J Mol Evol (1994) 38:583–592

Journal of Molecular Evolution

Highly Repetitive Structure and Its Organization of the Silk Fibroin Gene

Kazuei Mita,[1] Sachiko Ichimura,[1] Tharappel C. James[2]

[1] Division of Biology, National Institute of Radiological Sciences, Anagawa, Chiba 263, Japan
[2] Department of Molecular Biology and Biochemistry, Wesleyan University, Middletown, Connecticut 06459-0175, USA

Received: 27 August 1992

Abstract. We have sequenced a number of cDNAs representing the *Bombyx mori* silk fibroin heavy chain transcript. These reveal that the central region of the fibroin gene is composed of alternate arrays of the crystalline element **a** and the noncrystalline element **b.** The core region is partitioned by a homogeneous nonrepetitive amorphous domain of around 100 bp in length. The element **a** is characterized by repeats of a highly conserved 18-bp sequence coding for perfect repeats of the unit peptide Gly-Ala-Gly-Ala-Gly-Ser. The element **b** is composed of repeats of a less-conserved 30-bp sequence which codes for a peptide similar to that in element **a** except in that (1) Ser is replaced by Tyr and (2) there are irregular substitutions of Ala to Val or Tyr. Therefore, the structure of the fibroin gene core consists of three-step higher-order periodicities. Heterogeneities in numbers of repeats are observed in each step of periodicity. Boundary sequence appeared in each periodicity to be quite homogeneous. Sequence analysis indicates that the unit sequences of elements **a** and **b** have homology to those of recombination hotspots reported in other genes and a recombination event may frequently occur between the misaligned sister chromatids, resulting in heterogeneities in repeat numbers and duplication or deletion of repetitive sequences. The repetitive superstructure of the fibroin gene may have been a result of continuous unequal crossovers in a primordial gene during evolution. A couple of important features of the fibroin protein were proved by the present nucleotide sequencing. The amino acid representation of the amorphous domain is vastly different from that of the repetitive regions. The carboxy-terminal nonrepetitive region has three Cys and nine (Arg + Lys) residues that may be responsible for complex formation with the fibroin light-chain molecule. The present DNA analysis also clearly demonstrates that the tRNA population in the posterior silk gland strictly complements the frequency of codons in the fibroin mRNA, which may help to achieve a highly efficient translation of fibroin mRNA.

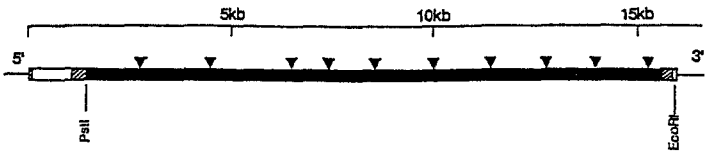
Key words: Silk fibroin gene — Unequal crossover — Three-tiered higher-order structure — Codon usage — Repetitive sequence

Correspondence to: K. Mita

Introduction

In a previous paper (Mita et al. 1988), we presented the sequence of a short fragment of *Bombyx mori* silk fibroin heavy-chain (H-chain) cDNA, showing that the fibroin coding region is composed of a reiterating bipartite unit sequence. Furthermore, the codon usage patterns are remarkably different in the two regions of the unit sequence and are preserved throughout the gene. Based on a partial restriction enzyme map of the fibroin gene, Gage and Manning (1980) postulated that the crystalline domains of the silk fibroin protein are partitioned by amorphous domains. Despite the very highly conserved nature of the "core" sequence of silk fibroin, they also found that alleles of the silk fibroin locus of 22 inbred stocks of *B. mori* differed in the size of the fibroin gene (Manning and Gage 1980). Such size differences could be accounted for by the length polymorphisms of the crystalline domains, which may have

(a)



(b)

```
1486 GGTGCTGCCGCTGGTTCTGGTGCGGGTGCCGGAGCTGGTTATGGAGCTGCTTCTGGTGCTGGTGCCGGTGCTGGGGCTGGTGCCGGAGCT
     G A A A G S G A G A G A G Y G A A S G A G A G A G A G A

1556 GGTTATGGAACTGGTGCAGGTGCAGGTGCCGGAGCTGGTTATGGAGCTGGTGCAGGTGCAGGTGCCGGAGCTGGTTATGGGGCTGGTGCA
     G Y G T G A G A G A G A G Y G A G A G A G A G A G Y G A G A

1646 GGTGCAGGTGCCGGAGCTGGTTATGGAGCTGGTGCAGGTGCAGGTGCCGGAGCTGGTTATGGGGCTGGTGCAGGTGCAGGTGCCGGAGCT
     G A G A G A G Y G A G A G A G A G A G A G Y G A G A G A G A G A

1736 GGTTATGGAGCTGGTGCGGGTGCCGGTGCCGGAGCTGGTTATGGAGCTGCCTCTGGTGCTGGTGCTGGCGCTGGGTACGGACAAGGAGTA
     G Y G A G A G A G A G A G Y G A A S G A G A G A G Y G Q G V

1826 GGAAGCGGAGCTGCTTCTGGAGCTGGTGCAGGTGCAGGAGCAGGTTCTGCCGCTGGTTCTGGGGCAGGTGCCGGTGCTGGTACCGGTGCT
     G S G A A S G A G A G A G A G S A A G S G A G A G A G T G A

1916 GGTGCAGGTTACGGAGCTGGTGCAGGTGCCGGTGCCGGAGCTGGTTATGGAGCTGCCTCTGGTACTGGAGCAGGTTATGGAGCTGGTGCC
     G A G Y G A G A G A G A G A G Y G A A S G T G A G Y G A G A

2006 GGAGCTGGTTACGGAGGTGCCTCTGGTGCTGGTGCTGGTGCCGGTGCTGGGGCTGGAGCCGGTTCTGGTTATGGAACTGGCGCTGGATAC
     G A G Y G G A S G A G A G A G A G A G A G S G Y G T G A G Y

2096 GGAGCAGGAGCCGGAGCAGGAGCCGGAGCAGGAGCTGGTGCTGGATACGGAGCAGGAGCTGGTGCTGGATACGGAGCAGGATATGGAGTA
     G A G A G A G A G A G A G A G A G Y G A G A G A G Y G A G Y G V

2186 GGAGCTGGTGCTGGATACGGAGCAGGATACGGAGCAGGAGCTGGAAGCGGAGCTGCCTCTGGTGCTGGTTCAGGTGCCGGTGCTGGTTCA
     G A G A G Y G A G Y G A G A G S G A A S G A G S G A G A G S

2276 GGTGCCGGTGCTGGTTCAGGTGCCGGTGCTGGTTCAGGTGCCGGTGCTGGTTCAGGTGCCGGTGCCGGTTCAGGTGCTGGTGCTGGTTCA
     G A G A G S G A G A G S G A G A G S G A G A G S G A G A G S

2366 GGTGCTGGTGCAGGCTCAGGTGCTGGTGCTGGTTCAGGTACTGGTGCTGGTTCAGGAGCTGGTGCTGGATACGGAGCAGGAGCTGGTGCT
     G A G A G S G A G A G S G T G A G S G A G A G Y G A G A G A

2456 GGATACGGAGCAGGAGCTGGTGCTGGATACGGAGCAGGAGCTGGTGCTGGATACGGAGCAGGAGCTGGCGTTGGATACGGAGCAGGAGCT
     G Y G A G A G A G Y G A G A G A G Y G A G A G V G Y G A G A

2546 GGAAGCGGAGCTGCCTCTGGTGCTGGTGCTGGTTCAGGTGCCGGTGCTGGTTCAGGTGCTGGTGCTGGTTCAGGTGCTGGTGCTGGTTCA
     G S G A A S G A G A G S G A G A G S G A G A G S G A G A G S

2636 GGTGCTGGTGCTGGTTCAGGTGCCGGTGCTGGTTCAGGTGCTGGTGCTGGTTCAGGAGCTGGTGCTGGTTCAGGTGCTGGTGCTGGTTCA
     G A G A G S G A G A G S G A G A G S G A G A G S G A G A G S

2726 GGAGCTGGAGTTGGATACGGAGCAGGAGTTGGTGCTGGATACGGAGCAGGATATGGAGCAGGAGCTGGTGCTGGATACGGAGCAGGAGCA
     G A G V G Y G A G V G A G Y G A G Y G A G A G A G Y G A G A

2816 GGAAGCGGAGCTGCCTCTGGTGCTGGTGCCGGTGCTGGAGCTGGTGCAGGAACAGGCTCTTCTGGATTTGGACCATATGTAGCAAATGAC
     G S G A A S G A G A G A G A G A G T G S S G F G P Y V A N D

2906 GGATATAGCAGAAGTGATGGCTACGAATACGCT
     G Y S R S D G Y E Y A
```

Fig. 1. Nucleotide and predicted amino acid sequences of the 5′ end and 3′ end cDNA clones of the silk fibroin heavy chain gene of *Bombyx mori.* **a** Schema of silk fibroin gene. This is a rendering of a restriction map of Gage and Manning (1980) showing the repetitive (*solid*) and the nonrepetitive (*hatched*) protein coding domains as well as intron (*open*). The *filled triangles* represent the locations of the amorphous domains (the positions of the first and the last two amorphous domains from the 5′ end are characterized by this work, while others are arbitrary); **b** the nucleotide sequence of the 5′ end repetitive coding region. *Numbering* is based on the partial sequence of the 5′ end region as shown by Tsujimoto and Suzuki (1979); **c** the nucleotide sequence of the 3′ end cDNA clone. The amorphous domains are *underlined* and the putative polyadenylation signal is *boxed.*

resulted from homologous unequal crossovers between the highly repetitive coding sequence of misaligned genes (Manning and Gage 1980).

It is believed that many repetitive genes, such as the rDNA (Petes 1980; Szostak and Wu 1980), the histone (Crawford et al. 1979; Jackson and Fink 1985), the human involucrin (Eckert and Green 1986), the mouse major histocompatibility complex (Steinmetz et al. 1986), the *Chironomus tentans* Balbiani Ring (Hoog et al. 1988), and the *B. mori* silk fibroin as well as the minisatellite DNAs (Jeffreys et al. 1985) evolved through unequal crossovers. While these repetitive sequences in long tandem arrays show surprisingly little sequence variation, one observes substantial allelic variations in the number of the repeat units. In the case of the human involucrin (Eckert and Green 1986) and the *Chironomus tentans* Balbiani Ring (Hoog et al. 1988) genes, different blocks, each comprising a repeat unit variant, are arranged in a regular array. Furthermore, it is widely accepted that the unequal crossing over resulting in duplication and/or deletion may be responsible for the molecular pathology of some inherited disorders in human (Collier et al. 1989; Sinnott et al. 1990). Although experimental verification for the involvement of unequal crossovers in the evolution of rDNA genes in yeast has been obtained (Petes 1980; Szostak and Wu 1980), direct evidence for such events in other higher organisms has not been so forthcoming.

(c)

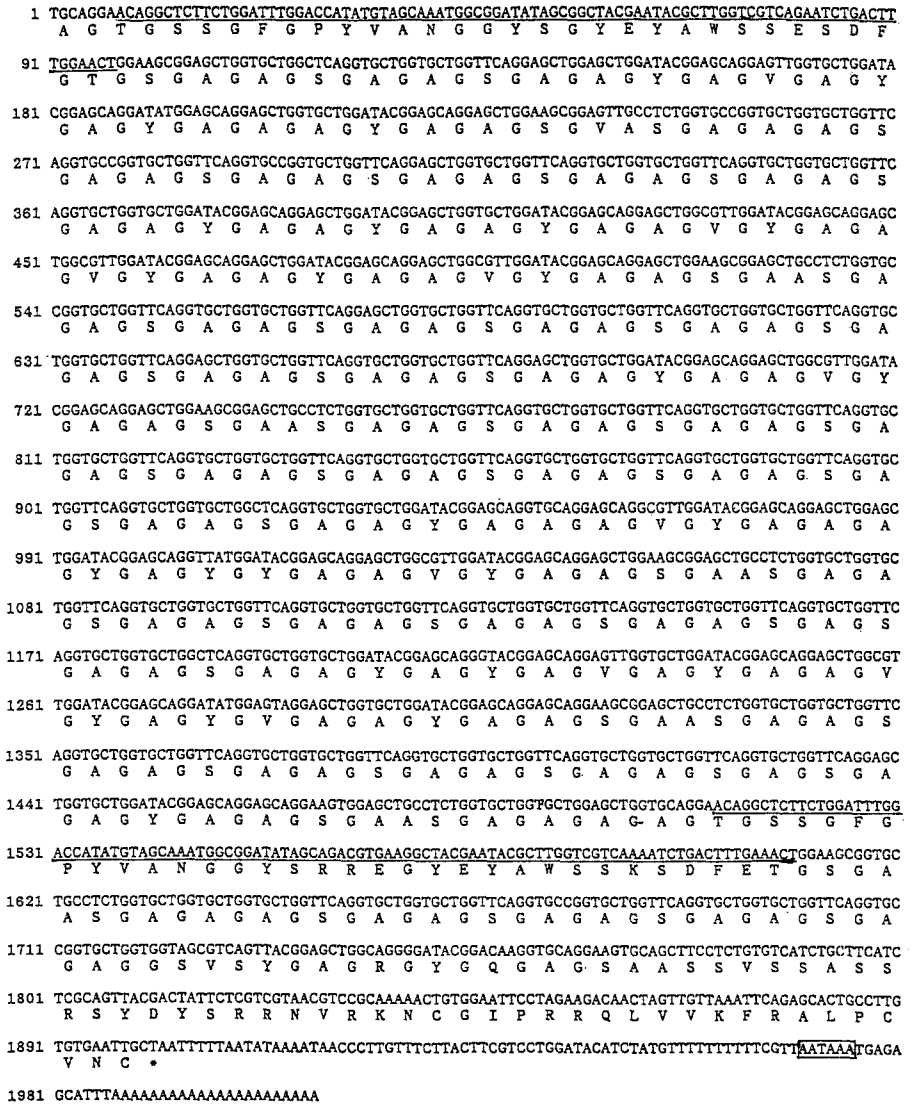


Fig. 1. Continued.

In the present work, we have sequenced the 5′ and 3′ ends of the core repetitive region of a number of silk fibroin H-chain cDNAs. The sequence analysis reveals that the repetitive region of fibroin has a three-tiered higher-order structure with significant heterogeneities in the repeat number. It is composed of alternate arrays of two domains, **a** and **b.** These alternate arrays are interspersed every 1–2 kb with a third amorphous domain. The domain **a** has the repeat of a highly conserved 18-base unit sequence encoding the hexapeptide Gly-Ala-Gly-Ala-Gly-Ser and corresponds to the crystalline region. The domain **b,** which corresponds to the non-crystalline region, is less conserved both in the repeat sequence and in the unit length. The nucleotide sequence of amorphous domain is also highly conserved; and its deduced amino acid sequence is completely different from that of the repetitive regions. Furthermore, the DNA sequence analysis of the various domains also shows the evolution of a highly specific codon usage pattern that is preserved all through the repetitive coding regions.

Materials and Methods

Materials. The F1 hybrids of *B. mori* were derived from the cross between Kinshu × Showa strains and were cultured on mulberry leaves. The M13 mp18 and mp19 cloning vectors were purchased from Takara Shuzo Co. For sequencing, we have used the 7-DEAZA Sequencing Kit and Deletion Kit, which were also obtained from Takara Shuzo Co. The cDNA Synthesis Kit, including AMV reverse transcriptase, RNase H, and *E. coli* DNA polymerase I, was purchased from Amersham, Inc. Restriction enzymes used were purchased from Boehringer-Mannheim and oligo(dT)-latex was provided by Japan Roche.

Isolation of Poly(A)$^+$RNA. Total RNA from the posterior silk gland (PSG) of 5th-instar, 3rd-day silkworm (in which fibroin gene

is actively transcribed) was extracted by the LiCl/urea procedure (Applebaum et al. 1981) and described in detail in the previous paper (Mita et al. 1988). The total PSG RNA was used to isolate poly(A)-containing RNA by selective binding to oligo(dT)-latex in 10 mM Tris-HCl (pH 7.5), 1 mM EDTA, 0.5 M NaCl, 0.1% SDS. Following a brief washing with the elution buffer (10 mM Tris-HCl, pH 7.5, 1 mM EDTA, 0.1% SDS) at room temperature, poly(A)$^+$RNA from oligo(dT)-latex was eluted by a 5-min incubation in elution buffer at 65°C.

cDNA Synthesis and Cloning. Single-stranded cDNA was synthesized using AMV reverse transcriptase and oligo(dT) primer. The second DNA strand was synthesized according to Gubler and Hoffman (1983) using RNase H and *E. coli* DNA polymerase I, followed by ligation to *Sma*I-cut M13 mp18 or mp19 vectors. These recombinant DNAs were used to transform *E. coli* JM105 cells. The recombinant plaques were grown individually and the phage DNAs were isolated. These DNAs were resolved on agarose gels by electrophoresis and transferred onto a GeneScreen filter (NEN Research Products). The filters were hybridized with ^{32}P-labeled cloned silk fibroin cDNA (Mita et al. 1988).

Synthesis of cDNAs Containing the Amorphous Domain. The sequence of the 3′ end of the amorphous domain had already been determined from the sequence of the 3′ end cDNA clone (Fig. 1c) in the present work. We reverse-transcribed the cDNAs containing the amorphous sequence from PSG poly(A)$^+$RNA using synthetic 17-mer primer complementary to the 3′ end of the amorphous domain. The 17-mer was synthesized by DNA synthesizer (model 380B, Applied Biosystems). These cDNAs were processed as described above.

DNA Sequencing. The nucleotide sequence was determined by the dideoxynucleotide chain-termination method (Sanger et al. 1977; Mizusawa et al. 1986). As the cDNA clones were 1.5–2 kb long and appropriate restriction sites were not available for subcloning, the deletion method was used to complete sequencing (Henikoff 1984; Yanisch-Perron et al. 1985).

Results

DNA Sequencing of Silk Fibroin Gene

One of the *B. mori* silk fibroin H-chain cDNA clones synthesized using the synthetic amorphous region primer has a sequence at the 5′ end that is completely identical to that reported by Tsujimoto and Suzuki (1979). Therefore, we believe that this clone represents the 5′ end of the fibroin gene "core." Figure 1b shows the nucleotide sequence of this clone beginning from the 5′ end to the first amorphous domain. The region 2,090–2,866 represents a typical repetitive coding region described previously (Mita et al. 1988) and consists of alternate array of two elements, **a** and **b,** corresponding to repetitive and joining components, respectively. The element **a** is characterized by repeats of highly conserved 18-base-unit sequence, GGTGCTGGTGCTGGTTCA, which encodes the hexapeptide Gly-Ala-Gly-Ala-Gly-Ser, a characteristic feature of fibroin. The element **b** is composed of repeats of a 30-base-unit sequence, GGTGCTGGATACGGAGCAGGAGCTGGCGTT, that is occasionally partially deleted in actual unit sequence (Ichimura and Mita 1992). Heterogeneities in repeat number are observed in both elements. The region 1,562–1,777 (Fig. 1b) is composed of six repeats of 36-base-unit sequence, GGAGCTGGTGCAGGTGCAGGTGCCGGAGCTGGTTAT.

We have cloned a cDNA encompassing about 2 kb of silk fibroin H-chain mRNA upstream from the 3′ end poly(A) tail. Figure 1c shows the nucleotide sequence of this cDNA clone. This clone also has two sets of amorphous domains each about 100 bases long. Each sequence is quite similar except for a couple of point mutations and a nine-nucleotide deletion within the amorphous domain. The region between the amorphous domains is composed of the reiterations of elements **a** and **b.** The numbers of repeat units in element **a** may vary from two to more than 10 throughout the fibroin coding region. (See also Figs. 1b and 1c.) The unit sequence of element **a** is highly homogeneous while that of element **b** is very heterogeneous throughout the 3′ and 5′ regions of the fibroin H-chain gene. Interestingly, the boundary from element **b** to **a** is defined by the 18-base sequence, GGAAGCGGAGCTGCCTCT.

Based on restriction mapping, Gage and Manning (1980) have shown that the 15-kb-long coding sequence is divided into about ten large repetitive coding domains separated by small, nonrepetitive, amorphous coding domains. At the restriction-enzyme-map level, this translates to a set of *Mbo*II-*Ava*II-*Hin*fI restriction sites within the 60–110 bases in the amorphous domain, repeating every 1–1.5 kb. Our sequence analysis confirms the above finding at the DNA sequence level. The 3′ end cDNA clone in Fig. 1c includes two sets of similar amorphous domains. The amorphous domain closer to the 5′ end in this clone has the characteristic *Mbo*II-*Ava*II-*Hin*fI sites in that order. A 'G-to-A' transition at the *Hin*fI site in the amorphous domain closer to the 3′ end causes it to lose that site. Gage and Manning (1980) had also pointed out the missing *Hin*fI site in the 5′ (first) and 3′ end (last) amorphous domain. To decipher the sequence of the central area of the fibroin gene "core," we have sequenced 22 different cDNA clones of fibroin H-chain mRNA. These were primed by a synthetic oligomer complementary to the 3′ end of amorphous domain. Figure 2 shows the nucleotide sequences of 10 different amorphous domains. Sequence analysis of all the 10 amorphous domains and their corresponding 5′ flanking regions indicates that they do not represent duplicate sequences of one or more clones. Based on their partial restriction map, Gage and Manning (1980) estimated that there are 10 amorphous domains dispersed in the fibroin gene "core." Thus the sequence of 10 amorphous domains presented here may account for all the amorphous domains found in the gene. From these studies we conclude that the DNA sequence of the amorphous domains is quite homogeneous throughout the fibroin H-chain gene with two possible exceptions. They are the amorphous domains

Fig. 2. The nucleotide sequences of the amorphous domains. The 17-mer nucleotide complementary to the 3′ end amorphous domain as indicated in the figure was synthesized by DNA synthesizer. More than 20 cDNA clones prepared using this primer were sequenced and sequences of 10 different amorphous domains are summarized. The top-panel amorphous sequence is found to be the first amorphous domain of the 5′ end. (Its complete sequence is presented in Fig. 1b.) The second one from the bottom corresponds to the second amorphous domain from the 3′ end of the gene, while the bottom one is the last amorphous domain of the fibroin gene. (Its complete sequence is included in the 3′ end cDNA clone in Fig. 1c.) A pair of direct repeats are *underlined.* A *dot* represents the same base as in the consensus.

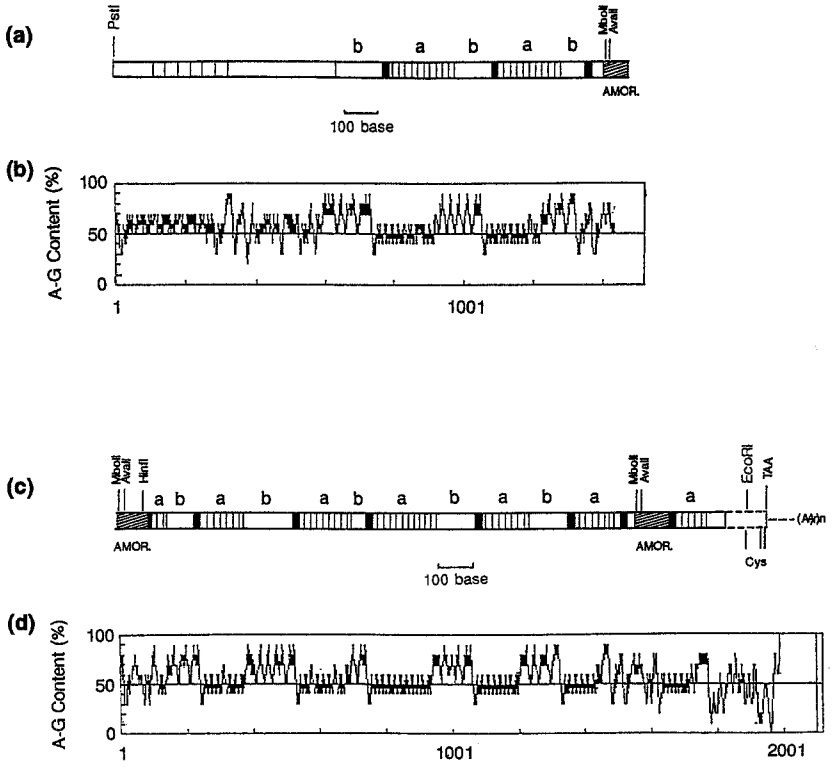


Fig. 3. Schematic representation of the 5′ and 3′ end cDNA clones. **a** The schematic diagram showing the 5′ end clone. **b** The A–G content (purine content, %) of the 5′ end clone. Average span of sequence is 10. **c** Similar to **a** above but of the 3′ end clone. **d** Similar to **b** above but of the 3′ end clone. The notations **a** and **b** in panels **a** and **c** denote the elements **a** and **b,** respectively (see text for details), and the *closed boxes* represent the 18-base boundary sequence GGAAGCGGAGCTGCCTCT. Other significant features are as marked in the panels.

at the 5′ (first) and 3′ (last) end where a 9-mer sequence has been inserted in the middle of the domain. The other conclusions that can be drawn from the sequencing data are (1) that the interamorphous areas are composed of the alternate arrays of element **a** and **b** as shown in Fig. 3c, (2) that the carboxy terminal 43 amino acids do not show the Gly-X dicodon repetition found in the repetitive core sequence, and (3) that the *Eco*RI site first identical by the mapping analysis of Ohshima and Suzuki (1977) is present 50 nucleotide upstream of the stop codon (Fig. 1c).

Higher-Order Repetitive Structure with Conserved Boundary Sequences

The present sequences reveal that the fibroin gene core possesses three-tiered higher-order structure as schematically shown in Fig. 4. It is noteworthy to mention that

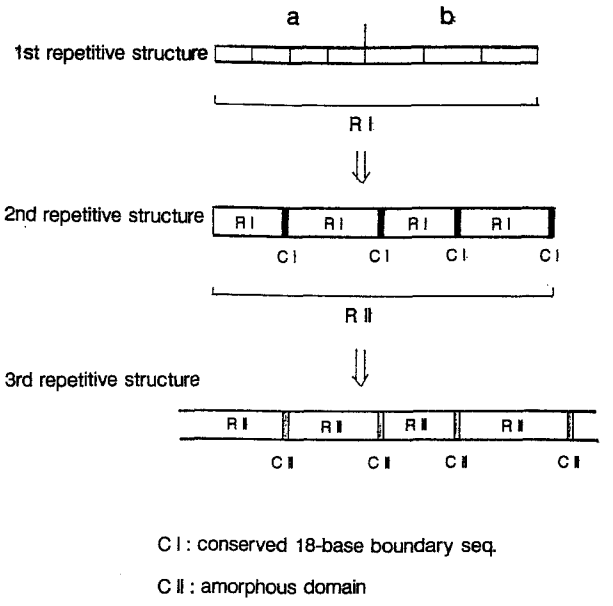


Fig. 4. Superstructure of silk fibroin gene core. *CI:* conserved 18-base boundary sequence; *CII:* amorphous domain; the notations **a** and **b** refer to elements **a** and **b,** respectively.

the boundary sequences (CI and CII in Fig. 4) are highly conserved. In actual fibroin gene core, heterogeneity in repeat number is observed in each step of periodicity. Manning and Gage (1980) postulated that the homologous unequal crossovers between the crystalline coding domains played a predominant role in the evolution of the fibroin gene from their *Hha*I and *Mbo*II cleavage patterns of silk fibroin locus from 22 inbred stock *B. mori.* In mouse immunoglobulin, switch recombination takes place in the S region containing frequent short-sequence motifs like GAGCT (Nikaido et al. 1982; Iwasato et al. 1990). The sequence is also found in element **b.** In addition, the *Chi* sequence (Dover 1989), a recombination signal in *E. coli,* appears to be similar to the 18-base unit sequence in element **a.** *Chi*-like sequences are also found near the recombination hotspots in the minisatellite DNAs (Jeffreys et al. 1985), the Balbiani Ring genes (Hoog et al. 1988), and the major histocompatibility complex genes (Steinmetz et al. 1986). These sequences could promote unequal crossing over at repetitive regions of the fibroin gene to generate new length polymorphisms and heterogeneities.

It has been suggested that every unequal sister chromatid crossover produces four recombination molecules each of which harbors either a deletion or a tandem duplication of the stretch of base pairs lying between the points of the crossover in the parent molecules (Smith 1976). To further our understanding of the evolution of fibroin genes, we have also partially sequenced the repetitive coding region downstream of the *Pst*I site of the silkworm *B. mandarina,* which is considered to be an ancestor of *B. mori* (Kusuda et al. 1986), and compared the sequence with that of *B. mori.* The results of this analysis shown in Fig. 5 suggest that there is an insertion of 144 bases at a position between 1,565 and 1,708 of the *B. mori* fibroin gene. The inserted sequence consists of four tandem repeats of a 36-mer sequence demonstrating a tandem duplication by an unequal crossover. Interestingly, around the putative recombination sites, signature sequences such as the GAGCT and the *Chi*-like sequences are also found.

The Amino Acid Sequence of Silk Fibroin

The amino acid sequences of all regions of fibroin can be deduced from the nucleotide sequences. Element **a** has perfect repeats of the unit Gly-Ala-Gly-Ala-Gly-Ser sequence, a well-known repetitive unit of the crystalline region of fibroin (Lucas et al. 1958). The unit repeat sequence of **b** is very much like that in element **a,** except that the Ser residue is replaced mostly by a Tyr residue. Often some variability in length of the repeat unit and the substitutions of Ala to either Val or Tyr is also observed (Fig. 6). Presumably, the bulkiness of the Tyr residues and the substantially irregular position of Tyr in element **b** would promote the disordered noncrystalline feature in specific domains of fibroin.

The amino acid sequence of the amorphous domains is vastly different from that of repetitive regions, notably for the presence of amino acids such as Asn, Asp, Glu, Lys, Phe, Pro, Thr, and Trp that are not found in the repetitive regions. Also, the amorphous domain and the downstream nonrepetitive coding regions at the 3′ end show the presence of a number of hydrophilic amino acids remarkably distinct from the rest of the fibroin protein. There are seven Arg and two Lys residues in the nonrepetitive 3′ end region. These basic residues may promote the complex formation with fibroin L-chain molecule through ionic interaction. The fact that fibroin L-chain has a preponderance of acidic amino acids (Shimura et al. 1976) supports this view. Furthermore, there are three cysteine residues located near the C-terminus of the fibroin H-chain. It is likely that these residues are involved in forming a complex with the fibroin L-chain through disulphide bonds (Ohmachi et al. 1982). The cluster of basic residues in this region would inhibit the formation of disulphide bonds within or between fibroin H-chain molecules.

Specific Codon Usage

We have estimated the overall codon choices for Gly, Ala, and Ser among synonymous codons by assuming that all interamorphous regions are composed of alternate reiteration of elements **a** and **b** in the same **a/b** ratio observed in the 3′ end cDNA clone (Table 1). Suzuki and Brown (1972) estimated that 22% and 16% of all codons in the fibroin mRNA are the glycine codons

Fig. 5. Comparison of nucleotide sequence of the 5′ end repetitive coding region of the silk fibroin gene between *B. mori* and *B. mandarina*. Kusuda et al. (1986) have cloned a part of the silk fibroin gene of *B. mandarina* from a genomic library that covers the 5′ end non-repetitive coding and intron regions. In this work, we obtained the *B. mandarina* clone from Dr. Kusuda and have sequenced a part of its repetitive region starting from the *Pst*I site.

GGU and GGA, respectively. These values agree with our DNA sequence analysis data. The existence of specific isoacceptor tRNA populations has been reported in the PSG (Chevallier and Garel 1979; Hentzen et al. 1981). The $tRNA_1^{Gly}$ carrying the GCC anticodon recognizes GGU > GGA codons, whereas the $tRNA_2^{Gly}$ of chmU-C-C anticodon recognizes GGA and GGG (Kawakami et al. 1979). Chevallier and Garel (1979) compared the $tRNA_1^{Gly}$/$tRNA_2^{Gly}$ levels during the start of the 5th instar and the 4th day of the 5th instar and obtained a ratio within the range of 1.6–2.0. From our data, the ratio of GGU/GGA codons in the fibroin H-chain mRNA is estimated to be 1.43 (Table 1). These two estimates are in good agreement, suggesting that there is proportional representation of tRNA population in PSG, relative to the frequencies of the corresponding codons. In another study, Hentzen et al. (1981) extracted $tRNA^{Ser}$ isoacceptors from the PSG of early and late 5th-instar larvae and fractionated them using DEAE-cellulose chromatography. They found that the $tRNA_2^{Ser}$ group is the predominant form in the PSG (82% during the secretion phase). It is known that the isoacceptor $tRNA_1^{Ser}$ is capable of translating AGU and AGC codons, whereas the other three types of $tRNA_2^{Ser}$, $tRNA_{2a}^{Ser}$, $tRNA_{2b}^{Ser}$, and $tRNA_{2c}^{Ser}$ are able to decode UCA, UCU, UCC, and UCG codons, respectively (Garel et al. 1976; Chavancy et al. 1979). Our sequence analysis reveals that the total fraction of UCA, UCC, UCG, and UCU codons in fibroin H-chain mRNA is 84% and is in excellent agreement with the ratio of the $tRNA_2^{Ser}$/$tRNA_1^{Ser}$ isoacceptors in the PSG of *B. mori*. Thus, the tRNA population in the PSG strictly complements the frequency of codons in the fibroin mRNA, and this may help to achieve a highly efficient translation of fibroin mRNA.

Discussion

The detailed organization of the fibroin gene core tempts us to speculations concerning the evolutionary process: In the first step, multiple duplications of [GG(A/U)GC(A/U)] may produce primordial repeats including an 18-base boundary sequence since it has been proposed that oligonucleotide repeats are the primordial source of all genes (Ohno 1984). Mutations occurred in one repeat followed by duplication to make element **a.** Other mutations in a repeat adjacent to element **a** lead to element **b** with a trace of primordial repeat at 3′ end—that is, an 18-base boundary sequence CI (Fig. 4). After the repetitive block RI is duplicated to form an alternate array of element **a** and **b** by unequal crossover, an amorphous sequence is inserted, probably at the CI sequence, by transposition, resulting in the direct repeat of a part of the CI sequence before and after the amorphous domain (Fig. 2). The CI sequence may be a recombination signal or target sequence. Interestingly, the CI sequence has a characteristic primary structure: the former half of the sequence is composed of purine stretch while the latter half is pyrimidine stretch. The 5′ end of the CI se-

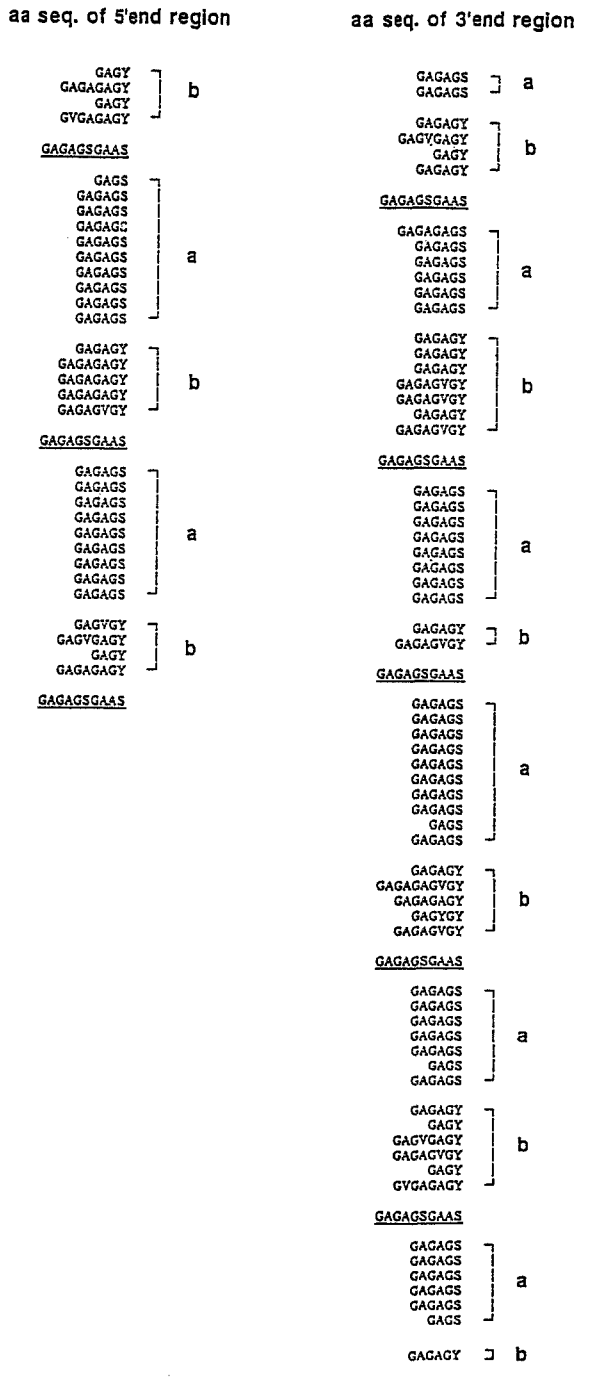


Fig. 6. Amino acid sequences of repetitive regions of the 5′ end and 3′ end cDNA clones. The amino acid (aa) sequences shown in the figure are restricted to the repetitive regions. The 5′ end aa sequence corresponds 2,084–2,833 of DNA sequence of the 5′ end cDNA clone in Fig. 1b, while the 3′ end aa sequence is derived from nucleotide number 104–1,451 of the 3′ end cDNA clone in Fig. 1c; the notations **a** and **b** refer to elements **a** and **b**, respectively. The *underlined sequence* was translated from the 18-base boundary sequence.

quence always links to a purine-rich region, element **b**, whereas the 3′ end is followed by rather a pyrimidine-rich sequence, element **a**. It is likely that the CI boundary sequence plays a role in switching the recombination site from purine-rich region to pyrimidine-rich sequence. The center of the CI sequence is GAGCT, which is the short-motif sequence of mouse immunoglobulin (Nikaido et al. 1982; Iwasato et al. 1990). Some recombination enzyme may attack AGCT, a specific 4-bp symmetrical sequence, just as restriction enzymes do. In the final stage, RII block including amorphous domain at its 3′ end is duplicated by unequal crossover events to generate the three-tiered higher-order periodicity.

A close analysis of the organization of the fibroin H-chain gene also reveals that in general all the amorphous domains, CII, preserve a certain sequence homology. However, the two outside amorphous domains with the insertion of 9-mer sequence and lack of *Hin*fI site are structurally distinguishable from the inside ones. This fact leads us to speculate on the order of duplication of RII in Fig. 4. First both outside amorphous domains are duplicated from a primordial sequence by an unequal crossover event. Subsequently, the inside amorphous ones were duplicated from either one of the outside ones. Judging from the deletion/insertion of the 6-mer sequence and a point mutation that appears in the 5′-flanking regions of the amorphous domains (Fig. 2), the first CII may be the source of the inside ones. Similar situations have been reported for human involucrin gene (Eckert and Green 1986) and Balbiani Ring genes (Hoog et al. 1988). It was also reported that the "aberrant" repeats are often observed toward the ends of tandem arrays (McCutchan et al. 1982; Eckert and Green 1986; Hoog et al. 1988) as well as in the silk fibroin gene (bases 1,562–1,777 in Fig. 1b). This agrees well with a previous theoretical evaluation of the evolution of repetitive genes (Smith 1976).

Hatfield et al. (1982) measured isoacceptor tRNAs of human reticulocytes by reverse-phase chromatography and showed that the occurrence of codons in globin mRNA is correlated with the codon recognition properties of the isoacceptor tRNAs. Quenzar et al. (1988) studied the relationship between the isoaccepting tRNA distribution and codon bias in tendon collagen of chicken. They noted that there is a relative increase in the levels of $tRNA_{IGG}^{Pro}$ (100%) and $tRNA_{GCC}^{Gly}$ (40%) in tendon as compared to other tissues, indicating a specialization of the tRNA population for collagen synthesis. However, no evidence suggesting tissue-specific coadaptation of codon usage and tRNA pools in three mouse actin mRNAs was found (Alonso et al. 1986). Thus the quantitative correlation shown here in *B. mori* between the mRNA may not be generalized to other eukaryotes, or it may be applicable to cases of enormous tissue-specific gene expression.

Acknowledgments. We thank Dr. Jun Kusuda, National Institute of Health, Japan, for his gift of a fibroin gene clone of *Bombyx mandarina*. This work was supported by a grant-in-aid for scientific research from the Ministry of Education, Science and Culture of Japan.

Table 1. Codon usage patterns for different regions and whole coding region of silk fibroin gene

Codon	Element **a**	Element **b**	5′-coding 1–1,447[a]	3′-coding nonrepeated[b]	Amorphous	Total[c]
Gly						
GGA	3	91	5	5	38	981
GGC	3	7	0	1	29	127
GGG	0	1	3	0	0	17
GGU	121	18	2	2	0	1406
Ala						
GCA	0	28	5	2	10	302
GCC	5	7	0	0	0	131
GCG	0	0	1	0	0	3
GCU	79	43	5	3	10	1237
Ser						
AGC	0	7	4	1	10	85
AGU	0	1	5	3	2	20
UCA	43	0	1	2	10	433
UCC	0	0	3	1	0	4
UCG	0	0	0	0	10	10
UCU	0	7	2	4	30	108

[a] Sequence from Tsujimoto and Suzuki (1979)

[b] Nonrepeated region is 1,721–1,900 of the 3′ end cDNA clone in Fig. 1c

[c] The calculation is based on the following assumptions: the length of total coding region is 15 kb, and the interamorphous regions are composed of reiteration of elements **a** and **b** with the same codon usage patterns found in the 3′ end cDNA clone

References

Alonso S, Minty A, Bourlet Y, Buckingham M (1986) Comparison of three actin-coding sequences in mouse, evolutionary relationships between the actin genes of warm-blooded vertebrates. J Mol Evol 23:11–22

Applebaum SW, James TC, Wreschner DH, Tata JR (1981) The preparation and characterization of locust vitellogenin messenger RNA and the synthesis of its complementary DNA. Biochem J 193:209–216

Chavancy G, Chevallier A, Fournier A, Garel JP (1979) Adaptation of iso-tRNA concentration to mRNA codon frequency in the eukaryote cell. Biochimie 61:71–78

Chevallier A, Garel JP (1979) Studies on tRNA adaptation, tRNA turnover, precursor tRNA and tRNA gene distribution in *Bombyx mori* by using two-dimensional polyacrylamide gel electrophoresis. Biochimie 61:245–262

Collier S, Sinnott PJ, Dyer PA, Price DA, Harris R, Strachan T (1989) Pulsed field gel electrophoresis identifies a high degree of variability in the number of tandem 21-hydroxylase and complement C4 gene repeats in 21-hydroxylase deficiency haplotypes. EMBO J 8:1393–1402

Crawford RV, Krieg P, Harvey RP, Hewish DA, Wells JRE (1979) Histone genes are clustered with a 15-kilobase repeat in the chicken genome. Nature 279:132–136

Dover GA (1989) Victims or perpetrators of DNA turnover? Nature 342:347–348

Eckert RL, Green H (1986) Structure and evolution of the human involucrin gene. Cell 46:583–589

Gage LP, Manning RF (1980) Internal structure of the silk fibroin gene of *Bombyx mori*. I. J Biol Chem 255:9444–9450

Garel JP, Hentzen D, Schlegel M, Dirheimer G (1976) Structural studies on RNA from *Bombyx mori L.* I. Biochimie 58:1089–1100

Gubler U, Hoffman BJ (1983) A simple and very efficient method for generating cDNA libraries. Gene 25:263–269

Hatfield D, Varricchio F, Rice M, Forget BG (1982) The aminoacyl-tRNA population of human reticulocytes. J Biol Chem 257:3183–3188

Henikoff S (1984) Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. Gene 28:351–359

Hentzen D, Chevallier A, Garel JP (1981) Differential usage of iso-accepting $tRNA^{Ser}$ species in silk glands of *Bombyx mori*. Nature 290:267–269

Hoog C, Daneholt B, Wieslander L (1988) Terminal repeats in long repeat arrays are likely to reflect the early evolution of Balbiani Ring genes. J Mol Biol 200:655–664

Ichimura S, Mita K (1992) Essential role of duplications of short motif sequences in the genomic evolution of *Bombyx mori*. J Mol Evol 35:123–130

Iwasato T, Shimizu A, Honjo T, Yamagishi H (1990) Circular DNA excised by immunoglobulin class switch recombination. Cell 62:143–149

Jackson JA, Fink GA (1985) Meiotic recombination between duplicated genetic elements in *Saccharomyces cerevisiae*. Genetics 109:303–332

Jeffreys AJ, Wilson V, Thein SL (1985) Hypervariable 'minisatellite' regions in human DNA. Nature 314:67–73

Kawakami M, Nishio K, Takemura S, Kondo T, Goto T (1979) 5-Carboxy-hydromethyluridine, a new modified nucleoside located in the anticodon of $tRNA_2^{Gly}$ from the posterior silk glands of *Bombyx mori*. Nucleic Acids Res Symp Ser 00:s53–s56

Kusuda J, Tazima Y, Onimaru K, Ninaki O, Suzuki Y (1986) The sequence around the 5′ end of the fibroin gene from wild silkworm, *Bombyx mandarina*, and comparison with that of the domesticated species, *B. mori*. Mol Gen Genet 203:359–364

Lucas F, Shaw JTB, Smith SG (1958) The silk fibroins. Adv Protein Chem 13:107–242

Manning RF, Gage LP (1980) Internal structure of the silk fibroin gene of *Bombyx mori*. II. J Biol Chem 255:9451–9457

McCutchan T, Hsu H, Thayer RE, Singer M (1982) Organization of African green monkey DNA at junctions between α-satellite and other DNA sequences. J Mol Biol 157:195–211

Mita K, Ichimura S, Zama M, James TC (1988) Specific codon us-

age pattern and its implications on the secondary structure of silk fibroin mRNA. J Mol Biol 203:917–925

Mizusawa S, Nishimura S, Seela F (1986) Improvement of the dideoxy chain termination method of DNA sequencing by use of deoxy-7-deaza guanosine triphosphate in place of dGTP. Nucleic Acids Res 14:1319–1324

Nikaido T, Yamawaki-Kataoka Y, Honjo T (1982) Nucleotide sequences of switch regions of immunoglobulin C_ε and C_γ genes and their comparison. J Biol Chem 257:7322–7329

Ohmachi T, Nagayama H, Shimura K (1982) The isolation of a messenger RNA coding for the small subunit of fibroin from the posterior silk gland of the silkworm. FEBS Lett 146:385–388

Ohno S (1984) Repeats of base oligomers as the primordial coding sequences of the primeval earth and their vestigates in modern genes. J Mol Evol 20:313–321

Ohshima Y, Suzuki Y (1977) Cloning of the silk fibroin gene and its flanking sequences. Proc Natl Acad Sci USA 74:5363–5367

Petes TD (1980) Unequal meiotic recombination within tandem arrays of yeast ribosomal DNA genes. Cell 19:765–774

Quenzar B, Agoutin B, Reinisch F, Weill D, Perin F, Keith G, Neyman T (1988) Distribution of isoaccepting tRNAs and codons for proline and glycine in collagenous and noncollagenous chicken tissues. Biochem Biophys Res Commun 150:148–155

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467

Shimura K, Kikuchi A, Ohtomo K, Katagawa Y, Hyodo A (1976) Studies on silk fibroin of *Bombyx mori*. I. J Biochem 80:693–702

Sinnott P, Collier S, Costigan C, Dyer PA, Harris R, Strachan T (1990) Genesis by meiotic unequal crossover of a *de novo* deletion that contributes to steroid 21-hydroxylase deficiency. Proc Natl Acad Sci USA 87:2107–2111

Smith GP (1976) Evolution of repeated DNA sequences by unequal crossover. Science 191:528–535

Steinmetz M, Stephan D, Lindahl KF (1986) Gene organization and recombinational hotspots in the murine major histocompatibility complex. Cell 44:895–904

Suzuki Y, Brown DD (1972) The genes for fibroin in *Bombyx mori*. J Mol Biol 63:409–429

Szostak JW, Wu R (1980) Unequal crossing over in the ribosomal DNA of *Saccharomyces cerevisiae*. Nature 284:426–430

Yanisch-Perron C, Vieira J, Messing J (1985) Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13 mp18 and pUC19 vectors. Gene 33:103–119