence to randomly permuted versions of the other. The comparison that yielded the highest z value (3.1) was between E2, peak 2 and E1, peak 3. According to Lipman and Pear-

Table II. Comparison of amino acid composition of E1 and E2 as deduced from cloned sequences (residues per 100 residues)

	E1 cDNA clone ^a (residues 17–123)	E2 cDNA clone ^b (residues 20–127)	E2 cDNA clone ^b (residues 128–end)
Cys	3.7	5.6	0.2
Asp	3.7	1.9	15.0
Thr	9.3	10.2	3.2
Ser	9.3	13.0	8.1
Glu	7.5	4.6	4.2
Asn	3.7	6.5	10.7
Gln	2.8	4.6	0.2
Pro	4.7	3.7	1.2
Gly	9.3	7.4	20.4
Ala	7.5	6.5	3.4
Val	8.4	4.6	10.5
Met	0.9	0.0	1.0
Ile	5.6	9.3	3.0
Leu	12.2	13.0	1.2
Tyr	3.7	1.9	0.2
Phe	0.0	0.0	0.2
His	0.0	0.9	0.4
Lys	3.7	3.7	16.2
Trp	0.0	0.0	0.0
Arg	3.7	2.8	1.0
$\Delta Q = 81^c$ $\Delta Q = 873$ $\Delta Q = 706$			

^aFrom Regier and Pacholski (1985).

^bSee Figure 1.

^c ΔQ values are an indicator of the degree of similarity between two proteins (Marchalonis and Weltman, 1971). They are calculated by separately squaring the difference in concentration of each amino acid and then summing. The smaller the number, the more similar the compositions. In a test sample, pairs of homologous proteins almost always had ΔQ values < 100 .

son (1985), this value is 'possibly significant'. At present, the comparisons are suggestive that sequences in and around hydrophobic peaks 2 and 3 may have arisen by duplication from a common ancestor.

Discussion

The evolution of *E* genes

The carboxy terminal three-fourths of E2 consists almost entirely of hexapeptide repeats that are evolutionarily related (Figures 1, 2 and 4). The central portion of this domain consists of the lysine repeats flanked on both sides by the asparagine repeats. The greater homogeneity of the lysine repeats suggests a more recent origin, perhaps by reduplication of a variant ancestral asparagine repeat. Variability in repeat length and in composition is greater at the ends of the repeat domain, as well as at the amino terminal border of the major lysine repeat (around residue 277). Variable repeat length may reflect later replication and/or deletion events that occurred out-of-register with the predominant hexapeptide sequence. Compositional heterogeneity may result from a lower rate of correction or homogenization at borders by unequal crossing-over or intergenic conversion (Smith, 1974; Petes, 1980; Klein and Petes, 1981).

The amino terminal domain of E2 has no obvious repeats. Its precise border with the repeat domain has been obscured through diversification of repeat sequences. We favour placing the border between residues 127 and 128, based on inspection of the sequence in Figure 1. However, it is possible that it could extend to residue 159.

The amino terminal 16% of E2 (up to residue 88) and 50% of E1 (up to residue 85) are clearly homologous (see Figures 5 and 6, and text). Beyond residue 88 of E2 there is no convincing evidence for homology with E1. In E2 residues 86–127 represents a relatively short transitional peptide (7% of total) between the region of homology with E1 and the hexapeptide domain.

How did E genes evolve? Most likely, the ancestral E gene encoded an amino terminal sequence similar to that in the modern-day E1 and E2. Whether this included part or all of the E2 transitional peptide or the E1 carboxy terminal half is unclear. After the ancestral gene duplicated *in toto*, one of the copies (the E2 gene ancestor) added on the hexapeptide repeat domain, either by reduplicating a sequence already present within the gene, perhaps in the ancestral transition zone, or by transposition of an unrelated sequence. Absence of an intron at the border of the repeat domain argues against an exon shuffling hypothesis. We are currently sequencing E genes from another silkworm species in hopes of better understanding the origin of the repeat domain.

A molecular model of filler

We have presented a model for how E1 protein might assemble to form filler, based on its sequence, hydropathy profile and predicted secondary structure (Regier and Pacholski, 1985). We suggested that peptides within hydrophobic peaks, which are predicted to form β -sheet and α -helix, are good candidates for assembling into the core of the filler ultrastructure visible in the electron microscope. Separating these generally hydrophobic residues are stretches of hydrophilic residues that act as spacers with hinges (the β -turns) that might permit a change in the orientation of the hydrophobic regions. Particular orientations of the hydrophobic regions could be stabilized by disulfide bonding of cysteine residues within spacers.

The E2 sequence supports this model and further suggests that its amino terminal domain may function in a way similar to that of E1. It is reasonable to suggest that E1 and E2 may bond to themselves and to each other via their first three, and possibly four, hydrophobic peak regions. Furthermore, their hydrophilic spacer regions may also interact, since disulfide bonds form between E1's, E2's and between E1 and E2 shortly after synthesis (Regier and Kafatos, 1985). We suggest that E1 and the amino terminal domain of E2 bond to more than one molecule to form an ordered but highly interconnected core network of filler. This would be consistent with the observation that filler can be washed off the surface of the chorion under denaturing, but nonreducing, conditions in an insoluble, highly polymerized form (J.Regier, unpublished observation). More divergence in hydrophilic valley sequences has been permitted during evolution than in the hydrophobic peak sequences, consistent with their principal role as spacers (Figures 5 and 6).

The higher-order structure of the repeat domain must be quite distinct from its amino terminal domain, given its rather uniformly hydrophilic nature and its distinct composition (Table II) and sequence. Two observations suggest that the repeat region may have a highly ordered structure. One is that the hexapeptide as a repeat unit is highly conserved. The other is that asparagine or lysine is present as the first residue in 64 out of 74 repeats, despite quite substantial diversification in other positions. With 37% charged residues (versus 14% for the amino terminal domain), individual repeat domains would appear well suited to hydrogen-bond extensively to themselves, to repeat domains of other E2's and/or to water. Bonding to water may be particularly impor-

tant as filler is loosely organized and collapses upon dehydration. These suggestions about the molecular nature of filler should now be tested experimentally.

Materials and methods

Sequence analysis

The construction of an *A. polyphemus* silkworm chorion cDNA library in the plasmid pBR322 and the isolation and characterization of E2-encoding cDNA clones have been described (Regier *et al.*, 1984). Clone pcvl 11 contained the longest insert [1890 bp minus oligo(G.C) tails] and was sequenced in its entirety after reinserting the insert into the plasmid pUC8 or after constructing M13 subclones (Messing *et al.*, 1981; Vieira and Messing, 1982). An *A. polyphemus* chromosomal clone, called AP1, was shown to contain a single copy of the E2 gene (Hatzopoulos and Regier, 1986). AP1 DNA was restricted with *Hind*III, subcloned into pBR322 and screened with nick-translated pcvl 11. All hybridization resided in a single subcloned fragment called AP1-s1, which had an insert size of 4766 bp. The entire sequence of this insert was deduced from analysis of pUC8 or M13 subclones, except for the 5' and 3' ends which were sequenced directly. Sequence analysis revealed that the 5' end of the E2 gene, including codons for the first 15 residues of the signal peptide, lay outside AP1-s1. Therefore, the relevant portion of the adjoining chromosomal fragment, subcloned in AP1-s2, was also sequenced. Sequencing was performed by the end-labeling/site-specific modification method (Maxam and Gilbert, 1980) and the M13/dideoxy chain termination method (Sanger *et al.*, 1977; Sanger and Coulson, 1978; Messing *et al.*, 1981; Biggin *et al.*, 1983).

Computer analysis

Sequence comparisons in Figures 1 and 2 were aided by the IBI/Pustell software programs adapted for the IBM PC (International Biotechnologies, Inc. user manual; Pustell and Kafatos, 1982; Pustell and Kafatos, 1984). The Microgenie software programs (Beckman Instrument's Microgenie user manual; Queen and Korn, 1984) were used for alignment of E1 and E2 coding regions (Figures 6 and 7). The Chou and Fasman (1978, 1979) and Garnier, Osguthorpe and Robson (1978) protein secondary structure predictive schemes were adapted for the VAX 11/785 computer and IBM PC, respectively.

Acknowledgements

I thank Mr John Watson and Ms Ann Durot for assistance with sequencing and Mr Nick Vlahos with computer analysis, Mr Antonis Hatzopoulos for providing clone AP1-s2 and for helpful discussions, Drs Israel Hanukoglu (Technion, Israel), Thomas Lalor (Yale University) and William Pearson (University of Virginia) for providing computer software, and Dr Gerald Fasman (Brandeis University) for helpful comments about secondary structure predictions. This work was supported by grants from the National Institutes of Health to JCR and to Bionet (No. 1 U41 RR-01685-02).

References

- Biggin, M.D., Gibson, T.J. and Hong, G.F. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 3963–3965.
- Chou, P.Y. and Fasman, G.D. (1978) *Adv. Enzymol. Related Areas Mol. Biol.*, **47**, 45–148.
- Chou, P.Y. and Fasman, G.D. (1979) *Biophys. J.*, **26**, 367–384.
- Eickbush, T.H., Rodakis, G.C., Lecanidou, R. and Kafatos, F.C. (1985) *Dev. Biol.*, **112**, 368–376.
- Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) *J. Mol. Biol.*, **120**, 97–120.
- Goldsmith, M.R. and Kafatos, F.C. (1984) *Annu. Rev. Genet.*, **18**, 443–487.
- Hamodrakas, S.J., Etmektzoglou, T. and Kafatos, F.C. (1985) *J. Mol. Biol.*, **186**, 583–589.
- Hatzopoulos, A.K. and Regier, J.C. (1986) *Mol. Cell. Biol.*, in press.
- Iatrou, K. and Tsitilou, S.G. (1983) *EMBO J.*, **2**, 1431–1440.
- Jones, C.W. and Kafatos, F.C. (1980) *Cell*, **22**, 855–867.
- Jones, C.W. and Kafatos, F.C. (1981) *Mol. Cell. Biol.*, **1**, 814–828.
- Jones, C.W. and Kafatos, F.C. (1982) *J. Mol. Evol.*, **19**, 87–103.
- Kasarda, D.D., Okita, T.W., Bernardin, J.E., Baecker, P.A., Nimmo, C.C., Lew, E.J.-L., Dietler, M.D. and Greene, F.C. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 4712–4716.
- Klein, H.L. and Petes, T.D. (1981) *Nature*, **289**, 144–148.
- Kyte, J. and Doolittle, R.F. (1982) *J. Mol. Biol.*, **157**, 105–132.
- Lipman, D.J. and Pearson, W.R. (1985) *Science*, **227**, 1435–1441.
- Marchalonis, J.J. and Weltman, J.E. (1971) *Comp. Biochem. Physiol.*, **38B**, 609–625.
- Maxam, A.M. and Gilbert, W. (1980) *Methods Enzymol.*, **65**, 499–560.

- Mazur, G.D., Regier, J.C. and Kafatos, F.C. (1980) *Dev. Biol.*, **76**, 305–321.
- Mazur, G.D., Regier, J.C. and Kafatos, F.C. (1982) In King, R.C. and Akai, H. (eds), *Insect Ultrastructure*, Plenum Press, New York, Vol. 1, pp. 150–185.
- Messing, J., Crea, R. and Seeburg, P.H. (1981) *Nucleic Acids Res.*, **9**, 309–321.
- Mount, S.M. (1982) *Nucleic Acids Res.*, **10**, 459–472.
- Paul, M. and Kafatos, F.C. (1975) *Dev. Biol.*, **42**, 141–159.
- Petes, T.D. (1980) *Cell*, **19**, 765–774.
- Pustell, J. and Kafatos, F.C. (1982) *Nucleic Acids Res.*, **10**, 4765–4782.
- Pustell, J. and Kafatos, F.C. (1984) *Nucleic Acids Res.*, **12**, 643–655.
- Pustell, J., Kafatos, F.C., Wobus, U. and Baumlein, H. (1984) *J. Mol. Evol.*, **20**, 281–295.
- Queen, C. and Korn, L.J. (1984) *Nucleic Acids Res.*, **12**, 581–599.
- Regier, J.C. and Kafatos, F.C. (1985) In Kerkut, G.A. and Gilbert, L.I. (eds), *Comprehensive Insect Physiology, Biochemistry and Pharmacology*. Pergamon Press, Oxford, Vol. 1, pp. 113–151.
- Regier, J.C. and Pacholski, P. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 6035–6039.
- Regier, J.C., Kafatos, F.C., Kramer, K.J., Heinrikson, R.L. and Keim, P.S. (1978) *J. Biol. Chem.*, **253**, 1305–1314.
- Regier, J.C., Mazur, G.D. and Kafatos, F.C. (1980) *Dev. Biol.*, **76**, 286–304.
- Regier, J.C., Mazur, G.D., Kafatos, F.C. and Paul, M. (1982) *Dev. Biol.*, **92**, 159–174.
- Regier, J.C., Hatzopoulos, A.K. and Durot, A.C. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 2796–2800.
- Regier, J.C., Hatzopoulos, A.K. and Durot, A.C. (1986) *Dev. Biol.*, in press.
- Sanger, F. and Coulson, A.R. (1978) *FEBS Lett.*, **87**, 107–110.
- Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463–5467.
- Smith, G.P. (1974) *Cold Spring Harbor Symp. Quant. Biol.*, **38**, 507–513.
- Southern, E.M. (1975) *J. Mol. Biol.*, **98**, 503–517.
- Vieira, J. and Messing, J. (1982) *Gene*, **19**, 259–269.
- von Heijne, G. (1984) *J. Mol. Biol.*, **173**, 243–251.

Received on 17 March 1986; revised on 3 June 1986