Copyright © 1991 by the Genetics Society of America

Sequence Identity in an Early Chorion Multigene Family Is the Result of Localized Gene Conversion

Barbara L. Hibner,[1] William D. Burke and Thomas H. Eickbush

Department of Biology, University of Rochester, Rochester, New York 14627

Manuscript received January 3, 1991
Accepted for publication March 16, 1991

ABSTRACT

The multigene families that encode the chorion (eggshell) of the silk moth, *Bombyx mori*, are closely linked on one chromosome. We report here the isolation and characterization of two segments, totaling 102 kb of genomic DNA, containing the genes expressed during the early period of choriogenesis. Most of these early genes can be divided into two multigene families, *ErA* and *ErB*, organized into five divergently transcribed *ErA/ErB* gene pairs. Nucleotide sequence identity in the major coding regions of the *ErA* genes was 96%, while nucleotide sequence identity for the *ErB* major coding regions was only 63%. Selection pressure on the encoded proteins cannot explain this difference in the level of sequence conservation between the *ErA* and *ErB* gene families, since when only fourfold redundant codon positions are considered, the divergence within the *ErA* genes is 8%, while the divergence within the *ErB* genes (corrected for multiple substitutions at the same site) is 110%. The high sequence identity of the *ErA* major exons can be explained by sequence exchange events similar to gene conversion localized to the major exon of the *ErA* genes. These gene conversions are correlated with the presence of clustered copies of the nucleotide sequence GGXGGX, encoding paired glycine residues. This sequence has previously been correlated with gradients of gene conversion that extend throughout the coding and noncoding regions of the *High-cysteine* (*Hc*) chorion genes of *B. mori*. We suggest that the difference in the extent of the conversion tracts in these gene families reflects a tendency for these recombination events to become localized over time to the protein encoding regions of the major exons.

MULTIGENE families of higher eukaryotes can exhibit high levels of sequence homogeneity within a species including uniform sequence features not present in related species. This concerted evolution can occur within sequences that appear to have no phenotypic effect on the host. In these cases concerted evolution has been suggested to operate distinct from natural selection in that the fixation of variants occurs by sequence exchange mechanisms that affect gene frequency in a non-Mendelian manner (SMITH 1973; OTHA 1980; DOVER 1982; ARNHEIM 1983). These exchange mechanisms include both reciprocal events (crossovers) and nonreciprocal events (gene conversion) (PETES 1980; SZOSTAK and WU 1980; JACKSON and FINK 1981; NAGYLAKI and PETES 1982).

The chorion locus of *Bombyx mori* is an excellent system in which to study the concerted evolution of multigene families. The chorion, or eggshell, proteins are encoded by over 150 genes (see review by GOLDSMITH and KAFATOS 1984), which can be placed into multigene families based on their sequence identities and period of expression (IATROU, TSITOLOU and KAFATOS 1982; EICKBUSH *et al.* 1985; LECANIDOU *et al.* 1986). Eggshell morphology and chorion protein composition are quite different between species of silk moths suggesting a rapid rate of evolution (reviewed in KAFATOS *et al.* 1977). For each species, however, the many chorion proteins must assemble into one rigid, semipermeable macromolecular structure, suggesting that this rapid evolution must be a coordinated process.

The two gene families expressed in *B. mori* during the late period of choriogenesis, *HcA* and *HcB*, are arranged in 15 divergently transcribed pairs containing one member of each family, clustered in a 140 kilobase pair (kb) region of chromosome 2 (EICKBUSH and KAFATOS 1982; EICKBUSH and BURKE 1985; 1986; BURKE and EICKBUSH 1986). Nucleotide sequence analysis of these genes indicated that a high degree of sequence identity in both their coding and noncoding regions exists within each family. Numerous sequence transfers resembling gene conversion events were detected between the gene pairs. A gradient of transfers was observed in which recombination appeared to initiate within the genes, in a region encoding a tandem array of cysteine-glycine-glycine amino acid repeats. Further support for these gene conversion-like events was obtained by comparison of the nucleotide sequences of the same gene pair and its flanking

[1] Present address: Cancer Center, University of Rochester Medical School, Rochester, New York 14627.

regions in two races of *B. mori* (XIONG, SAKAGUCHI and EICKBUSH 1988). Nucleotide differences between the two strains were more prevalent in the gene regions than the 3′ flanking regions; they were clustered in short conversion-like patches, and in most cases corresponded to nucleotide variants found in other members of the *HcA* and *HcB* families. The 3′ flanking regions were identical suggesting that these sequence transfers were not the result of unequal crossovers.

A short chromosomal segment containing early chorion genes has previously been cloned (HIBNER *et al.* 1988). Two of the genes on this segment were found to be arranged as a divergently oriented gene pair, *ErA/ErB*. In this paper we present the cloning and sequence analysis of all members of the *ErA* and *ErB* multigene families. The major exons of the five *ErA* genes contain high levels of nucleotide sequence identity, while the *ErB* major exons do not. We present evidence that sequence exchanges similar to gene conversion are again responsible for the high level of sequence identity in a chorion gene family. Unlike the *HcA* and *HcB* genes, however, these events in the *ErA/ErB* gene pairs are localized to only the major exon of the *ErA* genes.

MATERIALS AND METHODS

Genomic libraries and their screening: The cDNA inserts from m6C11, m6A2 and m2G12 (EICKBUSH *et al.* 1985) were used to screen 150,000 clones of a Charon 4 partial *Eco*RI library (EICKBUSH and KAFATOS 1982). Hybridizations were conducted at 75° in 2 × SSC (1 × SSC = 0.15 M NaCl, 0.015 M Na citrate), 0.1% bovine serum albumin, 0.1% Ficoll, 0.1% polyvinylpyrrolidone, 25 mM Na phosphate, pH 6.5, 1% sodium pyrophosphate, 0.1% SDS, 10% dextran sulfate, and 250 mg/ml denatured calf thymus DNA. The final wash of filters was in 0.1 × SSC, 0.1% SDS at 75°. Phage DNAs from positive plaques were purified, restriction digested with *Eco*RI, and clones sharing an *Eco*RI fragments were organized into arrays. Additional restriction enzymes were used to confirm the overlap between clones. In this manner all positive phage clones from the *Eco*RI library were placed into three arrays, one of which contained the two previously identified clones E1 and E2 (HIBNER *et al.* 1988). In an effort to link the three arrays of phage clones a partial *Sau*3A library was constructed. High molecular weight DNA was isolated from 10 sibling female moths of strain 703 and was partially digested with *Sau*3A. Fragments of DNA 15–30 kb in length were isolated on sucrose gradients and inserted into the vector Charon 35 (LOENEN and BLATTNER 1983). Phage DNA was packaged using the extracts and procedures of Promega Biotec. Screening of this library and the isolation and characterization of positive clones was similar to that of the *Eco*RI library.

Analysis of early chorion genes: Restriction fragments from the overlapping clones that hybridized to the cDNA clones m6A2 or m6C11 were subcloned into pUC13. Detailed restriction maps of the subclones were generated, and specific restriction fragments containing gene regions were placed into m13mp18 and m13mp19 vectors (YANISCH-PERRON, VIEIRA and MESSING 1985). The nucleotide sequence of these fragments was determined by the dideoxy-chain termination method (SANGER, NICKLEN and COULSON 1977). All protein encoding regions were sequenced on both strands. DNA sequences and deduced protein sequences were compiled and analyzed using the MacVector Analysis Software available from International Biotechnologies Inc. Regions of nucleotide similarity were originally localized using the matrix analysis programs available in the MacVector Analysis Software (Biotechnologies, Inc.). In the case of the *ErB* coding sequences, optimum alignments were aided by following the protein alignments, and involved the introduction of numerous gaps to increase the maximum identity in the amino-terminal and carboxyl-terminal regions of the proteins.

RESULTS

Cloning of the *ErA/ErB* gene families: The previously characterized *ErA.1* gene (HIBNER *et al.* 1988) had 96% nucleotide sequence identity to the cDNA clone m6C11 (LECANIDOU *et al.* 1986). *ErB.1* had approximately 60% nucleotide sequence identity to the cDNA clone m6A2 (LECANIDOU *et al.* 1983). To estimate the size of the *ErA* and *ErB* families, genomic blots were performed using m6C11 to probe for the *ErA* family, and both *ErB.1* and m6A2 to probe for the *ErB* family. The *ErB.1* probe corresponded to the two adjacent *Kpn*I fragments from the *ErB.1* major exon (see Figure 2). At the criteria we have previously used to determine the members of a chorion gene family, 75°, 0.3 M NaCl (EICKBUSH and KAFATOS 1982; EICKBUSH *et al.* 1985) the m6C11 probe hybridized to as many as five genomic bands (data not shown). In the case of the *ErB* gene probes, two genomic bands hybridized to m6A2, while only one genomic band, corresponding to the *ErB.1* gene itself, hybridized to the *ErB.1* probe. At low hybridization criteria (65°, 0.6 M NaCl) both the m6A2 and *ErB.1* probes hybridized to a large number of additional bands. These additional bands revealed at low hybridization criteria represented members of the middle *B* families (see also Figure 2C in LECANIDOU *et al.* 1983). The m6C11 probe did not hybridize to additional genomic bands at low criteria.

To clone the remaining genes of the *ErA* families, m6C11 was used to screen a Charon 4 partial *Eco*RI library (EICKBUSH and KAFATOS 1982) and a Charon 35 partial *Sau*3A library (see MATERIALS AND METHODS). To clone other possible early chorion genes, clones m6A2 and m2G12 were also used to probe these libraries. These two cDNA clones represented the only chorion probes which had early developmental kinetics (EICKBUSH *et al.* 1985) for which the corresponding genes were not yet isolated.

All positive genomic clones obtained from the two libraries using these three probes were assembled into two overlapping arrays as shown in Figure 1. One array was 77 kb in length and contained four *ErA* genes, the two genes which hybridized to m6A2 on genomic blots (labeled *ErB.3* and *ErB.4* in the figure, see below) as well as two genes hybridizing to m2G12.

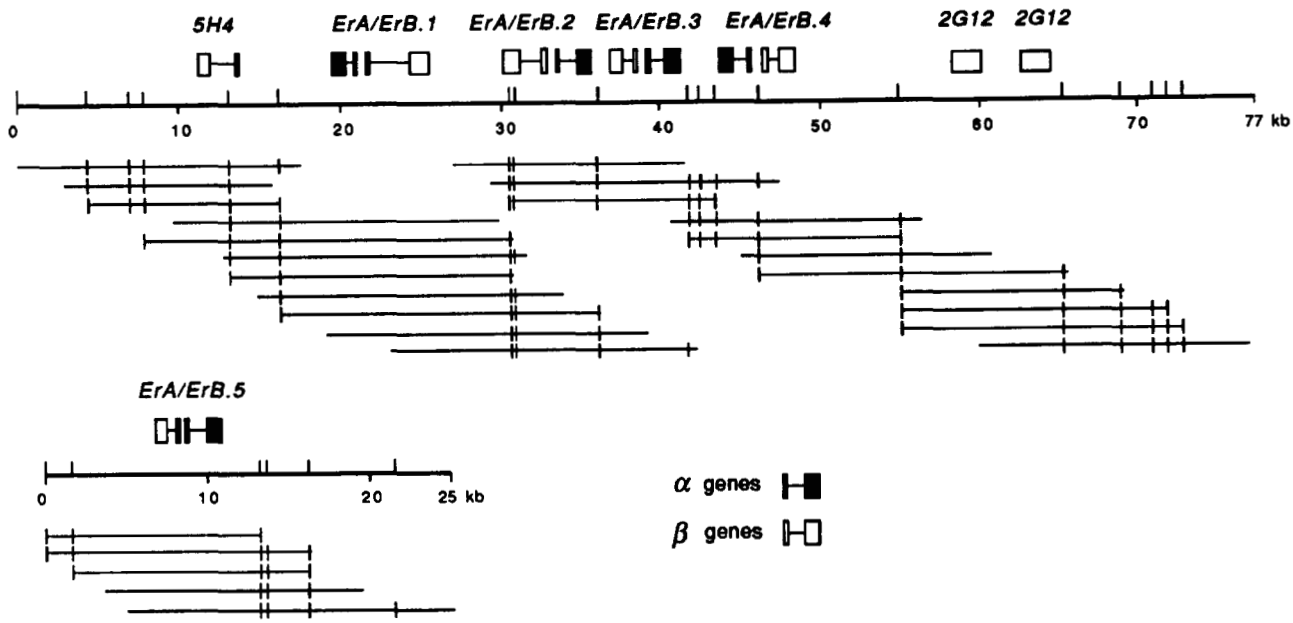

FIGURE 1.—Two cloned segments of the early chorion gene complex of *B. mori*. Genomic *Eco*RI sites are indicated by the short vertical lines. Representative overlapping phage clones, shown below the *Eco*RI map, were placed into two arrays based upon their restriction digestion pattern with *Eco*R1. Clones ending at *Eco*RI sites were isolated from a partial *Eco*RI lambda Charon 4 genomic library; clones not ending at *Eco*RI sites were isolated from a partial *Sau*3A lambda Charon 35 library. The positions of the chorion genes are indicated with boxes above the *Eco*RI map. Each gene contains two exons (boxes) separated by an intron (horizontal line) as determined by nucleotide sequence analysis. Genes homologous to the cDNA clone, m6C11, were named the *ErA* genes and are shown as solid boxes; genes that are divergently paired with these *ErA* genes, the *ErB* genes, are shown as open boxes. The 25-kb fragment containing *ErA/ErB.5* is located less than 60 kb from the left end of the 77-kb fragment. The orientation of the 25-kb segment to the 77-kb segment is not known. Also shown are the approximate locations of two genes homologous to the cDNA clone, m2G12.

One end of this segment contained the previously characterized *5H4*, *ErA.1* and *ErB.1* genes (HIBNER *et al.* 1988). The second region cloned was 24.5 kb in length and contained one *ErA* gene. These genes accounted for all of the *ErA.1* hybridizing bands detected on the genomic blots. As described below, each *ErA* gene is paired with another chorion gene. While the exact distance of the 24.5 kb fragment from the 77-kb fragment is not known, genomic blot analysis using pulse field gels has revealed that all five *ErA* genes are located on a single 145-kb *Not*I fragment (J. IZZO, unpublished data).

Organization and expression of the early chorion gene pairs: The approximate location of all chorion genes on the two genomic fragments was determined by low criteria hybridization to all characterized early cDNA clones (EICKBUSH *et al.* 1985) as well as total cDNA made by reverse transcription of early chorionogenic mRNA. The complete nucleotide sequence of each gene region was then determined (see MATERIAL AND METHODS). The newly cloned *ErA* genes on the 77-kb segment were named *ErA.2*, *ErA.3* and *ErA.4*, while the gene on the 24.5-kb segment was named *ErA.5*. Each of the *ErA* genes was found to be paired with a divergently transcribed gene that had low nucleotide similarity to *ErB.1*. These *ErB*-like genes were named *ErB.2* through *ErB.5*. The gene sequences are available from GenBank under the following accession numbers: *ErA.2*,X58445; *ErA.3*, X58446; *ErA.4*, X58447; *ErA.5*, X58448; *ErB.2*, X58449; *ErB.3*, X58450; *ErB.4*, X58451; and *ErB.5*, X58452. Analysis of the *2G12* gene sequences will be the subject of a separate report (J. IZZO and EICKBUSH, in preparation).

Comparison of the m6C11 and m6A2 cDNA clones with the various genomic sequences indicated that these cDNA clones were derived from transcripts of the *ErA.4* and *ErB.4* genes respectively. Consequently, a precise determination of the exon/intron structure of this gene pair was possible. The intron/exon structure for the remaining gene pairs was determined by their nucleotide similarities to the *ErA/ErB.4* pair. Each early gene consisted of two exons, the first small exon containing the 5′ untranslated region and all but four amino acids of the leader peptide, and the second, larger exon encoding the remaining four amino acids of the leader peptide, the entire mature protein, and the 3′ untranslated region. Diagrams of the five *ErA/ErB* gene pairs with detailed restriction maps are shown in Figure 2. As is the case with the *HcA/HcB* and *A/B* chorion gene pairs (EICKBUSH and BURKE 1986; SPOEREL *et al.* 1989) only a few hundred base pairs separate the 5′ ends of these divergently transcribed *ErA* and *ErB* genes. The lengths of the *ErA* and *ErB* introns are variable ranging from 375 to 1214 bp in the *ErA* genes and 470 to 1100 bp in

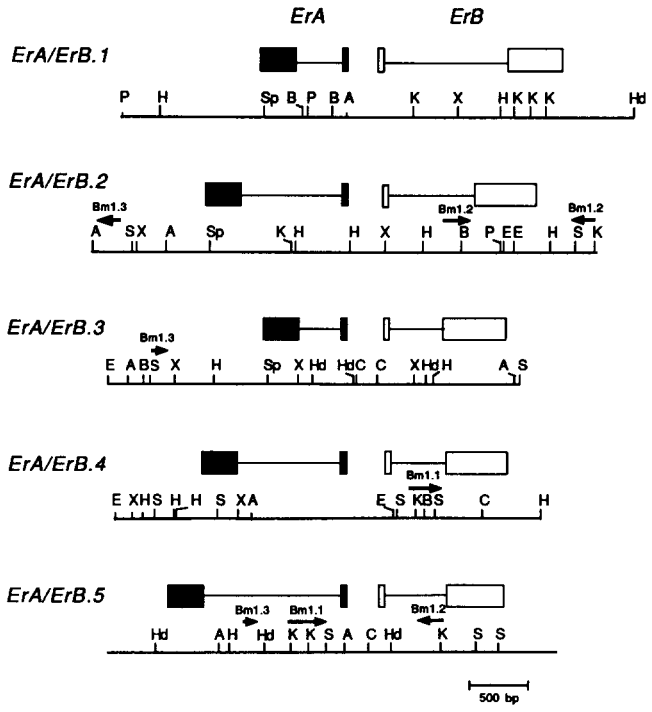

FIGURE 2.—Detailed diagrams of the *ErA/ErB* gene pairs. The extent of the chromosomal region sequenced from each gene pair is indicated by the horizontal line. Above each restriction map is the location of the protein encoding regions of the minor and major exons. In order that the corresponding regions of each gene pair can be directly compared, the orientation of gene pairs *ErA/ErB.2*, *ErA/ErB.3* and *ErA/ErB.5* is drawn opposite to that in Figure 1. Arrows correspond to middle repetitive Bm1 elements identified in *B. mori* (ADAMS *et al.* 1986). Elements labeled 1.1, 1.2 and 1.3 are 450, 250 and 125 bp, respectively, in length. The arrows point in the direction of the oligo-A tail at the 3′ end of each element. Restriction sites: A, *Ava*I; B, *Bgl*II; C, *Cla*I; E, *Eco*R1; H, *Hin*cII; Hd, *Hin*dII; K, *Kpn*I; P, *Pst*I; S, *Sst*I; Sp, *Sph*I; X, *Xba*I.

the *ErB* genes. Located within certain of these introns as well as 3′ of the genes are copies of the middle repetitive oligo-A terminated element Bm1 (ADAMS *et al.* 1986). Bm1 elements can be divided into three size classes of 450, 250 and 125 bp, labeled Bm1.1, Bm1.2 and Bm1.3, respectively. The variable location of these elements and their high levels of nucleotide identity indicate that they have inserted recently, well after the gene pair duplications which gave rise to the *ErA* and *ErB* families.

All *ErA* and *ErB* genes appear functional in that they encode appropriate leader peptides for secretion from the cell, appropriate splice-sites for RNA processing, and no premature termination codons. All early genes appear to have the same temporal transcription pattern with mRNA initially accumulating in follicle 1 and disappearing by follicle 5–8 (EICKBUSH *et al.* 1985). In the case of the five *ErA* genes, transcripts from all members of the family were detected using the m6C11 cDNA probe. In the case of the five *ErB* genes, transcripts were detected using the previously described probes *ErB.1* and m6A2 (HIBNER *et al.* 1988) as well as probes specific to *ErB.2* and *ErB.5* (data not shown).

Nucleotide sequence identities within the *ErA* and *ErB* families: Except for the Bm1 elements and short regions immediately adjacent to the *ErA* major exons (see below), no nucleotide sequence identity could be detected between the noncoding regions of the different gene pairs. This differs from the middle and late chorion genes where nucleotide similarities extend throughout both the coding and noncoding regions of the gene pairs (see BURKE and EICKBUSH 1986; SPOEREL *et al.* 1989). The lack of sequence identity in the noncoding regions of the early gene pair is most surprising for the 5′ flanking regions. *P* element-mediated transformation experiments in Drosophila have shown that the short intergenic region of the *A/B* gene pairs of *B. mori* contains elements sufficient for its correct temporal and tissue specific regulation (MITSALIS and KAFATOS 1985). In particular the hexanucleotide sequence TCACGT has been proposed to be an orientation-independent, tissue-specific regulatory element that is evolutionarily conserved in both lepidopterans and dipterans (MITSALIS and KAFATOS 1985; MITSALIS *et al.* 1987; KONSOLAKI *et al.* 1990). Within the 5′ flanking regions of the *ErA/ErB* gene pairs, either orientation of this hexanucleotide sequence was found only in one instance (48 bp 5′ of the *ErB.4* TATA box). While multiple 5 of 6 matches to this hexanucleotide sequence can be found in either orientation within the 5′ flanking region of all *ErA/ErB* gene pairs (see for example HIBNER *et al.* 1988), these matches did not occur at a uniform position within the 5′ flanking DNA, and are not more frequent than expected for random DNA sequences. We have not been able to identify conserved sequences specific to the *ErA/ErB* gene pairs; no pentanucleotide or longer sequence was found at a consistent position within the 5′ flanking region of all early gene pairs.

The nucleotide sequences of all regions of the *ErA/ErB* gene pairs with identifiable nucleotide similarity are shown in Figure 3 (*ErA* genes) and Figure 4 (*ErB* genes). Highest nucleotide sequence identity between the early genes (95–97%) was found for the coding regions of the *ErA* major exons. Nucleotide identity at a reduced level extended for approximately 50 bp 3′ of the termination codon of the *ErA* genes. Based upon its cDNA sequence the poly-A addition site for *ErA.4* is located 75 bp downstream of the termination codon (LECANIDOU *et al.* 1986). Thus this region of 3′ similarity in the *ErA* genes does not include the entire 3′ untranslated regions of the genes. For two genes (*ErA.4* and *ErA.5*) a 214 bp segment with high sequence identity was also found extending 5′ of the *ErA* major exon into the intron. The possible significance of this region will be described in greater detail below.

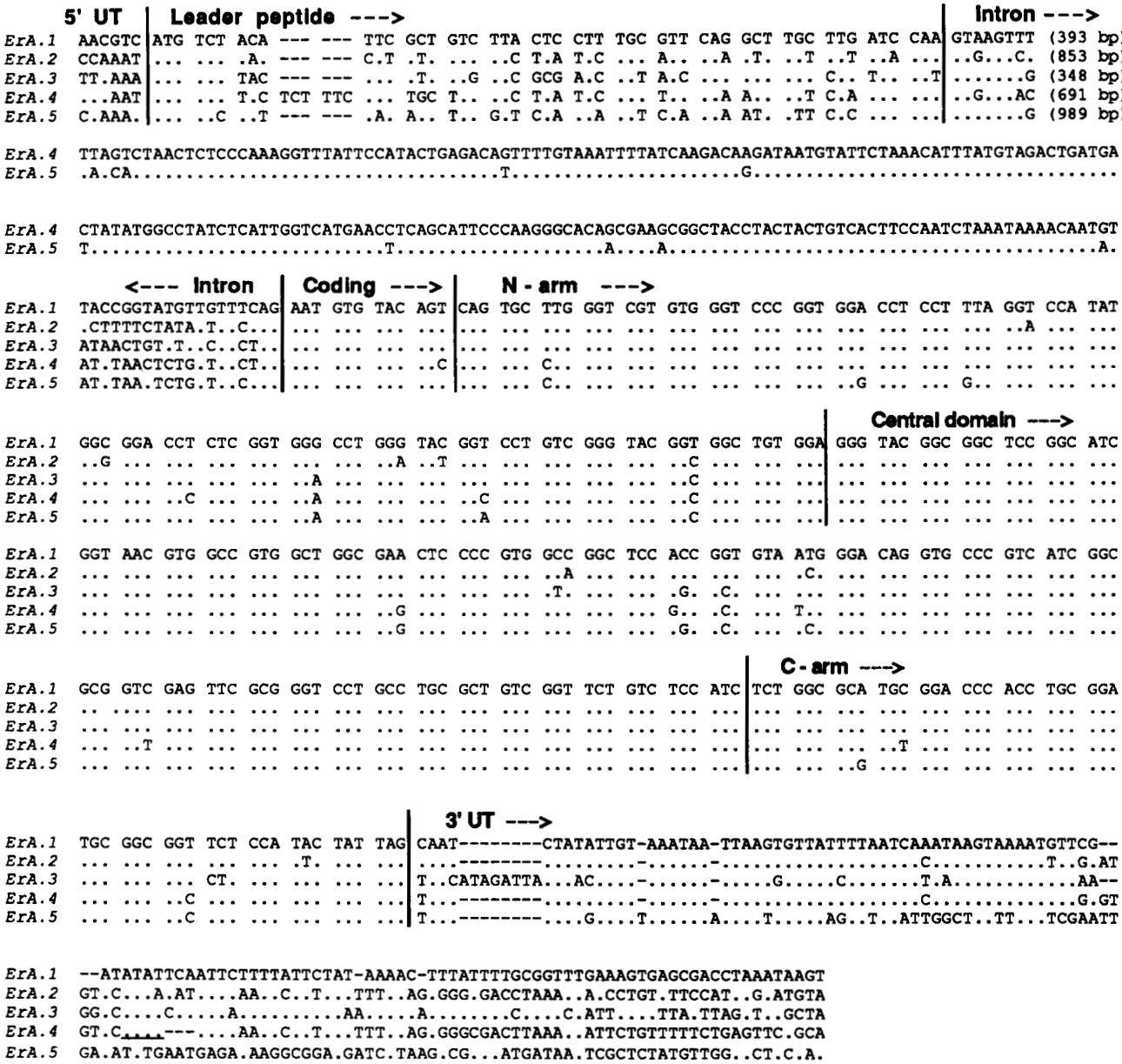

FIGURE 3.—Nucleotide sequences of the regions of sequence similarity within the *ErA* genes. All sequences are compared to *ErA.1*, with identical nucleotides shown as dots, nucleotide substitutions indicated, and gaps necessary for alignment by dashes. All codons of the minor and major exons are shown as triplets. An additional sequence of the *ErA* intron which is similar in the *ErA.4* and *ErA.5* genes is also shown. The number of nucleotides from each intron which are omitted from the comparison are indicated for each gene. The length of the 3′ untranslated region is only known for the *ErA.4* gene (LECANIDOU *et al.* 1986) and is indicated in the figure by underlining the last 4 bases of its 3′ untranslated region.

All other protein encoding regions of the early genes contained significantly lower levels of sequence identity than that of the *ErA* major exon. Nucleotide identities of 62% (*ErA* genes) and 55% (*ErB* genes) were detected for the leader peptide encoding regions of the minor exons, and 63% nucleotide identity was found for the protein encoding regions of the *ErB* major exons. Finally, no regions of sequence similarities were found which extended downstream of the termination codon of the *ErB* genes or 5′ of the major exon into the introns.

Selective pressure on the encoded proteins cannot explain the sequence identity within the *ErA* major exon: One possible explanation that could account for the high levels of sequence identity between the *ErA* major exons compared to the *ErB* major exons, is a higher selective pressure to maintain the sequence of the *ErA* proteins. To examine this possibility we have calculated the ratio of synonymous nucleotide substitutions to replacement substitutions for all pairwise combinations of *ErA* (Table 1) and *ErB* genes (Table 2). The ratio of synonymous to replacement substitutions for the *ErA* genes average 1.93. The large variation in this ratio (0.4–3.0) for the individual

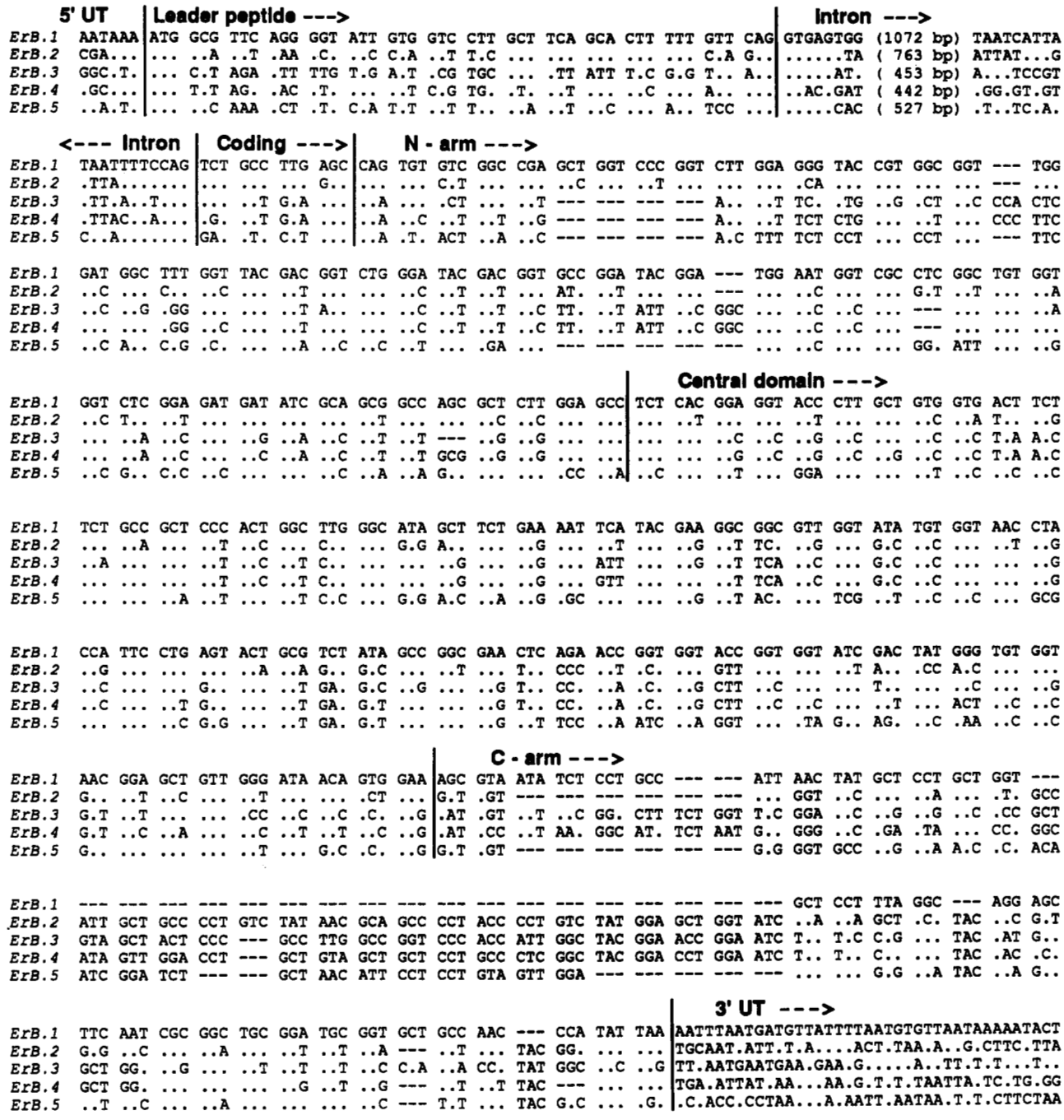

FIGURE 4.—Nucleotide sequence comparison of the regions of sequence similarity within the *ErB* genes. All sequences are compared to *ErB.1* with identical nucleotides shown as dots, nucleotide substitutions indicated, and gaps necessary for alignment by dashes. All codons of the minor and major exons are shown as triplets. The number of nucleotides of the intron omitted from the comparison is indicated for each gene.

ErA gene comparisons is due to the low number of total nucleotide substitutions present in these genes (range 7–19). The ratio of synonymous to replacement substitutions for the *ErB* genes average 1.46, somewhat lower than that of the *ErA* genes. However, this ratio is likely to be an underestimate of the true ratio for the *ErB* genes, because the total number of synonymous substitutions in the *ErB* genes is underestimated (see below).

As a second approach to determine if the observed difference in sequence conservation between the *ErA* and *ErB* genes was a result of selective pressure, we calculated the percentage of nucleotide substitutions at fourfold synonymous sites. Nucleotide changes at fourfold synonymous positions are not subject to direct selective pressure on the encoded proteins. The average percent divergence at fourfold synonymous positions in the *ErA* genes was 7.7% (Table 1), and 57.8% in the *ErB* genes (Table 2). If one corrects for multiple substitutions at the same site (JUKES and

TABLE 1

The ratio of synonymous to replacement substitutions for the *ErA* genes (lower half of the matrix), and the percentage nucleotide divergence at fourfold synonymous sites (upper half of the matrix)

	ErA.1	*ErA.2*	*ErA.3*	*ErA.4*	*ErA.5*
ErA.1	—	6.5	2.6	9.1	7.9
ErA.2	3.0	—	6.6	13.0	11.8
ErA.3	0.4	0.9	—	6.6	5.3
ErA.4	3.3	2.8	2.0	—	7.9
ErA.5	2.0	3.0	1.2	1.4	—

TABLE 2

The ratio of synonymous to replacement substitutions for the *ErB* genes (lower half of the matrix), and the percentage nucleotide divergence at fourfold synonymous sites (upper half of the matrix)

	ErB.1	*ErB.2*	*ErB.3*	*ErB.4*	*ErB.5*
ErB.1	—	44.5	66.7	62.3	52.4
ErB.2	1.7	—	61.5	60.7	63.4
ErB.3	1.7	1.7	—	33.3	68.7
ErB.4	1.7	1.5	1.7	—	64.1
ErB.5	1.2	1.2	1.1	1.1	—

CANTER 1969) these values become 8.2 and 110%, respectively. Thus the *ErB* genes have 13 times the level of divergence at fourfold synonymous sites found in the *ErA* genes.

One possible mechanism of selecting for particular nucleotides at fourfold synonymous sites is tRNA abundance in the tissue of expression. The *ErA* and *ErB* gene families encode proteins with approximately the same amino acid composition, which are synthesized at the same time in the same tissue at roughly the same level. One would expect their codon preference, therefore, to be similar. A comparison of the codon usage for the five most abundant amino acids in the *ErA* and *ErB* genes are shown in Table 3. These five amino acids account for 69% of all codons in the *ErA* genes. Codon usage for the remaining amino acids are not included in this table since they are present at too low a level to evaluate codon preferences. The *ErA* and *ErB* genes have similar codon preferences, with one exception, the *ErA* genes utilize more frequently the valine codon GTG, while *ErB* genes utilize more frequently GTT. This difference affects on average only four codons per gene, thus codon bias can not explain the markedly higher conservation of nucleotides at fourfold synonymous positions in the *ErA* genes.

Evidence for sequence transfers between members of the *ErA* and *ErB* families: Sequence transfers between the Hc genes have been postulated to explain the maintenance of a high degree of sequence identity within the two late chorion gene families (EICKBUSH and BURKE 1985; 1986; XIONG, SAKAGUCHI and Eickbush 1988). Evidence for such sequence exchanges between the *ErA* genes can also be found in several instances where these exchange events appear to have extended beyond the coding region of the major exon. In the intron, 5′ of the major exon, nucleotide similarity was detected between the *ErA* genes in only one instance: the *ErA.4* and *ErA.5* genes contained 96% identity for a 214-bp region (Figure 3). In the region 3′ of the *ErA* major exon coding regions, approximately 85% sequence identity was found for the first 50 bp downstream of the termination codons. Beyond this region sequence identity between the *ErA* genes disappears rapidly except for two pairwise comparisons: 78% for a 47-bp region between *ErA.1* and *ErA.3*, and 88% for a 59-bp region between *ErA.2* and *ErA.4*. Thus evidence for three different sequence exchanges can be found in the *ErA* genes: one between *ErA.4* and *ErA.5* that extended into the intron, a second between *ErA.1* and *ErA.3* extending into the 3′ flanking region, and a third between *ErA.2* and *ErA.4* also extending into the 3′ flanking region.

TABLE 3

Codon usage comparison of the *ErA* and *ErB* genes for the five most abundant amino acids

Amino acid	Codon	*ErA* genes	*ErB* genes
Glycine	GGT	0.32	0.42
	GGC	0.35	0.34
	GGA	0.21	0.18
	GGG	0.12	0.06
Valine	GTT	0.02	0.33
	GTC	0.40	0.39
	GTA	0.10	0.13
	GTG	0.48	0.15
Proline	CCT	0.48	0.49
	CCC	0.35	0.22
	CCA	0.17	0.22
	CCG	0.00	0.07
Alanine	GCT	0.30	0.47
	GCC	0.33	0.30
	GCA	0.12	0.11
	GCG	0.26	0.12
Tyrosine	TAT	0.32	0.31
	TAC	0.68	0.69

In the case of the *ErB* genes no evidence was found for recent sequence exchange events. No sequence similarities outside the protein encoding regions were detected for any pairwise comparisons of *ErB* genes. For the protein encoding regions themselves, the highest level of sequence identity was always between *ErB.3* and *ErB.4*, suggesting that they represent the last gene duplication event within the family. The lowest level of identity was between *ErB.5* and all other *ErB* genes.

DISCUSSION

With the cloning of 102 kb of chromosomal DNA containing 13 early chorion genes the previously iden-

tified early cDNA clones (EICKBUSH *et al.* 1985) have now been analyzed at the gene level. We assume this region corresponds to a major segment of the Ch3 locus defined by GOLDSMITH and CLERMONT-RATTNER (1979, 1980). Our attempts to isolate additional early chorion cDNA families by probing a follicle cell cDNA library at low criteria with total synthesized cDNA from early follicles have not resulted in any cDNA clones that do not hybridize with already characterized cDNAs at low criteria (B. L. HIBNER and T. H. EICKBUSH, unpublished data). We conclude that if other early chorion genes exist, they represent minor transcripts that have little sequence identity with those previously characterized.

Sequence exchange between the *ErA* genes: Analysis of the nucleotide sequences of the *ErA/ErB* gene pairs suggests that the significantly higher level of sequence identity of the *ErA* major exon (96%) compared to that of the *ErB* major exon (63%) is not a result of higher selective pressure on the encoded *ErA* proteins, since the high sequence identity of the *ErA* genes include fourfold synonymous positions. Two major categories of recombination mechanisms are known that can maintain high levels of sequence identity between the members of a multigene family. Unequal crossing over, in expanding and contracting the number of genes in a tandem array, can lead to the eventual homogenization of those genes (SMITH 1973). A number of multigene families are believed to be undergoing unequal crossing over including the rDNA genes (PETES 1980; SZOSTAK and WU 1980; COEN, STRACHAN and DOVER 1982; SEPERACK, SLATKIN and ARNHEIM 1988), visual pigment genes (VOLLRATH, NATHANS and DAVIS 1988), amylase genes (GUMUCIO *et al.* 1988) and proline-rich protein genes (LYONS, STEIN and SMITHIES 1988). The absence of significant nucleotide sequence similarity outside the coding regions suggests however, that the *ErA/ErB* gene pairs have not undergone recent unequal crossover events.

The second recombination mechanism that can increase sequence identity between members of a multigene family is gene conversion. Gene conversion-like events have been implicated in the evolution of a growing number of multigene families (RUPPERT, SCHERER and SCHUTZ 1984; POWERS and SMITHIES 1986; ATCHISON and ADESNIK 1986; REYNAUD *et al.* 1987; CRAIN *et al.* 1987; LE BLANCQ *et al.* 1988; GELIEBTER and NATHENSON 1988; PARHAM *et al.* 1988). Unlike the gene conversions found in fungi (JUDD and PETES 1988; BORTS and HABER 1989), the conversion-like events in higher eukaryotic multigene families are usually less than a few hundred base pairs in length and are typically not associated with reciprocal exchanges (crossovers) (POWERS and SMITHIES 1986; REYNAUD *et al.* 1987; PARHAM *et al.* 1988; GELIEBTER and NATHENSON 1988). By integrating murine major histocompatibility genes into yeast chromosomes it has recently been confirmed that the short conversion-like tracts detected in these genes in mice are authentic gene conversion events (WHEELER *et al.* 1990).

The levels of sequence identity in and around the *ErA* genes are clearly best explained by gene conversions occurring between the five members of the family. In most cases these events are localized to the major exons of the genes, however in at least three instances these conversions have extended outside these exons: high levels of sequence identity were found between the *ErA.4* and *ErA.5* genes extending into the intron, and between the *ErA.1* and *ErA.3* genes and *ErA.2* and *ErA.4* extending into the 3′ untranslated region. It should be noted that *ErA.5* is located 30–80 kb distant from the 25-kb segment containing the other four *ErA* genes. Since the level of sequence identity of the *ErA.5* major exon is similar to that of the other four *ErA* genes, the sequence transfers between the *ErA* genes appear to be largely independent of their distance along the chromosome.

Why *ErB* genes are not undergoing sequence exchanges: We have previously suggested that gene conversion-like events are responsible for the high levels of sequence identities in the late (*HcA* and *HcB*) and middle (*A* and *B*) chorion gene families (EICKBUSH and BURKE 1985; 1986; XIONG, SAKAGUCHI and EICKBUSH 1988; SPOEREL *et al.* 1989). Therefore, perhaps the more surprising result of our analysis of the early chorion gene families is not that the *ErA* family was undergoing gene conversions, rather that the *ErB* family was *not* undergoing these events. Analysis of the *ErB* gene sequences suggests that these genes are diverging independently of each other. The *ErB.5* gene appears to be the oldest member of the family since it uniformly has the lowest levels of homology to the other genes, while *ErB.3* and *ErB.4* have the highest levels of homology suggesting that they resulted from the most recent duplication event to occur in this family. We can find evidence for only one sequence transfer between the *ErB* genes since their expansion into a gene family. The region of the *ErB.1* and *ErB.2* genes that encode the N-arm of the mature protein (see Figure 4) contains 84% nucleotide identity, while the regions encoding the central domain and C-arm of the proteins contain, respectively, 75% and 59% nucleotide identity. In all other pairwise comparisons of the *ErB* genes, the region encoding the N-arm exhibit a level of identity intermediate between that of the central domain and the C-arm (data not shown). Thus unless one assumes a higher level of selective pressure on the N-arms of just these two *ErB* proteins, a relatively old sequence transfer has occurred between the *ErB.1* and *ErB.2* genes.

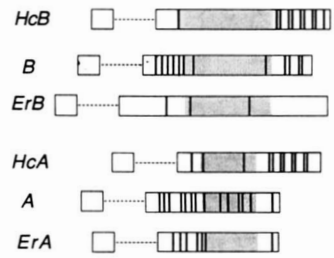

FIGURE 5.—The location of the nucleotide sequence GGXGGX in the six major chorion gene families (X = A, T or C). The protein encoding regions of the exons are shown as boxes, the introns as horizontal lines. Only the position of the introns are indicated with no attempt made to draw their length to scale. The length of the protein encoding regions correspond to the average length for that gene family. The region of the major exon encoding the central domain of the encoded protein is shaded. The location of GGXGGX sequences are indicated by the dark vertical bars.

One explanation for why the *ErB* genes have only rarely undergone gene conversions, is based on the finding that the *ErB* genes do not contain the putative recombination hotspots previously identified in the *Hc* chorion gene families which do undergo these events. The sequence transfers in the late chorion genes were not uniformly distributed along the *HcA/HcB* gene pairs (EICKBUSH and BURKE 1986). Sequence transfers were highest near the 3′ end of each gene and lowest in the common 5′ region between the divergently transcribed genes. A model was presented that explained these gradients by assuming that the events leading to gene conversion preferentially initiated in a short DNA repeat. The resulting heteroduplexes could then extend to distances influenced by features of the sequence affecting their stability. The DNA repeat, TG(T/C)GGXGGX (where X = A, T or C) encoded tandem copies of the amino acid sequence cysteine-glycine-glycine. As diagramed in Figure 5, each *Hc* gene has on average 8 of these repeats clustered in the region encoding the C-arm of the *Hc* protein.

Gene conversions have also been implicated in maintaining the high levels of sequence identities in the *A* and *B* chorion gene families (SPOEREL *et al.* 1989). We have searched the *A* and *B* chorion genes for sequences corresponding to the putative late chorion gene recombination hotspot. While only a few copies of the full repeat were found, these genes contained a large number of examples of a portion of this sequence, GGXGGX, as shown in Figure 5. The *A* genes have on average seven copies of this repeat clustered in the region encoding the N-arm, and another six copies in the central domain. The *B* genes have six clustered copies in the N-arm encoding region and another four copies in the C-arm encoding region. In all cases the GGXGGX sequences in the *A* and *B* genes correspond to paired glycine codons. Those repeats within the region encoding the N- and C-arms of the proteins encode part of the larger amino acid sequence, glycine-(tyrosine or leucine)-glycine-glycine, that are the major component of the N- and C-arms of all *A* and *B* proteins, instead of the cysteine-glycine-glycine repeats present in the *Hc* proteins.

We have also searched the *ErA* and *ErB* genes for GGXGGX sequences. Six copies of this sequence are clustered in the region encoding the N-arm of the *ErA* protein and one copy in the region encoding the C-arm. In all cases these repeats encoded paired glycine residues, usually as part of a tyrosine-glycine-glycine or cysteine-glycine-glycine repeat like that in the *Hc*, *A* and *B* chorion proteins. The *ErB* genes, on the other hand, contain only three copies of the GGXGGX nucleotide sequence. These copies are widely separated in the *ErB* genes, one in the region encoding the N-arm and at either end of the region encoding the central domain. From one to four additional copies of GGXGGX sequences can also be found in the *ErB* genes, however these copies are not conserved between the different genes.

Thus all five of the major chorion gene families which appear to be undergoing gene conversions contain closely spaced copies of the simple nucleotide sequence, GGXGGX. This simple nucleotide sequence encodes paired glycine residues, a major repeat present in the N- and C-arms of most chorion proteins. This correlation supports our original hypothesis based solely on the gradients of sequence transfers within the *Hc* genes (EICKBUSH and BURKE 1986). Because the only gene family that does not undergo frequent gene conversions, *ErB*, contains a few poorly conserved copies of the GGXGGX repeat, it appears that either the density of these repeats or their precise spacing must be important for the efficient promotion of exchange events.

Several other eukaryotic recombination hotspots have been previously correlated with reciprocal or nonreciprocal recombination events; a region next to the *Eβ* gene in the mouse histocompatibility locus (KOBORI *et al.* 1986; STEINMETZ, STEPHAN and LINDAHL 1986); deletion mutants in the *ADE8* gene of *Saccharomyces cerevisiae* (WHITE *et al.* 1988), and within the phosphoglycerate kinase gene of *Trypanosoma brucei* (LE BLANCQ *et al.* 1988). The sequences of these recombination hotspots are compared in Figure 6 with those from the chorion genes, as well as with the prokaryotic recombination signal, *chi* (SMITH *et al.* 1981). These putative hotspots have in common the occurrence of paired guanine nucleotides, separated by either 1 or 2 bp. In the case of the phosphoglycerate kinase gene the hotspot occurs within the protein encoding region corresponding to three consecutive glycine codons. Recently, a high level of sequence identity between two collagen genes in *Caenorhabditis elegans* has been attributed to gene conversion (PARK and KRAMER 1990). Each of these genes

Chi (bacterial)	G C T GG T GG
HcA/HcB chorion genes	T G Y GG X GG X
A/B chorion genes	T A Y GG X GG X
ErA chorion genes	GG X GG X
Phosphoglycerate kinase genes	GG T GG T GG T
ADE8 gene	G C Y X GGG C R GG X T
MHC - *Eβ*	GG A GG T A GG (CAGG)17
Human minisatellite DNA (core)	GGG C A GG A X G
Human VNTR markers (core)	GGG N N G T GGGG

FIGURE 6.—Recombination hotspots identified within or flanking eukaryotic genes. All sequences can be characterized as having paired G residues separated by one or two nucleotides. Paired G residues are also found in the chi sequence of *E. coli*. In the case of the chorion genes and the phosphoglycerate kinase genes the hotspots are within the protein encoding regions, with the paired G separated by one base corresponding to glycine codons. *N* = any nucleotide; *X* = A, T or C; *Y* = pyrimidine; *R* = purine. Sequences are from: *chi* DNA (SMITH *et al.* 1981); phosphoglycerate kinase gene (LE BLANCQ *et al.* 1988); *ADE8* gene (WHITE *et al.* 1988); *Eβ* gene (KOBORI *et al.* 1986; STEINMETZ, STEPHAN and LINDAHL 1986); core sequence of minisatellite DNA (JEFFERYS, WILSON and THEIN 1985); and the core common sequence found in a series of VNTR markers (NAKAMURA *et al.* 1987).

contains within a 600-bp region, five examples of the GG*X*GG*X* sequence, and six examples of its inverse sequence, CC*X*CC*X* (where *X* = A, T or G). The former sequences encode paired glycine residues in the collagen protein while the latter encodes paired proline residues. Similarly many examples of CC*X*CC*X* encoding paired proline residues and GG*X*GG*X* encoding paired glycine residues can also be found in the human salivary proline-rich protein genes, which undergo frequent intragenic unequal crossover events (LYONS, STEIN and SMITHIES 1988). Finally, G-rich DNA is associated with the core sequences of the hypervariable human minisatellite DNA (JEFFREYS, WILSON and THEIN 1985), and human variable number of tandem repeat (VNTR) markers used for gene mapping (NAKAMURA *et al.* 1987). However, it has recently been suggested that it is the repetitious nature of these latter sequences causing slippage mutations, rather than their *chi* similarity stimulating recombination, that accounts for their hypervariability (JEFFREYS, NEUMANN and WILSON 1990; DOVER 1990).

There are now enough examples to predict that any clustered gene family that encodes frequently paired glycine or proline residues will be undergoing concerted evolution promoted by gene conversion. While the location of such sequences in noncoding regions flanking the genes would also stimulate conversions, this would be an unstable situation, since insertion or deletion events could separate or remove the hotspot from the gene. Two explanations could account for the failure of the *ErB* genes to accumulate GG*X*GG*X* repeats. First, the function of the *ErB* proteins in the formation of the eggshell maybe incompatible with clusters of paired glycine residues in either their N- or C-arm. Second, the different *ErB* genes may serve unique functions in the formation of the eggshell. Sequence transfers between these genes would result in the loss of function and would be selected against.

Differences between the patterns of gene conversions in the early and late chorion gene pairs: Although the gene conversions in the *HcA*, *HcB* and *ErA* families share a number of similar features, one major difference exists. Within the *ErA* genes, few of the conversion events extend beyond their origin in the major exon, whereas in the *HcA* and *HcB* families the introns, minor exons and 5′ flanking regions are also involved (EICKBUSH and BURKE 1985; 1986). Based on hybridization experiments (SPOEREL *et al.* 1989) the gene conversions in the *A* and *B* chorion families probably also extend throughout the gene. However, since less sequence data is available from these large families to confirm this conclusion, the following discussion will be limited to a comparison of the *Hc* and *ErA* families.

The difference in the extent of the gene conversions between the *ErA* and *Hc* families could be explained simply by the conversion events occurring less frequently between the *ErA* genes. One can estimate the relative rates of conversion between the different families by comparing nucleotides at fourfold synonymous sites. These sites are not under selective restraint, thus each family should accumulate sequence changes at these sites at similar rates. The extent to which these changes are fixed or eliminated from within each family is a relative estimate of the conversion rate for each family. As described in this report, the level of divergence between the *ErA* genes at fourfold synonymous sites is 7.7%. In the case of the *HcA* and *HcB* genes the divergence at these sites is 6.7% and 7.9%, respectively (data calculated from BURKE and EICKBUSH 1986). Thus using nucleotide sequences at fourfold synonymous sites as a measure of the rate of sequence exchange, the frequency of conversion events in the protein encoding regions of the *ErA* genes is similar to that in the *HcA* and *HcB* families. Clearly an explanation other than rate of exchange must account for why these conversion tracts extend outside their origin in the major exon of the *Hc* genes, but not in the *ErA* genes.

It is possible to explain this difference between the extent of conversions in the two families as being simply the result of the same conversion process acting in an old *vs.* a young gene family. The early gene families are likely to be older than the *Hc* families since the early proteins (originally called C proteins) are found in all species of silk moths, while the *Hc* families appear to be a special adaptation of *B. mori*, which allows it to diapause as eggs (KAFATOS *et al.* 1977). Being older the early families could have accumulated a greater number of insertion/deletion differences in noncoding regions. These mutations would serve as a barrier to the passage of conversion

events. The late gene families may have simply expanded too recently to have accumulated sufficient insertion/deletion differences to significantly affect the passage of conversion events. There is an interesting example in the late genes that lends support to this explanation. In gene *HcA.3* a 0.9-kb insertion is located in the middle of its intron (EICKBUSH and BURKE 1986). The portion of the intron on the major exon side of the insertion has the same frequency of shared variants as found in all other *HcA* genes, suggesting that it continues to undergo gene conversion events. On the minor exon side of this insertion, the frequency of shared variants is only half that found for any other *HcA* gene, suggesting that sequence transfers are much less efficient at homogenizing this region of *HcA.3*. This large insertion appears to be acting as a barrier to gene conversions. Unless this insertion is eliminated (for example by unequal crossovers) the intron and minor exon of the *HcA.3* gene will become increasingly more divergent from the other members of the family.

Thus the *Hc* families reflect the gene conversion patterns that are possible in a newly expanded gene family, whereas in the early genes, conversion events have been limited by insertion/deletion differences to the regions where they originate, the major exon. The homogenization of the *ErA* major exons despite the extensive divergence of the *ErB* genes with which they are paired indicates that the conversion process, at least in cases involving recombination hotspots, can successfully maintain DNA sequence homology in localized regions of a gene over a significant evolutionary period. These conversions are inefficient at maintaining sequences that are subject to insertions/deletions, such as noncoding regions.

This work was supported by U.S. Public Health Service grant GM31867 from the National Institutes of Health. We thank JOHN IZZO for the use of unpublished data and for comments on the manuscript.

LITERATURE CITED

ADAMS, D. S., T. H. EICKBUSH, R. J. HERRERA and P. M. LIZARDI, 1986 A highly reiterated family of transcribed oligo(A)-terminated, interspersed DNA elements in the genome of *Bombyx mori*. J. Mol. Biol. **187:** 465–478.

ARNHEIM, N., 1983 Concerted evolution of multigene families, pp. 38–61 in *Evolution of Genes and Proteins*, edited by M. NEI and R. K. KOEHN. Sinauer, Sunderland, Mass.

ATCHISON, M., and M. ADESNIK, 1986 Gene conversion in a cytochrome P-450 gene family. Proc. Natl. Acad. Sci. USA **83:** 2300–2304.

BORTS, R. H., and J. E. HABER, 1989 Length and distribution of meiotic gene conversion tracts and crossovers in *Saccharomyces cerevisiae*. Genetics **123:** 69–80.

BURKE, W. D., and T. H. EICKBUSH, 1986 The silkmoth late chorion locus. I. Variation within two paired multigene families. J. Mol. Biol. **190:** 343–356.

COEN, E. S., T. STRACHAN and G. A. DOVER, 1982 Dynamics of concerted evolution of ribosomal DNA and histone gene families in the *melanogaster* species subgroup of *Drosophila*. J. Mol. Biol. **158:** 17–35.

CRAIN, W. R., M. F. BOSHAR, A. D. COOPER, D. S. DURICA, A. NAGY and D. STEFFEN, 1987 The sequence of a sea urchin muscle actin gene suggests a gene conversion with a cytoskeletal actin gene. J. Mol. Evol. **25:** 37–45.

DOVER, G. A., 1982 Molecular drive: a cohesive mode of species evolution. Nature **299:** 111–116.

DOVER, G. A., 1990 Mapping 'frozen accidents.' Nature **344:** 812–813.

EICKBUSH, T. H., and W. D. BURKE, 1985 Silkmoth chorion gene families contain patchwork patterns of sequence homology. Proc. Natl. Acad. Sci. USA **82:** 2814–2818.

EICKBUSH, T. H., and W. D. BURKE, 1986 The silkmoth late chorion locus. II. Gradients of gene conversion in two paired multigene families. J. Mol. Biol. **190:** 357–366.

EICKBUSH, T. H., and F. C. KAFATOS, 1982 A walk in the chorion locus of *Bombyx mori*. Cell **29:** 633–643.

EICKBUSH, T. H., G. C. RODAKIS, R. LECANIDOU and F. C. KAFATOS, 1985 A complex set of early chorion DNA sequences from *Bombyx mori*. Dev. Biol. **112:** 368–376.

GELIEBTER, J., and S. G. NATHENSON, 1988 Microrecombinations generate sequence diversity in the murine major histocompatibility complex: analysis of the K^{bm3}, K^{bm4}, K^{bm10}, and K^{bm11} mutants. Mol. Cell. Biol. **8:** 4342–4352.

GOLDSMITH, M. R., and E. CLERMONT-RATTNER, 1979 Organization of the chorion genes of *Bombyx mori*, a multigene family. II. Partial localization of three gene clusters. Genetics **92:** 1173–1185.

GOLDSMITH, M. R., and E. CLERMONT-RATTNER, 1980 Organization of the chorion genes of *Bombyx mori*, a multigene family. III. Detailed marker composition of three gene clusters. Genetics **96:** 201–212.

GOLDSMITH, M. R., and F. C. KAFATOS, 1984 Developmentally regulated genes in silkmoths. Annu. Rev. Genet. **18:** 443–487.

GUMUCIO, D. L., K. WEIBAUER, R. M. CALDWELL, L. C. SAMUELSON and M. H. MEISLER, 1988 Concerted evolution of human amylase genes. Mol. Cell. Biol. **8:** 1197–1205.

HIBNER, B. L., W. D. BURKE, R. LECANIDOU, G. C. RODAKIS and T. H. EICKBUSH, 1988 Organization and expression of three genes from the silkmoth early chorion locus. Dev. Biol. **125:** 423–431.

IATROU, K., S. G. TSITILOU and F. C. KAFATOS, 1982 Developmental classes and homologous families of chorion genes in *Bombyx mori*. J. Mol. Biol. **157:** 417–434.

JACKSON, J. A., and G. R. FINK, 1981 Gene conversion between duplicated genetic elements in yeast. Nature **292:** 306–307.

JEFFREYS, A. J., V. WILSON and S. L. THEIN, 1985 Hypervariable 'minisatellite' regions in human DNA. Nature **314:** 67–73.

JEFFREYS, A. J., R. NEUMANN and V. WILSON, 1990 Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and maturation by single molecule analysis. Cell **60:** 473–485.

JUDD, R. S., and T. D. PETES, 1988 Physical lengths of meiotic and mitotic gene conversion tracts in *Saccharomyces cerevisiae*. Genetics **118:** 401–410.

JUKES, T. H., and C. R. CANTOR, 1969 Evolution of protein molecules, pp. 21–132 in *Mammalian Protein Metabolism*, edited by M. N. MUNRO. Academic Press, New York.

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, pp. 45–145 in *Results and Problems in Cell Differentiation*, Vol. 8, edited by W. BEERMANN. Springer-Verlag, Berlin.

KOBORI, J. A., E. STRAUSS, K. MINARD and L. HOOD, 1986 Molecular analysis of the hotspot of recombination in the murine major histocompatibility complex. Science **234:** 173–179.

KONSOLAKI, M., K. KOMITOPOULOU, P. P. TOLIAS, D. L. KING, C. SWIMMER and F. C. KAFATOS, 1990 The chorion genes of the medfly, *Ceratitus capitata*. I. Structural and regulatory conservation of the *s36* gene relative to two *Drosophila* species. Nucleic Acids Res. **18:** 1731–1737.

LE BLANCQ, S. M., B. W. SWINKELS, W. C. GIBSON and P. BORST, 1988 Evidence for gene conversion between the phosphoglycerate kinase genes of *Trypanosoma brucei*. J. Mol. Biol. **200:** 439–447.

LECANIDOU, R., T. H. EICKBUSH, G. C. RODAKIS and F. C. KAFATOS, 1983 Novel B family sequence from an early chorion cDNA library of *Bombyx mori*. Proc. Natl. Acad. Sci. USA **80:** 1995–1959.

LECANIDOU, R., G. C. RODAKIS, T. H. EICKBUSH and F. C. KAFATOS, 1986 Evolution of the silkmoth chorion gene superfamily: Gene families *CA* and *CB*. Proc. Natl. Acad. Sci. USA **83:** 6514–6518.

LOENEN, W. A. M., and F. R. BLATTNER, 1983 Lambda Charon vectors (Ch32, 33, 34 and 35) adapted for DNA cloning in recombination-deficient hosts. Gene **26:** 171–179.

LYONS, K. M., J. H. STEIN and O. SMITHIES, 1988 Length polymorphisms in human proline-rich protein genes generated by intragenic unequal crossing over. Genetics **120:** 267–278.

MITSIALIS, S. A., and F. C. KAFATOS, 1985 Regulatory elements controlling chorion gene expression are conserved between flies and moths. Nature **317:** 453–456.

MITSIALIS, S. A., N. SPOEREL, M. LEVITEN and F. C. KAFATOS, 1987 A short defined DNA region is sufficient for developmentally correct expression of moth chorion genes in *Drosophila*. Proc. Natl. Acad. Sci. USA **84:** 7987–7991.

NAGYLAKI, T., and T. D. PETES, 1982 Intrachromosomal gene conversion and the maintenance of sequence homogeneity among repeated genes. Genetics **100:** 315–337.

NAKAMURA, Y., M. LEPPERT, P. O'CONNELL, R. WOLFF, T. HOLM, M. CULVER, C. MARTIN, E. FUJIMOTO, M. HOFF, E. KUMLIN and R. WHITE, 1987 Variable number of tandem repeat (VNTR) markers for human gene mapping. Science **235:** 1616–1622.

OHTA, T., 1980 *Evolution and Variation of Multigene Families*. Springer-Verlag, Berlin.

PARHAM, P., C. E. LOMEN, D. A. LAWLOR, J. P. WAYS, N. HOLMES, H. L. COPPIN, R. D. SALTER, A. M. WAN and P. D. ENNIS, 1988 Nature of polymorphism in HLA-A, -B, and -C molecules. Proc. Natl. Acad. Sci. USA **85:** 4005–4009.

PARK, Y.-S., and J. M. KRAMER, 1990 Tandemly duplicated *Caenorhabditis elegans* collagen genes differ in their modes of splicing. J. Mol. Biol. **211:** 395–406.

PETES, T. D., 1980 Unequal meiotic recombination within tandem arrays of yeast ribosomal RNA genes. Cell **19:** 765–774.

POWERS, P. A., and O. SMITHIES, 1986 Short gene conversions in the human fetal globin gene region: a by-product of chromosome pairing during meiosis? Genetics **112:** 343–358.

REYNAUD, C.-A., V. AUQUEZ, H. GRIMAL and J.-C. WEILL, 1987 A hyperconversion mechanism generates the chicken light chain preimmune repertoire. Cell **48:** 379–388.

RUPPERT, S., G. SCHERER, and G. SCHUTZ, 1984 Recent gene conversion involving bovine vasopressin and oxytocin precursor genes suggested by nucleotide sequence. Nature **308:** 554–557.

SANGER, F., S. NICKLEN, and A. COULSON, 1977 DNA sequencing with chain terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:** 5463–5467.

SEPERACK, P., M. SLATKIN, and N. ARNHEIM, 1988 Linkage disequilibrium in human ribosomal genes: implications for multigene family evolution. Genetics **119:** 943–949.

SMITH, G. P., 1973 Evolution of repeated DNA sequences by unequal crossovers. Science **191:** 528–535.

SMITH, G. R., S. M. KUNES, D. W. SCHULTZ, A. TAYLOR and K. L. TRIMAN, 1981 Structure of *Chi* hotspots of generalized recombination. Cell **24:** 429–436.

SPOEREL, N., H. T. NGUYEN, T. H. EICKBUSH and F. C. KAFATOS, 1989 Gene evolution and regulation in the chorion complex of *Bombyx mori*: Hybridization and sequence analysis of multiple developmentally middle *A/B* chorion gene pairs. J. Mol. Biol. **209:** 1–19.

STEINMETZ, M., D. STEPHAN and K. F. LINDAHL, 1986 Gene organization and recombination hotspots in the murine major histocompatibility complex. Cell **44:** 895–904.

SZOSTAK, J. W., and R. WU, 1980 Unequal crossing over in the ribosomal DNA of *Saccharomyces cerevisae*. Nature **284:** 426–430.

VOLLRATH, D., J. NATHANS and R. DAVIS, 1988 Tandem array of human visual pigment genes at Xq28. Science **240:** 1669–1671.

WHEELER, C. J., D. MALONEY, S. FOGEL and R. S.GOODENOW, 1990 Microconversion between murine *H-2* genes integrated into yeast. Nature **347:** 192–194.

WHITE, J. H., J. F. DIMARTINO, R. W. ANDERSON, K. LUSNAK, D. HILBERT and S. FOGEL, 1988 A DNA sequence conferring high postmeiotic segregation frequency to heterozygous deletions in *Saccharomyces cerevisiae* is related to sequences associated with eucaryotic recombination hotspots. Mol. Cell. Biol. **8:** 1253–1258.

XIONG, Y., B. SAKAGUCHI and T. H. EICKBUSH, 1988 Gene conversion can generate sequence variants in the late chorion multigene families of *Bombyx mori*. Genetics **120:** 221–231.

YANISCH-PERRON, C., J. VIEIRA and J. MESSING, 1985 Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. Gene **33:** 103–119.

Communicating editor: W.-H. LI