J Mol Evol (2000) 51:329–338
DOI: 10.1007/s002390010095

JOURNAL OF MOLECULAR EVOLUTION

Dynamic Rearrangement Within the *Antheraea pernyi* Silk Fibroin Gene Is Associated with Four Types of Repetitive Units

Hideki Sezutsu, Kenji Yukuhiro

Department of Insect Genetic Breeding, National Institute of Sericultural and Entomological Science, 1-2 Ohwashi, Tsukuba, Ibaraki 305-8634 Japan

Received: 18 February 2000 / Accepted: 30 June 2000

Abstract. We characterized a full-length gene encoding wild silkmoth *Antheraea pernyi* fibroin (Ap-fibroin) to clarify the conformation of repetitive sequences. The gene consisted of a first exon encoding 14 amino acid residues, a short intron (120 bp), and a long second exon encoding 2,625 amino acid residues. Three amino acids, alanine, glycine, and serine, amounted to 81% of the Ap-fibroin sequence. The Ap-fibroin, except for 155 residues of the amino terminus, was composed of 80 tandemly arranged polyalanine-containing units (motifs). A motif was a doublet of a polyalanine block (PAB) and a nonpolyalanine block (NPAB). Seventy-eight of the 80 motifs were classified into four types based on differences in the NPAB sequences. Although respective motifs were significantly conserved, many rearrangements were observed within the second exon, i.e., the triplication of a 558-bp-long sequence and other duplication events of shorter sequences. Chi-like sequences, GCTGGAG, might contribute to the rearrangement within the gene as described in human minisatellite loci, because they were found at specific sites of NPAB-encoding sequences in three of four types of motifs. The present results support the idea that the Ap-fibroin gene is unstable like minisatellite sequences and that the evolution of this gene is strongly associated with its instability.

Key words: Silk fibroin — *Antheraea pernyi* — Repetitive motifs — Polyalanine block (PAB) — Nonpolyalanine block (NPAB) — Rearrangement by duplications — Chi sequence — Coding minisatellite sequence

Correspondence to: K. Yukuhiro

Introduction

Silks play important roles in the lives of arthropods, for example, cocoon silks produced by lepidopteran insects, silk egg stalk by Neuropteran insects, underwater silk prey capture nets by Trichopteran insects, and web building by spiders (reviewed by Craig 1997). Silks are composed of one or more proteins called fibroins. For example, the dragline silk of the spider *Nephila clavipes* consists of two components, spidroins 1 and 2 (Xu and Lewis 1990; Hinman and Lewis 1992), but a single fibroin protein has been suggested to form the silk of the *Antheraea* moth (Tamura et al. 1987). Fibroins are large proteins composed of several simple amino acid sequences organized in reiterated arrays (Lucas and Rudall 1968) and can be classified into many types by structures deduced from X-ray diffraction analysis (Lucas and Rudall 1968).

Genes that encode proteins with repetitive structures are often called coding-minisatellite sequences (e.g., Paulsson et al. 1992). They seem to have properties similar to minisatellite sequences, which are tandemly repeated sequences of 10–100 bp units (Jeffreys et al. 1985). Minisatellite sequences are highly unstable components of genomes. Their instability is associated with changes in number and arrangement of repeated units induced by frequent unequal crossing-over and/or gene conversion events (Jeffreys et al. 1985). Jeffreys et al. (1985) suggested that the Chi sequence (GCTGGTGG) mediates the instability of human minisatellite loci. The Chi sequence is a RecBCD-mediated recombination hot spot in *Escherichia coli* (Lam et al. 1974). Genomic rearrangement through unequal crossing-over and/or

gene conversion and other transposition-like events are significant driving forces of evolution, in events such as exon shuffling.

To date no fibroin genes have been fully structurally characterized because of its instability when cloned. Characterization of a full-length fibroin gene would enable us to determine whether its instability is similarly mediated as minisatellite sequences and to understand how they have evolved.

We determined the entire sequence of a fibroin gene from a species of giant silkmoth, *Antheraea pernyi*, also known as the Chinese oak silkmoth. This species produces a quite different fibroin from that of the domesticated silkmoth, *Bombyx mori.* Eighty polyalanine block (PAB)–containing units were tandemly arranged in *A. pernyi* fibroin (Ap-fibroin). We defined a PAB-containing unit as a "motif." The motifs were classified into four types by differences in the sequences of non-polyalanine block (NPAB) parts. They were not distributed at random; that is, particular types of motifs were coupled preferentially. We found that a 558-bp-long sequence corresponding to seven motifs repeated three times with no differences in nucleotide sequence. Other duplications of shorter sequences were also observed. It is notable that individual NPAB encoding sequences in three of four types of motifs possessed a Chi-like sequence (GCTGGAGG), and the remaining motif lacked a Chi-like sequence and was always coupled with a different motif. Present results clearly show that dynamic rearrangements occurred within the Ap-fibroin gene and that this gene is potentially unstable like minisatellite sequences.

Materials and Methods

Samples. Genomic DNA was prepared from a pair of silkglands of a final instar larva of *A. pernyi* based on a standard technique (Sambrook et al. 1989). Larvae were the kind gift of Dr. S. Hayasaka.

Genomic DNA Library Preparation and Screening. Genomic DNA was partially digested by *Mbo* I and fractionated through agarose gel electrophoresis, and 13- to 23-kb-long fragments were isolated and cloned into a Lambda DASH II phage vector (Stratagene). GigaPack® III gold (Stratagene) was used for in vitro packaging. We used the recombination deficient *SURE Escherichia coli* strain (Stratagene) to decrease the possibility of genetic rearrangement of cloned DNAs when propagating phages or plasmids. We screened this library using a 1350-bp-long Ap-fibroin cDNA (Yukuhiro et al. 1997) as a probe. Seven clones were isolated and one clone, Ap2, yielded a full-length fibroin gene. The other six clones contained truncated fibroin genes.

Sequencing Strategy and Analysis. The Ap2 insert was subcloned into p*Bluescript* II SK(+) (Stratagene). We prepared deletions of this plasmid using a Kilo-Sequence Deletion Kit (Takara) and selected about 100 clones by size using long agarose gel electrophoresis. These were sequenced by combining a DyeTerminator Cycle Sequencing FS Ready Reaction Kit (PE Biosystems) and ABI Autosequencers 373S and 377. We then aligned these nucleotide sequences using a Sequence Navigator (PE Biosystems).

The amino acid sequence and the hydrophobic/hydrophilic plot for this sequence were deduced using Genetyx Mac Ver. 10.1 (SDC). We also used Genetyx Mac Ver. 10.1 for codon usage analysis and Harrplot 2.0 (SDC) for HarrPlot analysis (Starden 1982).

Using MEGA (Kumar et al. 1993), we estimated the average heterozygosity per nucleotide site for NPAB sequences of the same type of motif by estimating the mean p-distance (Kumar et al. 1993). All gap sites were excluded from the analysis. An average heterozygosity per amino acid site for NPABs of the same type of motif was also estimated by a similar procedure.

Results

Ap-fibroin Gene Structure and Deduced Amino Acid Sequence

In clone Ap2, we identified a 9282-bp nucleotide sequence including a full-length *A. pernyi* fibroin gene (about 8.1 kb) and flanking sequences (Fig. 1). The Ap-fibroin gene consisted of the first exon, which encoded 14 amino acid residues, an intron (120 bp), and a second exon encoding 2625 residues (Fig. 1). The deduced amino acid sequence is shown in Fig. 2. Three amino acids, alanine (A) (43%), glycine (G) (27%), and serine (S) (11%), together accounted for 81% of the Ap-fibroin sequence, which is consistent with analysis of the purified protein (Kirimura 1962; Fraser and MacRae 1973).

The first exon and the 5′-region of the second exon encoded 155 residues of amino terminal sequence (Fig. 2), which was unique along the Ap-fibroin. Three amino acids, alanine, glycine, and serine, were not abundant in this region. Thus, this region of the Ap-fibroin gene showed different features of hydrophobicity from the rest (Fig. 1).

Except for the first 155 residues, the Ap-fibroin sequence consisted of 80 polyalanine-containing units (motifs). A PAB and NPAB constituted a motif (Fig. 2). The amino acid sequences encoded by NPABs were more hydrophilic than the PAB sequences; therefore there was alternation of hydrophobic and hydrophilic regions (Fig. 1).

Seventy-eight of the 80 motifs were classified into four types (Types 1, 2, 3, and 4) based on differences in the NPAB sequences. Consensus amino acid sequences of respective types are shown in Fig. 3a. In Types 1, 2, and 3, the NPAB sequences are similar in the first segment (SGAGG or GSGAGG) and the last segment (GGYGSDS or GGYGSGSS), but differ in the middle of the block (Fig. 3a). When the three motifs are ordered, a PAB and the similar portions of adjacent NPABs are regarded as a constant domain, whereas the region between the two constant domains is a variable domain (Fig. 3b). Note that the nucleotide sequence GCTGGAGG encodes an alanine-glycine-glycine triplet (Fig. 3a) in most of the constant domains and is similar to a Chi site.

Type 1 motifs carried nine residues, including a tryp-

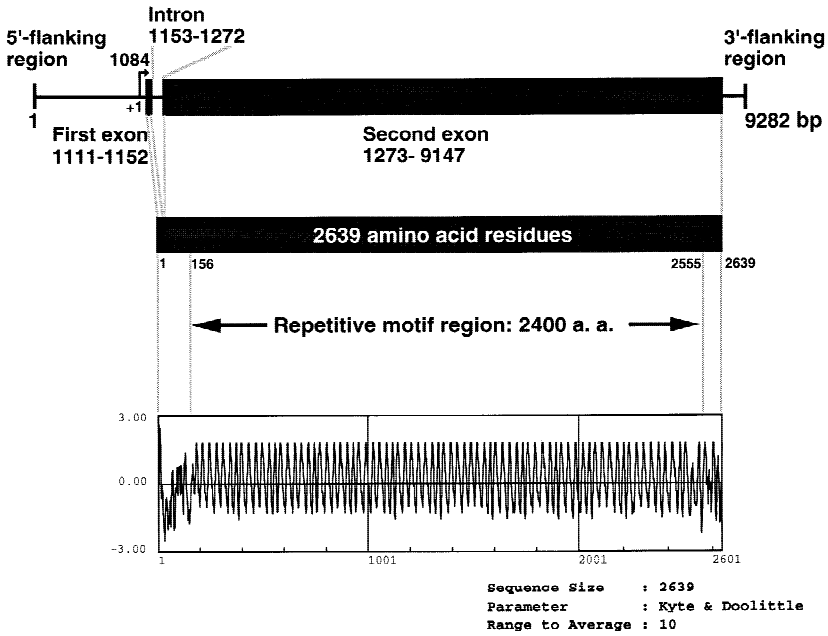

Fig. 1. Structure of the Ap-fibroin gene and hydrophobic/hydrophilic plot of deduced amino acid sequence. The Ap-fibroin gene structure in a genomic clone Ap2 is shown at the top. Closed boxes indicate the coding sequences. The number of amino acid residues in the deduced sequence is illustrated in the middle. The position 1084 is a possible transcription initiation site according to Tamura et al. (1986) and thus is marked as +1. The repetitive motif region corresponds to the amino acid sequence from position 156 to 2555 in Fig. 2. The hydrophobic/hydrophilic plot of the deduced amino acid sequence is shown at the bottom according to Kyte and Doolittle (1982).

tophan, in a variable domain and were further classified into four subtypes (1S, 1A, 1V, and 1R) based on differences in the sixth or seventh amino acid sites of the NPAB sequence (Fig. 3a). Type 2 motifs differed in the number of Gly-Gly-Tyr (GGY) triplets, inferring the occurrence of replication slippage events. Each Type 3 motif carried an Arg-Gly-Asp (RGD) triplet, which is the cell adhesion signal of fibronectin (Hynes 1987). NPAB sequences of Type 4 motifs were highly heterologous to those of other motifs, corresponding to the Type 2 motif described by Yukuhiro et al. (1997).

NPAB sequences of two motifs at the carboxyl terminus (79th and 80th motifs) were different in the appearance of amino acid residues that were rare in the four basic types of motifs. For example, together the two exceptional motifs included three cysteine residues and three leucines, which were not seen in the other 78 motifs (Fig. 2).

Motif Distribution and Dynamic Rearrangement Within the Second Exon

Figure 3d shows the arrangement of individual motifs along the gene. A Type 3 motif is always followed by a Type 4 motif, although no other preferential coupling of other motifs was observed. The coupled array of Type 3 and Type 4 is shown in Fig. 3c. It is notable that the Type 4 motif lacks a constant NPAB domain, which means there is no Chi-like sequence in the Type 4 motif. Most of these coupled arrays (9/12) are linked to a Type 1S, subtype of the Type 1 motif.

Harrplot analysis (Starden 1982) along the Ap-fibroin DNA sequence detected some sequence fragments more than once (Fig. 4). We used a sliding window to clarify the distribution of respective motifs: A dot is plotted when 95 of 100 bases match (Fig. 4). Short ranges of multiplication were dispersed along the gene, although they were limited to the repetitive motif region described in Fig. 1. We also found a few long stretches of sequences that encoded several motifs. Fragment A in this figure, for example, is 558 bp long and is tandemly repeated three times. The triplicated sequences showed no variation in the nucleotide sequence (data not shown).

Variation of Motif Sequence

The amino acid sequences in the NPABs of respective motifs are aligned in Fig. 5. Although extensive uniformity of respective motifs was observed, there were a few variants, i.e., 1st, 72nd, 74th, 76th, and 78th motifs. These variant motifs tended to be located at the carboxyl terminus. They were excluded from the following analysis (see Fig. 5).

We identified amino acid differences at seven sites in the NPAB sequence of Type 1 motifs: One site was highly polymorphic, with four different amino acid residues observed (Fig. 5). Insertion or deletion of a glycine residue was polymorphic at the first amino acid site. All Type 1 motifs next to a Type 4 motif showed the deletion at this site (Fig. 2). An additional serine residue was found at the end of the 77th motif.

In Type 2 motifs, amino acid changes were observed at two sites (Fig. 5). Insertion or deletion polymorphism except for the repetition of a GGY triplet was also observed at the first amino acid site.

Type 3 motifs were highly conserved; a variation was seen only in the first amino acid site of the seventh motif. Three variable amino acid sites were seen in Type 4 motifs.

Fig. 2. Deduced Ap-fibroin amino acid sequence. Sequences encoded by the first and second exons are separated. The polyalanine-containing units (motifs) are shown on every other line and numbered (1 to 80). The polyalanine block (PAB) and nonpolyalanine block (NPAB) sequences are aligned, using a dot as a gap. The amino acids other from alanine, glycine, and serine are written in bold letters. A cysteine residue is underlined. Nucleotide and protein sequence data is in the Genbank database under accession number AF083334.

The number of variable amino acid sites is summarized in Table 1. Average heterozygosities per amino acid site of four types of motifs are also tabulated. The score for the Type 3 motif was 0.0107 ± 0.0031 smaller than any other scores. We also found a very low variation in nucleotide sequences encoding NPABs of Type 3 motifs; the average heterozygosity per nucleotide site was 0.0157 ± 0.0017.

Biased Codon Usage

Codon usage patterns of the Ap-fibroin gene were strongly biased toward A- or U-ended isocodons (Table 2). The GGU isocodon was used at 299 of 720 glycine residues, and UCA isocodons (171/297) were preferred for serine residues. The GCA alanine isocodon was most abundant (662/1137), which consists of 25% of the cod-

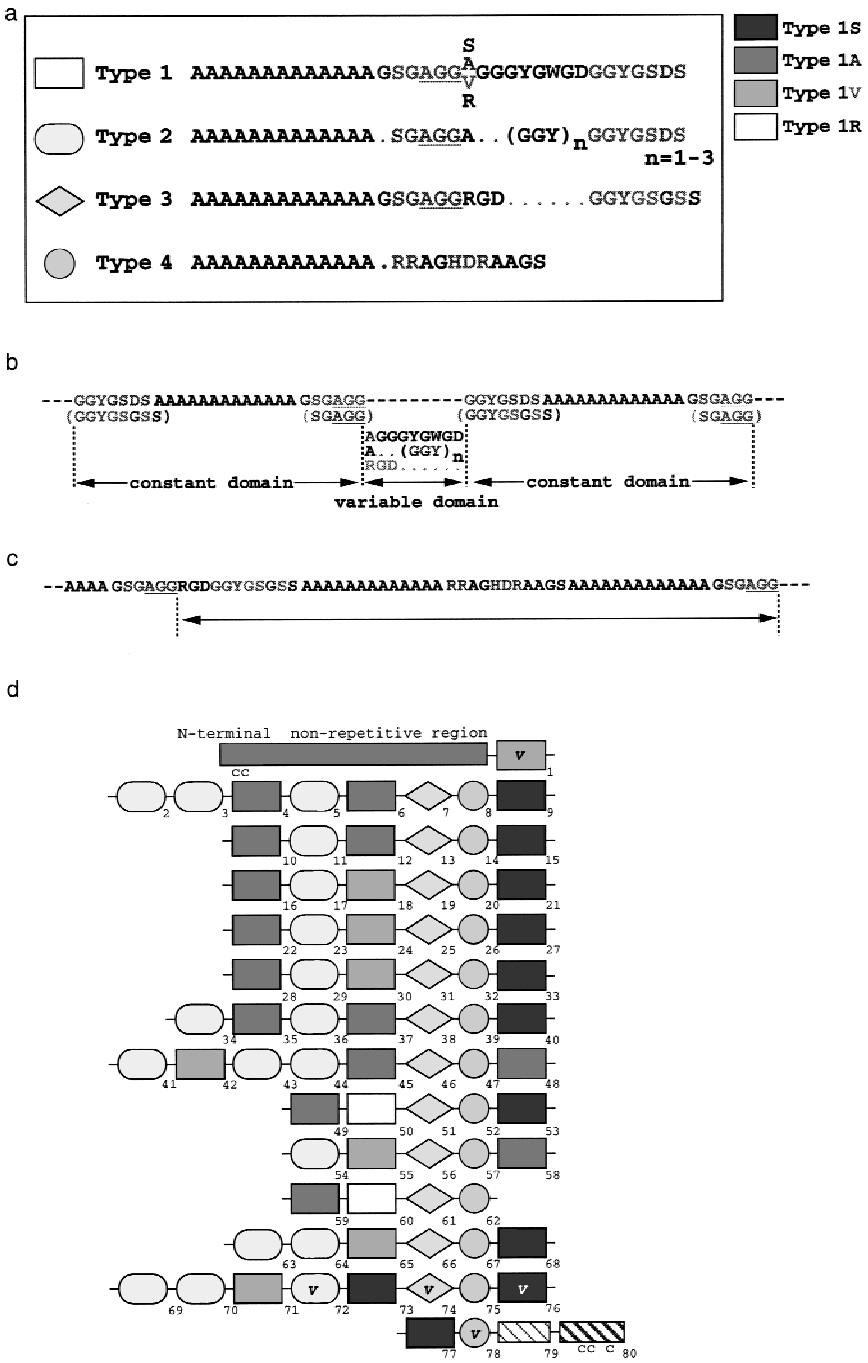

Fig. 3. Definition and distribution of motifs. **a:** Definition of four types of motifs. Symbols and aligned consensus amino acid sequences of motifs are illustrated. A dot means a gap. The AGG triplet encoded by the Chi-like sequence is underlined. A polymorphic amino acid site in the Type 1 motif is indicated by four residues after the AGG triplet. The symbols of the four subtypes are presented on the right. For Type 2 motifs, n means the repetition number of the GGY triplet. In Type 4 motifs, R, H, and D are hydrophilic amino acid residues. **b:** Constant domain and variable domain in Types 1, 2, and 3. **c:** Coupled array of the Types 3 and 4. **d:** Distribution of motifs. Arrangement of motifs is expressed using symbols. The number in the lower right of the symbol corresponds to the motif number in Fig. 2. Subtypes of Type 1 motifs are shown by different patterns. The letter V indicates a variant sequence motif. The nonrepetitive region at the amino terminus and two nonrepetitive motifs at the carboxyl terminus are shown by other rectangles. The letter C below the two termini indicates a cysteine residue.

ing region, in contrast to that of the *Bombyx* fibroin heavy chain (FHC) gene, where the GCU isocodon was frequently used (Mita et al. 1994). Preference of A- or U-ended isocodons might be interpreted to reduce a total G + C content (counterbalance effect: Nakamura et al. 1991).

Characteristics of PAB Sequences

The number of alanine residues in PABs ranged from 10 to 15 except for the 1st and 77th motifs (Fig. 2). Most of the PABs consisted of 12 or 13 residues, that is, the number of PABs consisting of 12 alanine residues was 25, and that of 13 residues was 41. Type 3 motifs tended to have a PAB with 12 alanine residues. Although PABs of 14th, 20th, and 26th motifs included a serine residue, these motifs were associated with the triplication events described above.

To classify PABs based on differences in nucleotide sequence, we aligned and compared PAB nucleotide sequences depending on their 5′ motif (Fig. 6). Although information was limited in the third codon position, we found the following features: (1) The maximum number of repeated GCA isocodons in each PAB was seven, and

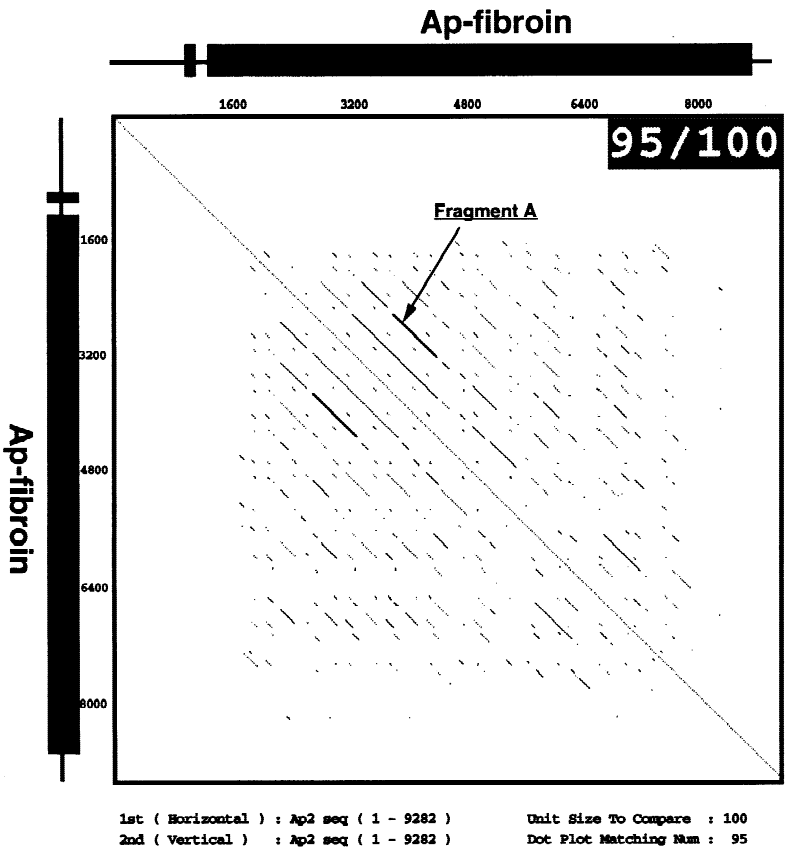

Fig. 4. Harrplot analysis of Ap-fibroin nucleotide sequence. A dot is plotted when 95 of 100 bases match. Fragment A is a 558-bp-long sequence, which is tandemly triplicated. This sequence encodes 186 amino acid residues, corresponding to NPAB of motif 13 to PAB of motif 19, NPAB of motif 19 to PAB of motif 25, and NPAB of motif 25 to PAB of motif 31, in Fig. 2.

no PABs were occupied by a single isocodon. (2) Identical PAB sequences repeatedly appeared along the repetitive motif region. This repetition might be associated with the rearrangement within the second exon.

5′-Flanking and Intronic Sequence

We determined a 1110-bp-long 5′-flanking sequence showing a high similarity (91.0%) in nucleotide sequence to that of the *Antheraea yamamai* fibroin (Ay-fibroin) gene (Genbank accession number: X05578; Tamura et al. 1987). Tamura et al. (1987) suggested that several elements in the 5′-flanking sequence of the Ay-fibroin gene correspond to the transcriptional regulatory elements of the *Bombyx* FHC gene (Genbank accession number: V00094; Tsujimoto and Suzuki 1979). These elements were also present in the Ap-fibroin 5′-flanking sequence.

The intronic sequence of the Ap-fibroin gene also showed high similarity to the Ay-fibroin intronic sequence (109/120 = 0.908) (Genbank accession number: X05578; Tamura et al. 1987). The Ay-fibroin intron was longer than that of the Ap-fibroin because the Ay-fibroin intron carried a 30-bp-long A + T–rich fragment, which is probably an insertion.

The Ap-fibroin intron showed no nucleotide sequence similarity to that of the *Bombyx* FHC gene (970 bp) (V00094; Tsujimoto and Suzuki 1979). The nucleotide sequence of the *Bombyx* FHC second exon (Mita et al. 1994) also showed little similarity to the Ap-fibroin second exonic sequence.

Discussion

The Ap-fibroin consists of three regions: a nonrepetitive amino terminal region, a long region of 78 repetitive motifs, and a region of 2 unique motifs at the carboxyl terminus. The amino acid sequence for the first exonic sequence shows a high identity to the Ay-fibroin (Tamura et al. 1987), with 13 of the 14 residues identical, and is identical at 11 of the 14 residues of the *Bombyx* FHC (Tsujimoto and Suzuki 1979). The sequence conservation of this region indicates its functional significance, particularly at the two conserved cysteine residues. For example, Tamura et al. (1987) have suggested that the Ay-fibroin forms a homodimer mediated by disulfide bonds. It is also possible that the conserved amino-terminal sequence may work as a signal peptide essential for fibroin secretion.

Four types of motifs that differ in NPAB sequences were found in the repetitive motif region (Fig. 1). The partial sequence of the Ay-fibroin gene carried motifs significantly similar to Ap-fibroin motifs (Yukuhiro et al. 1997). Kirimura (1962) reported that fibroins of five Saturniidae species included multiple PABs, although the

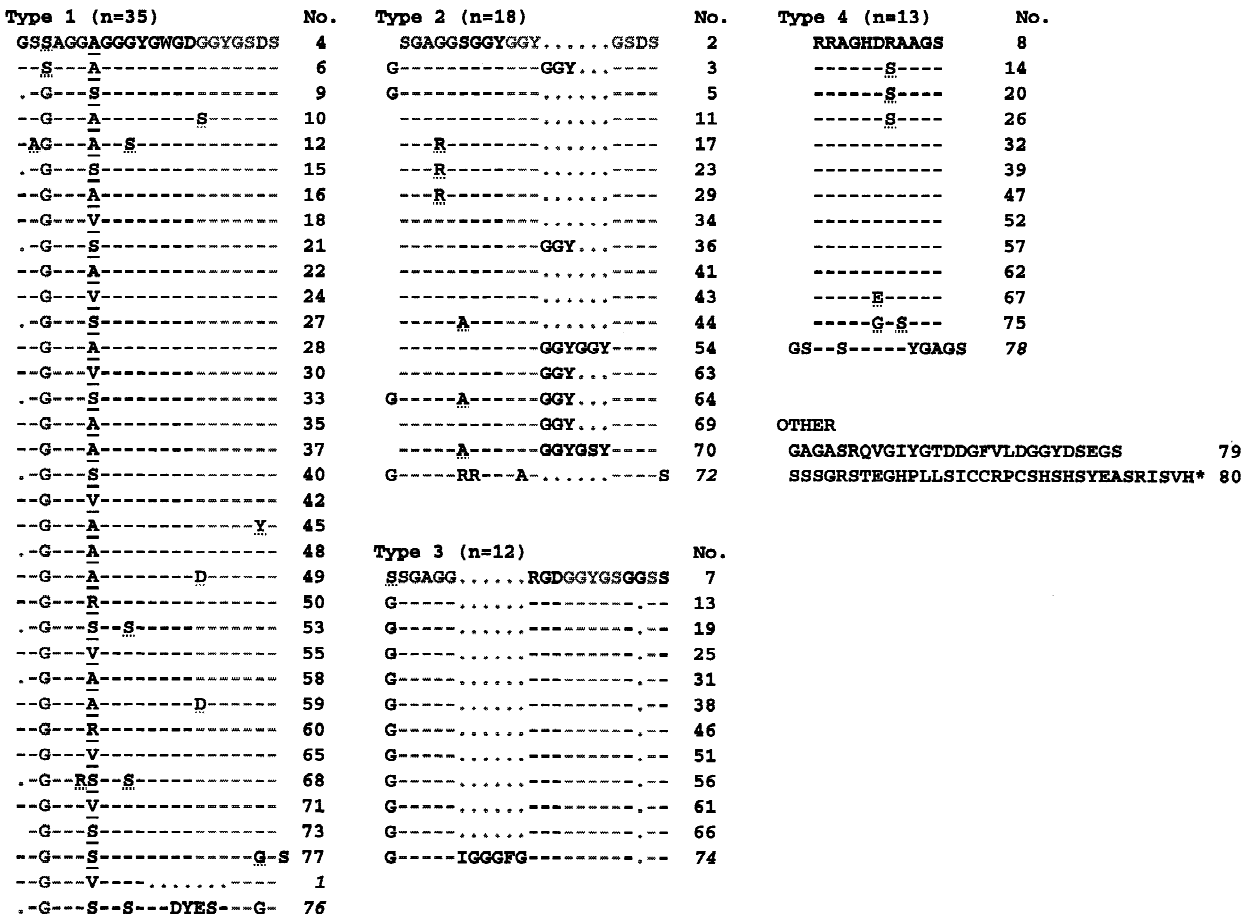

Fig. 5. Aligned data set of NPAB amino acid sequence for each type of motif. n is the number of motifs. Motifs are aligned based on their position from the amino terminus to the carboxyl terminus. Polymorphic sites in Type 1 NPABs are underlined. Other variable sites are shown with dotted underline. A dot means a gap site, which is regarded as a deletion or insertion of an amino acid residue. – indicates an identical residue. The motifs whose numbers are written in italic were eliminated from the analysis of Table 1.

Table 1. Variation in NPAB amino-acid and nucleotide sequence of motifs*

	No. of NPABs examined	No. of variable amino acid sites	Average heterozygosity per amino acid site (±SE)	No. of variable nucleotide sites	Average heterozygosity per nucleotide site (±SE)
Type 1	33	7	0.0697 ± 0.0018	24	0.0807 ± 0.0016
Type 2	17	2	0.0386 ± 0.0033	9	0.0591 ± 0.0024
Type 3	11	1	0.0107 ± 0.0031	3	0.0157 ± 0.0017
Type 4	12	3	0.0812 ± 0.0094	8	0.0707 ± 0.0079

* Insertion or deletion polymorphisms are not shown
SE indicates standard error

composition of amino acids was different. The divergence of amino acids contents might be attributable to differences in NPAB sequences. These observations strongly suggest that PAB sequences were highly conserved while NPAB are under less stringent selective constraint. From another viewpoint, the properties of fibroins might depend on NPAB sequences.

Repeated PABs coupled with Gly-rich NPABs are also seen in spider dragline-silk fibroins (Xu and Lewis 1990; Hinman and Lewis 1992; Guerette et al. 1996), which also prefer the same alanine isocodon (GCA). However, these fibroins show some different features: The numbers of alanine residues in PABs (4 to 10) are fewer than those of the Ap-fibroin and the some of spider fibroins are much richer in proline but poor in serine (Hinman and Lewis 1992; Guerette et al. 1996). We cannot conclude that spider fibroins contain two or more different motifs because this work was based on analysis of complementary DNAs (cDNAs), and therefore these data do not cover the entire coding sequence.

The expression of *Bombyx* FHC and Ay-fibroin genes is not observed in tissues other than posterior silkglands

Table 2. Codon usage of *A. pernyi* fibroin gene

Codon		No. (%)	Codon		No. (%)	Codon		No. (%)	Codon		No. (%)
Phe (F)	UUU	2 (0.08)	**Ser** (S)	UCU	43 (1.63)	Tyr (Y)	UAU	102 (3.86)	Cys (C)	UGU	1 (0.04)
	UUC	4 (0.15)		UCC	15 (0.57)		UAC	37 (1.40)		UGC	4 (0.15)
Leu (L)	UUA	4 (0.15)		**UCA**	**171 (6.48)**	Ter*	UAA	1 (0.04)	Ter*	UGA	0 (0.00)
	UUG	2 (0.08)		UCG	29 (1.10)	Ter*	UAG	0 (0.00)	TRP (W)	UGG	34 (1.29)
Leu (L)	CUU	3 (0.11)	Pro (P)	CCU	0 (0.00)	His (H)	CAU	20 (0.76)	Arg (R)	CGU	16 (0.61)
	CUC	1 (0.04)		CCC	0 (0.00)		CAC	5 (0.19)		CGC	2 (0.08)
	CUA	0 (0.00)		CCA	3 (0.11)	Gln (Q)	CAA	3 (0.11)		CGA	28 (1.06)
	CUG	0 (0.00)		CCG	2 (0.08)		CAG	2 (0.08)		CGG	0 (0.00)
Ile (I)	AUU	4 (0.15)	Thr (T)	ACU	4 (0.15)	Asn (N)	AAU	4 (0.15)	**Ser** (S)	AGU	34 (1.29)
	AUC	3 (0.11)		ACC	1 (0.04)		AAC	1 (0.04)		AGC	5 (0.19)
	AUA	4 (0.15)		ACA	3 (0.11)	Lys (K)	AAA	2 (0.08)	Arg (R)	AGA	23 (0.87)
Met (M)	AUG	1 (0.04)		ACG	1 (0.04)		AAG	1 (0.04)		AGG	2 (0.08)
Val (V)	GUU	3 (0.11)	**Ala** (A)	GCU	169 (6.40)	Asp (D)	GAU	41 (1.55)	**Gly** (G)	**GGU**	**299 (11.33)**
	GUC	5 (0.19)		GCC	78 (2.95)		GAC	77 (2.92)		GGC	198 (7.50)
	GUA	12 (0.45)		**GCA**	**662 (25.08)**	Glu (E)	GAA	14 (0.53)		GGA	192 (7.27)
	GUG	1 (0.04)		GCG	228 (8.64)		GAG	3 (0.11)		GGG	31 (1.17)

Numbers in parentheses are frequencies of the codon
Ter* indicates a termination codon

(Suzuki et al. 1986). In contrast, spiders produce silks of different composition by gland-specific expression. For example, Guerette et al. (1996) reported four kinds of fibroin cDNAs of a spider, *Araneus diadematus.* In a different species of spider, *Nephila clavipes,* the dragline silk consists of at least two different fibroins (Xu and Lewis 1990; Hinman and Lewis 1992). These results indicate that gene duplication and consequent sequence divergence adapted for tissue-specific roles might be associated with structurally varied spider fibroins. To determine whether fibroin genes of *A. pernyi* and spiders share a common ancestral gene and use a similar mechanism of gene expression, we need to know complete sequences of the spider fibroin genes, including the 5′-flanking sequences.

We detected multiple occurrences of duplication events, including a triplication of a 558-bp sequence. This strongly indicates that dynamic rearrangements occurred within the gene, which means that the Ap-fibroin gene is unstable. The distribution patterns of repeated units of the Ap-fibroin gene are much more complicated than minisatellite sequences; that is, four types of repeats corresponding to conserved motifs disperse along the Ap-fibroin gene, whereas a simple repeated unit is tandemly arranged in a minisatellite sequence.

Although recombination sites or hot spots in most eukaryotes remain unclear, Jeffreys et al. (1985) suggested the significant effect of the Chi sequence (GCTGGTGG) as a recombination "hot spot" in human minisatellite loci. The Chi sequence in *E. coli* plays a key role in the RecBCD-mediated recombination (Lam et al. 1974). Most of Type 1, 2, and 3 motifs contain a Chi-like sequence at the same position, corresponding to the AGG triplet in the constant domain (Fig. 3a). Because Type 4 motifs lack the Chi-like sequence, it cannot contribute to the hypothesized Chi-like sequence–mediated rearrangement; this can account for the coupling of Type 3 with Type 4 (Fig. 3c).

We found very low nucleotide sequence variation in NPABs of the Type 3 motif. Although a strict functional constraint on Type 3 motifs could induce low levels of variation in amino acid sequence, it cannot explain this result. This indicates more efficient turnover or gene conversion covering the Type 3 NPABs.

Sequence rearrangement, including unequal crossing-over and gene conversion, seems not only to shuffle the organization of motif-coding sequences but also to generate new types of motifs. As described above, NPABs of three types of motifs, Types 1, 2, and 3, can be divided into two domains, a constant domain and a variable domain. Classification of Types 1, 2, and 3 was mainly dependent on the difference in sequence of variable domains. If a nucleotide sequence that encodes the Type 1 motif is an original one, a rearrangement in sequence coding for the variable domain can generate the sequence for the Type 2 motif. That is, mismatching on nucleotide sequence GGXGGX (X meaning base), encoding paired glycine residues, and subsequent recombination between two Type 1 motif coding sequences induces a deletion of a nucleotide sequence that encodes six amino acid residues, GYGWGD, to produce the nucleotide sequence coding for a Type 2 motif. Such a mismatching mediated by a GGXGGX nucleotide sequence was suggested by Hibner et al. (1991), where gene conversion in the *B. mori* chorion locus was discussed.

A combination of this sort of rearrangement and additional amino acid substitutions can generate the Type 3 motif. Some chimeric motifs, e.g., the 1st and 78th motifs, seem to result from internal rearrangement of the motif-encoding sequence. This suggests another scenario to account for the formation of the Type 3 encoding sequence: an unequal cross over between a Type 1-en-

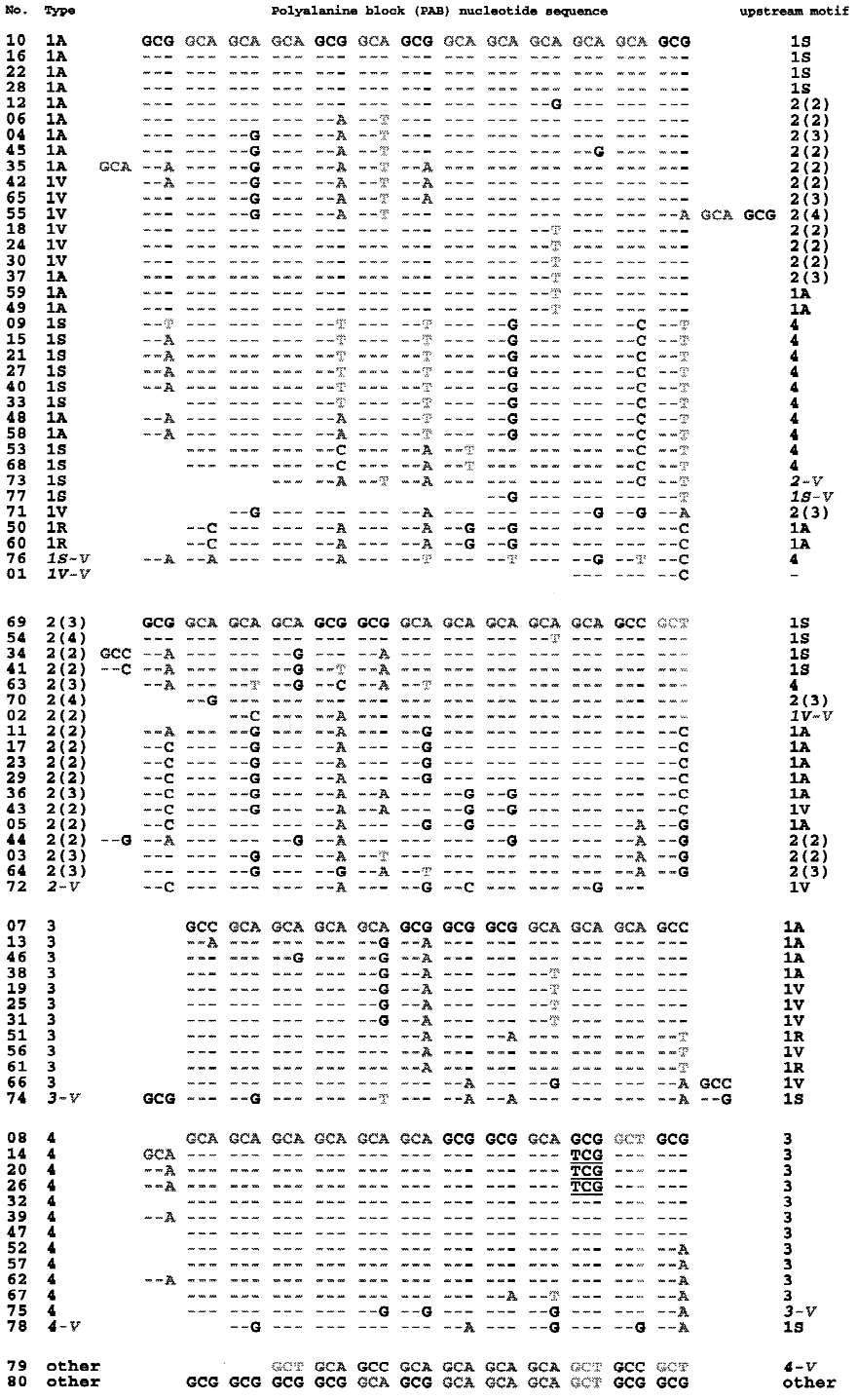

Fig. 6. Nucleotide sequences encoding PABs. Letters in the first column indicated the numbers of the motifs corresponding to those in Fig. 2. In the second column, classes of motifs are shown. The number in parentheses of Type 2 shows the number of GGY triplets. The motif written in italic indicates that it is a variant sequence motif. Nucleotide sequences for the PABs are presented in the third column. A hyphen (-) indicates an identical nucleotide. TCG codons in PAB of motif 14, 20, and 26 encode serine residues. In the final column, classes of motifs upstream from the PABs are shown.

coding sequence and a sequence that coded for a motif that was lost from the gene resulted in the generation of the Type 3 motif. Similar mechanisms can produce the Type 4 motifs. However, the Type 4 motifs are highly different in amino acid sequence from other three basic motifs, strongly suggesting that the contribution to the secondary structure different from other motifs. It is possible that the generation of Type 4 motifs was associated with positive natural selection.

Different alanine isocodons were preferentially used between Ap-fibroin and *Bombyx* FHC genes, although preferential use of A- or U-ended isocodons is a common

feature of fibroin genes (Mita et al. 1994; Xu and Lewis 1990; Hinman and Lewis 1992; Guerette et al. 1996). Most alanine residues in Ap-fibroin consist of PABs, whereas the *Bombyx* FHC gene lacks PABs (Mita et al. 1994). Therefore, the polyalanine structure probably contributes to the difference in the codon usage for alanine.

As a polyalanine block could be regarded as a sort of trinucleotide repeat sequence (Caskey et al. 1992), we had expected large variations in the number of alanine residues in the block due to replication slippage events. The average number of alanine residues in a PAB was, however, about 13 and variation was small (Fig. 6), suggesting a significant degree of constraint on the size of PABs. There are no PABs that consist of the GCA isocodon alone despite the abundance of the GCA isocodon. Other alanine isocodons appeared in the middle of a PAB, preventing occupation by GCA isocodons. In stable alleles of trinucleotide repeats, one to three point mutations, or interruptions, occur to disrupt the perfect repeat tract (Rolfsmeier and Lahue 2000). The interruption of the perfect GCA tracts in PABs may be related to the stabilizing mechanism of trinucleotide repeats to suppress the expansion of GCA tracts.

Though amino acid sequence uniformity of respective motifs indicates strong functional constraint on them, duplication and subsequent rearrangement of sequences with multiple motifs infers that the rearrangement rate is effectively large relative to the magnitude of negative natural selection. As another possible explanation, variation in arrangement of motifs might be selectively neutral or adaptive. To understand how fibroin genes have evolved, we need more detailed information on structural variations in fibroin genes within and between species. It is particularly important to clarify how the four types of motifs contribute to fibroin protein conformation.

Acknowledgments. We thank Dr. S. Hayasaka for the *A. pernyi* samples. We also appreciate two anonymous reviewers for helpful comments. We thank Dr. M.G. Goldsmith for a critical reading of the manuscript. This work was supported by the Enhancement of Center of Excellence, Special Coordination Funds for Promoting Science and Technology, Science and Technology Agency, Japan.

References

Caskey CT, Pizzuti A, Fu YH, Fenwick RG Jr, Nelson DL (1992) Triplet repeat mutations in human disease. Science 256:784–789

Craig CL (1997) Evolution of arthropod silks. In: Mittelr TE, Radovsky FJ, Resh VH (eds) Annual review of entomology vol. 42. Palo Alto, CA: Annual Reviews, pp 231–267

Fraser RDB, MacRae TP (1973) Conformation in fibrous proteins and related synthetic polypeptide. San Diego, CA: Academic Press

Guerette PA, Ginzinger DG, Weber BHF, Goslone JM (1996) Silk properties determined by gland-specific expression of a spider fibroin gene family. Science 272:112–114

Hibner BL, Burke WD, Eickbush TH (1991) Sequence identity in an early chorion multigene family is the result of localized gene conversion. Genetics 128:595–606

Hinman MB, Lewis RV (1992) Isolation of a clone encoding a second dragline silk fibroin. J Biol Chem 267:19320–19324

Hynes RO (1987) Integrins: a family of cell surface receptors. Cell 48:549–554

Jeffreys AJ, Wilson V, Thein SL (1985) Hypervariable "minisatellite" regions in human DNA. Nature 315:67–73

Kirimura J (1962) Studies on amino acid composition and chemical structure of silk protein by microbiological determination. Bull Sericul Exp Sta 17:447–522 (in Japanese)

Kumar S, Tamura K, Nei M (1993) MEGA: molecular evolutionary genetics analysis ver. 1.0. University Park, PA: Pennsylvania State University

Kyte J, Doolittle RF (1982) A simple method for displaying the hydrophathic character of a protein. J Mol Biol 157:105–132

Lam ST, Stahl MM, McMilin KD, Stahl FW (1974) Rec-mediated recombinational hot spot activity in bacteriophage lambda II: a mutation which causes hot spot activity. Genetics 77:425–433

Lucas F, Rudall KM (1968) Extracellular fibrous proteins: the silks. In: Florkin M, Stotz EH (eds) Comprehensive biochemistry, vol. 26B. Amsterdam: Elsevier, pp 475–558

Mita K, Ichimura S, James TC (1994) Highly repetitive structure and its organization of the silk fibroin gene. J Mol Evol 38:583–592

Nakamura T, Suyama A, Wada A (1991) Two types of linkage between codon usage and gene-expression levels. FEBS Lett 289:123–125

Paulsson G, Hoog C, Bernholm K, Wieslander L (1992) Balbiani Ring 1 gene in *Chironomus tentans* sequence organization and dynamics of a coding minisatellite. J Mol Biol 225:349–361

Rolfsmeier ML, Lahue RS (2000) Stabilizing effects of interruptions on trinucleotide repeat expansions in *Saccharomyces cerevisiae.* Mol Cell Biol 20:173–180

Sambrook J, Fritsch FE, Maniatis T (1989) Molecular cloning. New York: Cold Spring Harbor Laboratory Press

Starden R (1982) An interactive graphics program for comparing and aligning nuclear acid and amino acid sequences. Nucleic Acid Res 10:2951–2961

Suzuki Y, Tsuda M, Hirose S, Takiya S (1986) Transcription signals and factors of the silk genes. Adv Biophys 21:205–215

Tamura T, Inoue H, Suzuki Y (1987) The fibroin genes of the *Antheraea yamamai* and *Bombyx mori* are different in the core regions but reveal a striking sequences similarity in their 5′-ends and 5′-flanking regions. Mol Gen Genet 206:189–195

Tsujimoto Y, Suzuki Y (1979) The DNA sequence of *Bombyx mori* fibroin gene including the 5′ flanking, mRNA coding, entire intervening and fibroin protein coding regions. Cell 18:591–600

Xu M, Lewis RV (1990) Structure of a protein superfiber: spider dragline silk. Proc Natl Acad Sci USA 87:7120–7124

Yukuhiro K, Kanda T, Tamura T (1997) Preferential codon usage and two types of repetitive motifs in the fibroin gene of the Chinese oak silkworm, *Antheraea pernyi.* Insect Mol Biol 6:89–95