ELSEVIER

Comparative Biochemistry and Physiology Part B 133 (2002) 493–507

CBP

www.elsevier.com/locate/cbpb

Review

Comparative architecture of silks, fibrous proteins and their encoding genes in insects and spiders☆

Catherine L. Craig[a,b,c,d,*], Christian Riekel[e]

[a]*Museum of Comparative Zoology, Harvard University, Cambridge, MA, USA*
[b]*Department of Biology, Tufts University, Medford, MA, USA*
[c]*Department of Chemical Engineering, Tufts University, Medford, MA, USA*
[d]*Biotechnology Center, Tufts University, Medford, MA, USA*
[e]*European Synchrotron Radiation Facility, Grenoble, France*

Received 31 December 2001; received in revised form 23 April 2002; accepted 16 May 2002

Abstract

The known silk fibroins and fibrous glues are thought to be encoded by members of the same gene family. All silk fibroins sequenced to date contain regions of long-range order (crystalline regions) and/or short-range order (non-crystalline regions). All of the sequenced fibroin silks (*Flag* or silk from flagelliform gland in spiders; *Fhc* or heavy chain fibroin silks produced by Lepidoptera larvae) are made up of hierarchically organized, repetitive arrays of amino acids. *Fhc* fibroin genes are characterized by a similar molecular genetic architecture of two exons and one intron, but the organization and size of these units differs. The *Flag*, *Ser* (sericin gene) and *BR* (Balbiani ring genes; both fibrous proteins) genes are made up of multiple exons and introns. Sequences coding for crystalline and non-crystalline protein domains are integrated in the repetitive regions of *Fhc* and *MA* exons, but not in the protein glues *Ser1* and *BR-1*. Genetic 'hot-spots' promote recombination errors in *Fhc*, *MA*, and *Flag*. Codon bias, structural constraint, point mutations, and shortened coding arrays may be alternative means of stabilizing precursor mRNA transcripts. Differential regulation of gene expression and selective splicing of the mRNA transcript may allow rapid adaptation of silk functional properties to different physical environments.

Keywords: *A. pernyi*; *B. mori*; Evolvability; *N. clavipes*; Gene; Genetic architecture; Fibrous protein; Silk

1. Introduction

Evolvability is the capacity of the genome to produce new genotypes (Kirschner and Gerhard, 1998). Whether gene changes and new interactions result in adaptive change depends on how tightly the genotype is mapped onto an organismal phenotype (Wagner and Altenberg, 1996). One of the most direct genotype–phenotype maps is that charting the relationship between the molecular genetic organization of the silk gene and the molecular organization and function of silk proteins. As in venoms, plant defense compounds, and proteinaceous flower pigments, the gene product is used externally in a way that is easily measured.

The two most important periods of speciation in the Araneae—the divergence of the Araneomorphae from the Mygalomorphae in the Middle Triassic (Seldon and Gall, 1992) and the divergence of the derived Araneoidea from their ances-

---

☆ Contributions to a special issue of CBP on Comparative Functional Genomics.

*Corresponding author. Museum of Comparative Zoology, Harvard University, 26 Oxford Street, Cambridge, MA 02138, USA. Tel.: +1-617-496-8146; fax: +1-617-642-0931.

E-mail address: ccraig@oeb.harvard.edu (C.L. Craig).

PII: S1096-4959(02)00095-7

tral taxa the Deionopoidea during the early Cretaceous (Seldon, 1989)—correlate with the evolution of two types of silk-producing glands: the major ampullate gland (*MA*) and the flagelliform gland (*Flag*). Despite the importance of these events, little is known about the selective and molecular mechanisms that resulted in the evolution of the *MA* and *Flag* silk proteins. In this review we will compare the molecular genetic organization of *MA* and *Flag* genes and proteins produced by spiders to two types of fibroin silks produced by Lepidoptera larvae: the heavy chain fibroin (*Fhc*) spun by the larval moth *Bombyx mori* (*Bm-Fhc*; Bombycidae) and that spun by its sister taxa *Antherea pernyi* (*Ap-Fhc*; Saturniidae). In addition, the molecular genetic organization of the fibrous glue sericin (*Ser*; produced by *B. mori*) and the Balbiani Ring (*BR*) gene proteins produced by the larval fly *Chironomus tentans* (Diptera) as possible models for the fibrous protein glues produced by spiders will be discussed.

The fibroin genes and proteins differ at the level of their molecular genetic architecture, in amino acid sequence, and in structural configuration of the protein. In addition, all of the silk genes are inherently prone to recombination errors due to their highly repetitive nature and codon biases, and display considerable allelic variation that is attributed to a high frequency of unequal crossover (Gage and Manning, 1980; Mita et al., 1994; Hayashi and Lewis, 2000; Sezutsu and Yukuhiro, 2000). Despite their architectural differences, sequence differences, and instability, there are emergent properties of the organization of silk genes and proteins that suggest that a dynamic evolutionary conflict between genetic processes and natural selection may have played a role in silk evolution (Hayashi and Lewis, 2000; Gatsey et al., 2001).

2. The known silk fibroins and fibrous glues are encoded by members of the same gene family

The early divergence of the Araneomorphae (150 mya) and the Araneoidea (65 mya) from their sister taxa correlate with the evolution of *MA* and *Flag* silks. Despite their ancient roots, the repetitive nature of spider silk proteins are largely attributed to the variable proportions of the four amino acid motifs A_n, GA, GGX, and $GPG(X)_n$. While the sequence $GPG(X)_n$ is unique to the *MA2* and *Flag* silks produced by the araneoids, even the silks and fibrous glues of Lepidoptera larvae contain A_n, GA and GGX components (Table 1). Therefore, despite the fact the silks spun by insects and spiders evolved independently from one another, they display some remarkable similarities that could suggest convergence and stabilizing selection have been important factors in their evolution (Gatsey et al., 2001).

Within the Lepidoptera and Araneidae, all silk proteins and fibrous proteins are proposed to be members of a common gene family (Mita et al., 1994; Guerette et al., 1996; Hayashi and Lewis, 1998). The silk proteins these genes encode vary in structural organization but in general are made up of different combinations of β-pleated sheets tightly packed into crystalline arrays, loosely associated β-sheets, α-helices, β-spirals, and spacer regions of anomalous amino acid sequence (i.e. Xu and Lewis, 1990; Beckwitt and Arcidiacono, 1994; Guerette et al., 1996; Simmons et al., 1996; Hayashi and Lewis, 1998).

The *Fhc* silk spun by *B. mori* is the product of three genes. The first identified silk gene encodes for a very large, insoluble protein designated as the heavy-chain fibroin (*Fhc*, 350 kDa). Two additional genes, the L-chain (25 kDa designated as *Fl*) and the P25 protein, encode two subunits of the *B. mori* fiber that are linked to the fibroin by disulfide bonds. Together, the protein products of these three genes make up the water insoluble silk core of the cocoon silk fiber (Grzelak, 1995).

B. mori also produces sericin, a protein glue that ensures that the fibroin cocoon threads adhere to one another. Five sericin genes (*Ser1*, *Ser2*, *MSGS3*, *MSGS4*, *MSGS5*) have been cloned and partially sequenced (Grzelak, 1995). The structure of the *Ser1* gene is best known. It encodes six mRNA transcripts that result from selective splicing of the primary mRNA transcript (Garel et al., 1997; Michaille et al., 1986; Couble et al., 1987; Hayashi and Lewis, 1998). Selective splicing of mRNA transcripts may be an important mechanism for producing alternative proteins with alternative properties in other silk proteins as well. All of the fibroin proteins and *Ser1* share homologous 5′-flanking sequences (Grzelak, 1995).

The Diptera fly larva *C. tentans* secretes fibrous proteins that it uses to construct an underwater tube from which it feeds. Five genes make up the Balbiani ring (BR) gene proteins, (*BR1*, *BR2.1*, *BR2.2*, *BR3*, *BR6*). Four of these encode an uninterrupted block of 100 repeat units that are tandem-

Table 1
Comparison of repetitive units of fibroins and fibrous proteins produced by insects and spiders

	Primary repetitive motifs*	Source	Ref
Bombyx mori (*Bm-Fhc*)	(GGNGCN)$_n$ GGTTCA n = 0 ∼ 6, W = A of T = A,G,T,Cn = 0 ∼ 6, W = A of T = A,G,T,C GGNGYN)$_N$ GGNTAY n = 0 ∼ 8, Y = T,C	Genomic DNA, complete	(Zhou et al., 2000)
Antheraea pernyi (*Ap-Fhc*)	Constant domains GGYGSDS(A)$_{13}$ GSG$\underline{\text{AGG}}$ GGYGSGSS(A)$_{13}$ SG$\underline{\text{AGG}}$ Variable domains AGGGYGWGGD A (GGY)$_n$ RGD RRAGHDRAAGS	Genomic DNA, complete	(Sezutsu and Yukuhiro, 2000)
Nephila clavipes(*Ma-1*)	Possible domain organizations (A)$_N$ $\underline{\text{GGA}}$(Repetitive with variation)GAG(Variable)	cDNA, partial	(Xu and Lewis, 1990)
Nephila clavipes (*Ma-2*)	Possible domain organizations (A)$_N$ (variable)$\underline{\text{PGG}}$YGPGQQGG$\underline{\text{PGGY}}$(Variable)	cDNA, partial	(Hinman and Lewis, 1992)
Nephila clavipes (*Flag*)	GPGG(X)$_n$ (GGX)$_n$ (Variable and nonrepetitive)(GGX)$_n$ (GPGGX)$_n$	cDNA, partial genomic, partial	(Hayashi and Lewis, 2000)
B. mori (*SerI*)	LSEDSSEVDIDLGGNLGWWWNSDNKAQRAAGGATKSEASSSTQ (SRTSGGTSTYGYSSSHRGGSVSSTGSSSNTDSSTKNAG)$_n$	Genomic, cDNA	(Garel et al. 1997)
Chirnomous tentans(Br1)	AAATCTGGACCAAGATCAAGC AAACCTGAAAAGACCAAGCAAATCAGGACCT AAACCAAGTAAGGGATCTAAACCTAGACCAGAG	cDNA, partial genomic, partial	(Paulson et al., 1992)

[*] Primary repeat motifs are abstracted from full sequences. See reference for complete sequences and polymorphic sites.

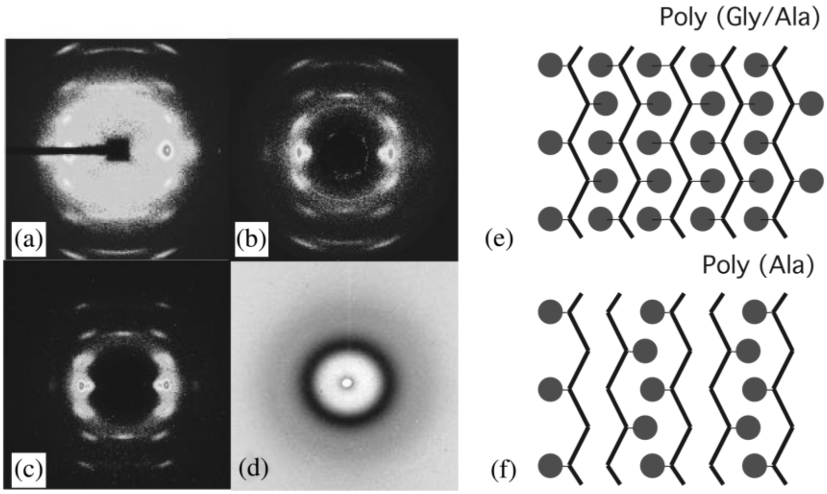


Fig. 1. X-Ray diffraction patterns of *Fhc* and *MA* silks spun by (a) *Bombyx mori*, (b) *Nephila clavipes*, (c) *Argiope argentata* show crystalline and non-crystalline regions; (d) X-ray diffraction patterns of *Flag* silk spun by *Micrathena gracilis* show no crystalline regions. The fiber diffraction patterns of *N. clavipes* and *A. argentata MA* silk illustrate a poly-alanine crystal organization, while the fiber diffraction pattern from *B. mori* cocoon silk illustrates a poly(glycine-alanine) crystal organization. (e,f) Illustration of how poly-A and poly-GA amino acid organization affects crystal packing. The diffraction pattern for *M. gracilis Flag* silks shows only an amorphous halo.

ly arrayed and almost identical. The gene *BR3*, however, is characterized by a different structural organization. *BR3* encodes a 10.9 kb transcript that is spliced into 5.5 kb mRNA. Fifty-eight introns separate coding units that vary in size between 17 and 678 bp. However, the positions of the introns relative to the repeat structure of the protein are the same.

Adult female araneoid spiders have seven different types of glands that yield four fibroin silks and three types of protein glue (Kovoor, 1987; Craig, 1997). The fibroin silk genes in spiders are expressed in the major ampullate gland (*MA*), the minor ampullate gland, (*MiA*), the flagelliform gland (*Flag*), and the cylindrical gland (*Cy*). Protein glues are produced in the piriform gland (*Pyr*), the aggregate gland (*Ag*), and the aciniform glands (*Ac*). Aciniform gland silk has some crystalline organization; the physical structure of the pyriform proteins is unknown. Recent X-ray diffraction patterns of aggregate gland proteins suggest that they may contain some crystalline β-sheet material as well (Craig and Riekel, unpublished).

Both the Lepidoptera and Diptera produce silks or fibrous proteins in modified salivary glands when they are larvae. Spiders produce silks in multiple glands and throughout their lives. The differences in number of silks, gland morphology, and life stages in which production occurs suggest that there have been different selective factors acting on Lepidoptera and spiders affecting silk organization and function. For example, in spiders, the greater number of proteins is the result of divergence among the silk genes as well as the evolution of specialized protein processing systems (Vollrath and Knight, 2001).

3. Most silk fibroins contain crystalline and non-crystalline regions

X-Ray diffraction and NMR data show that the *MA* silks spun by *Nephila* and *Fhc* silks spun by *B. mori* and *A. pernyi* contain crystalline and non-crystalline components (Fig. 1). The proposed molecular organization of the *MA* genes is similar to that of the *Ap-Fhc* gene, but both are fundamentally different from that of *Bm-Fhc.* The crystalline domains of the *Bm-fhc* fibroins are largely composed of repeating glycine and alanine that are arrayed in two related patterns. *Ap-fhc*, *Ma1*, and *Ma2*, in contrast, contain strings of poly-alanine that result in their crystal-forming domains (Table 1).

In all of the silks above, crystalline regions are interspersed by domains of 34–40 amino acids that make up the non-crystalline regions of the protein. Different silks, however, have different proportions of non-crystalline and crystalline fractions. Approximately 40–50% of *B. mori* silk is made up of protein crystals (Iisuka, 1965). When hydrated, only approximately 15% of the total volume of MA silks is crystalline (Gosline et al., 1999) which can be related to the swelling of the non-crystalline fraction. X-Ray diffraction data on *Nephila* silk suggest a size of <6 nm for the crystalline β-sheet domains (Grubb and Jelinski, 1997; Riekel et al., 1999a,b). More complex models were developed based on analytical transmission electron microscopy. Thus the existence of large β-sheet crystallites of 70–100 nm size was proposed for *Nephila* silk (Thiel et al., 1994). In related experiments, the notion of non-periodic lattice crystals (NPL) was introduced and the possibility that the β-sheet structure could change locally from crystalline to aperiodic (Thiel and Viney, 1997; Thiel et al., 1997). X-Ray fiber diffraction data (Thiel et al., 1997) suggesting the existence of more complex structural features than the classical β-sheet structure have, however, not been confirmed by single fiber X-ray diffraction (Riekel et al., 1999a,b).

The non-crystalline fraction of silk is currently the focus of many investigations of silk structure. For example, *MA* silk contains at least two different types of proteins that differ in their amino acid content (Fig. 2). The cDNA designated as *Ma1* is made up of repeated motifs largely composed of glycine and alanine. The cDNA designated as *Ma2* is made up of repeated motifs largely composed of the amino acids glycine, alanine, and proline. Thus, although *Ma1* contains almost no proline, proline makes up approximately 15–17% of the repeat motif in *Ma2* (Gosline et al., 1999). Proline is thought to affect the organization of the non-crystalline regions of the proteins and hence silk elasticity (Hayashi and Lewis, 1998; Gosline et al., 1999).

For example, using NMR analyses, van Beek et al. (2000) proposed that the glycine-rich, non-crystalline domains of *Ma1*, the proline-free silk, are organized into aggregated, 3_1-helix or Type I β turn that reinforces the highly oriented, polyalanine polymer network. X-Ray diffraction data suggest the presence of two types of non-crystalline material: (i) randomly oriented material; and (ii) oriented material (Grubb and Jelinski, 1997; Riekel et al., 1999a,b). The oriented non-crystalline fraction has been suggested to form a fibrillar system with the crystalline fraction (Yang et al., 1997; Riekel and Vollrath, 2001). The model of NPL crystallites (see above) is an interesting hypothesis but has not been verified by other groups.

The models of the two silks suggest that while proline enables the large-scale organization in *Ma2,* it disrupts the structural organization of the glycine-rich helices in *Ma1* (Gosline et al., 1999). Furthermore, the conserved, poly-alanine sequences that make up the crystalline regions of *MA1* and *MA2* in conjunction with the divergent non-polyalanine sequences that make up the non-crystalline regions of the silks may suggest that any diversification in *MA* silk structure and function is the result of reorganization and variation in the non-crystalline regions of *MA* proteins. Therefore, the non-crystalline regions of *MA* silks are probably the primary site of evolutionary innovation (Craig et al., 2000).

Silk produced in the flagelliform glands of spiders do not contain polyalanine, nor do they contain domains of β-pleated sheets. In the few samples of flag silk so far examined, X-ray diffraction has not shown a crystalline fraction (Fig. 1d; Craig and Riekel, unpublished), which suggests that *Flag* silks are structurally different form any other silk proteins studied to date.

4. All of the sequenced fibroin silks (Fhc, MA, and Flag) are made up of hierarchically organized, repetitive arrays of amino acids

The cocoon silks produced by the Lepidoptera moth larvae *Bombyx mori* (Bombycidae) (*Bm-Fhc*) and *Antheraea pernyi* (*Ap-Fhc*), spider *MA* silk, and *Flag* silk are made up of repetitive amino acid sequences that are highly conserved both within and, in some cases, across taxa (Figs. 2–6; Table 1). Cocoon silks produced by *Bombyx mori* are made up of sequences of glycine and alanine $(GA)_n$ organized into tandem units of GAGAG(S) and GAGAGX (X varies between tyrosine and valine; Zhou et al., 2000). Even though they are sister taxa, cocoon silk sequences from *A. antheraea* (*Ap-fhc*) (Saturniidae) are fundamentally different from *B. mori* sequences (Table 1; Fig. 1). *Ap-fhc* silks are composed of strings of polyalanine (hereafter referred to as PAB, polyalanine

A. cDNA for Major ampullate protein 1 *(Ma1)*. Possible genetic hot-spot (GGA) in bold. Note relative paucity of amino acid proline.

Q**GAG**
AAAAAA**GGA**GQGGYGGLGGQGAGQGGYGGLGGQGAGQGAG
AAAAAAA**GGA**GQGGYGGLGSQGAGRGGQGAG
AAAAAA**GGA**GQGGYGGLGSQGAGRGGLGGQGAG
AAAAAAA**GGA**GQGGYGGLGNQGAGRGGQG
AAAAAA**GGA**GQGGYGGLGSQGAGRGGLGGQGAG
AAAAAA**GGA**GQGGYGGLGGQGAGQGGYGGLGSQGAGRGGLGGQGAG
AAAAAAA**GGA**GQGGLGGQGAGQGAGAS
AAAA**GGA**GQGGYGGLGSQGAGRGGEGAG
AAAAAA**GGA**GQGGYGGLGGQGAGQGGYGGLGSQGAGRGGLGGQGAG
AAAA**GGA**GQGGLGGQGAGQGAG
AAAAAA**GGA**GQGGYGGLGSQGAGRGGLGGQGAGAV
AAAAA**GGA**GQGGYGGLGSQGAGRGGQGAG
AAAAAA**GGA**GQRGYGGLGNQGAGRGGLGGQGAG
AAAAAAA**GGA**GQGGYGGLGNQ**GAG**RGGQG
AAAAA**GGA**GQGGYGGLGSQGAGRGGQGAG
AAAAAAV**GAG**QEGIRGQGAGQGGYGGLGSQGSGRGGLGGQGAG
AAAAAA**GGA**GQGGLGGQGAGQGAG
AAAAAAGGVRQGGYGGLGSQGAGRGGQGAG
AAAAAA**GGA**GQGGYGGLGGQGVGRGGLGGQGAG
AAAA**GGA**GQGGYGGVGSGASAASAAASR
LSPQASSRVSSAVSNLVASGPTNSAALSSTISNVVSQIGASNPGLSGCDVLIQALLEVV
SALIQILGSSSIGQVNYGSAGQATQIVGQSVYQALG"

B. cDNA for Major ampullate protein 2 *(Ma2)*. Proline in bold is abundant; possible genetic hot-spot (PGG) in bold.

PGGYGPGQQGPGGYGPGQQGPSGPGS
AAAAAAAAAAG**PGG**YGPGQQGPGGYGPGQQGPGRYGPGQQGPSGPGS
AAAAAAGSGQQG**PGG**YGPRQQGPGGYGQGQQGPSGPGS
AAAASAAASAESGQQG**PGG**YGPGQQGPGGYGPGQQGPGGYGPGQQGPSGPGS
AAAAAAAASGPGQQG**PGG**YGPGQQGPGGYGPGQQGPSGPGS
AAAAAAAAS**GP**GQQG**PGG**YGPGQQGPGGYGGQQGLSGPGS
AAAAAAAGPGQQG**PGG**YGPGQQGPSGPGS
AAAAAAAAAG**PGG**YGPGQQG**PGG**YGPGQQGPSGAGS
AAAAAAAGPGQQGLGGYGPGQQG**PGG**YGPGQQG**PGG**YGPGSAS
AAAAAAGPGQQG**PGG**YGPGQQGPSGPGSAS
AAAAAAAAG**PGG**YGPGQQG**PGG**YAPGQQGPSGPGSAS
AAAAAAAAG**PGG**YGPGQQG**PGG**YAPGQQGPSGPGS
AAAAAAAAAG**PGG**YGPAQQGPSGPGIAASAASAG**PGG**YGPAQQGPAGYGPGSAVAASAGAGSAGYGPGSQA
SAAASRLASPDSGARVASAVSNLVSSGPTSSAALSSVISNAVSQIGAS**NPG**LSGCDVLIQALLEIVSACVT
ILSSSSIGQVNYGAASQFAQVVGQSVLSAFX

Fig. 2. Amino acid sequence of cDNA for *Ma1* (A) and *Ma2* (B) silks produced by *Nephila clavipes*. The cDNA shows that proline is virtually absent from the *Ma1* cDNA sequence although it is abundant in the *Ma2 cDNA* sequence. Furthermore, both *MA* sequences contain DNA 'hot-spots' (GGA, *Ma1*; PGG, *Ma2*). The arrangement of the *MA* PAB (crystalline) and NPAB (non-crystalline) domains and may be similar to the structure proposed for the silk fibroin *A. pernyi.*

block) with one of four, unique, non-polyalanine blocks (hereafter referred to as NPAB, non-polyalanine) (Sezutsu and Yukuhiro, 2000). In addition, the NPABs encode four repetitive motifs that show varying amounts of polymorphism. The variable and repetitive of amino acid sequences that make up the NPAB tails are the result of slippage and unequal pairing during cross-over (Sezutsu and Yukuhiro, 2000).

Spider *MA* silk is also composed of PABs and NPABs. The PAB blocks of *MA* silks sequenced to date have fewer alanine residues (4–10) than the sequence motifs that make up larval *Ap-fhc* silk (Table 1). However, like *Ap-fhc,* a variable amino acid domain and a constant amino acid domain follow the polyalanine regions of both *Ma1* and *Ma2.* Analogous to *Ap-fhc*, the variable domain in spider *MA* silk is proceeded by repetitive sequence of nucleotides may result in a genetic hotspot that initiates frequent rearrangements of the gene (Jeffreys et al., 1985; Sezutsu and Yukuhiro, 2000). The number and organization of the

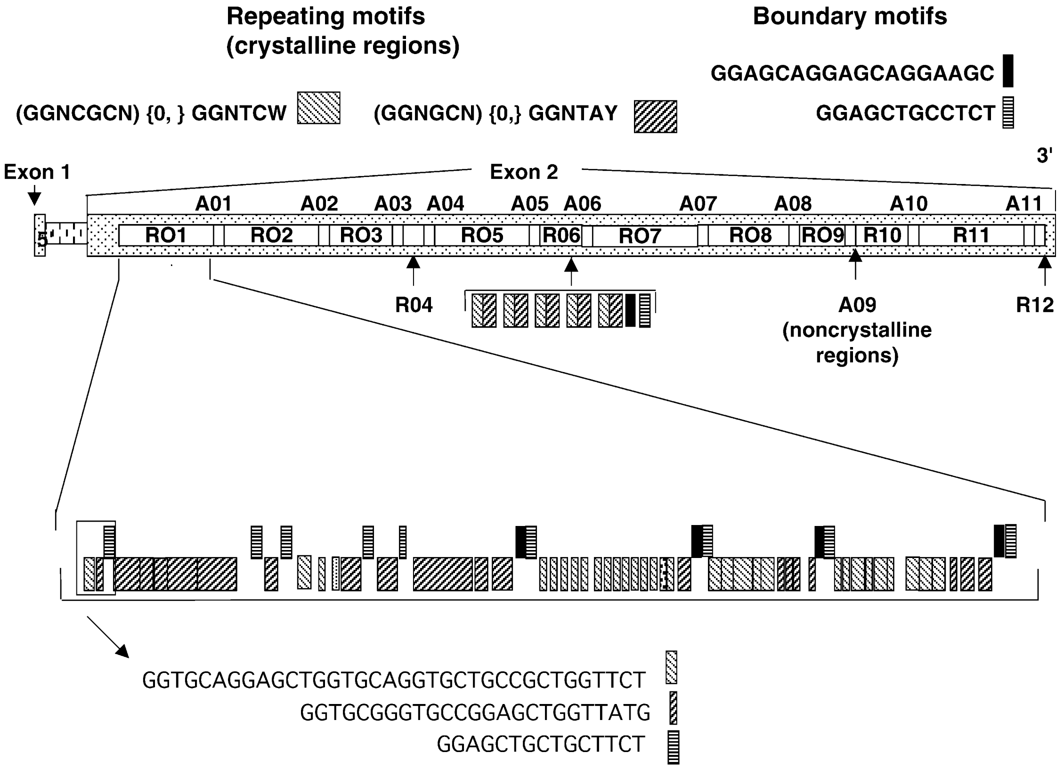


Fig. 3. The repeating sequence structure of silks spun by *B. mori* is hierarchically organized. The diagram illustrates the two-exon, one-intron repetitive makeup of the *B. mori* silk gene. Exon 2 illustrates the integrated and repetitive crystalline, non-crystalline, and border domains of the gene. It also shows that repeating arrays are variable in size. Figure modified from Zhou et al., 2000.

potential NPAB motifs in *Ma1* and *Ma2* silk, however, cannot be determined due to the limited sequence data available. Nevertheless, the preliminary data suggest that the NPAB amino acid motifs of *Ma2* and *Ma1* differ in both sequence and length. In addition, although both contain polymorphic sites, *Ma2* silks are significantly more polymorphic than *Ma1* silks. The primary repeating motifs for *Ma2* silk are GPGXX/GPGQQ, GP(SYG), and A_n. The primary repeating motifs for *Ma1* are (GA), A_n, GGX.

Flag silks spun by *N. clavipes* do not contain PABs but do share the repetitive motifs GGX (Alanine, Serine) and GPGGX with the *Ma2* fibroin. Although some of the *Flag* sites are polymorphic, the sizes of each of the repetitive regions or exons of *Flag* silks are more similar to each other than to the size of the repetitive regions in either *MA* or *Fhc* silks. This suggests that that *Flag* exons are either subject to homogenization or that they are evolving under a more stringent selection than either *MA* or *Fhc.*

5. Fhc fibroin genes (and perhaps MA genes) are characterized by a similar molecular genetic architecture of two exons and one intron, but the organization and size of these units differs

Three different types of molecular genetic architecture have been identified for silk fibroins and their related fibrous proteins. These include the molecular genetic architecture of the *Fhc* silks (Figs. 3 and 4), of *Flag* silk (Fig. 5), and of the fibrous proteins such as *Ser1* (Fig. 6).

The *Bm-Fhc* gene is made up of two exons in which the hexapeptides are hierarchically organized (Fig. 3). Each encodes for both crystalline and non-crystalline protein domains that result in a 5263 amino acid polypeptide. *B. mori* exon 1 contains 25bp of untranslated domain and a 42 bp region that encodes for 14 amino acids (Zhou et al., 2000). The repetitive domain of exon 2, separated from exon 1 by an intron of 970 bp, is organized into 12 subdomains that vary in size from 111 to 255 bp. The repetitive motif of exon

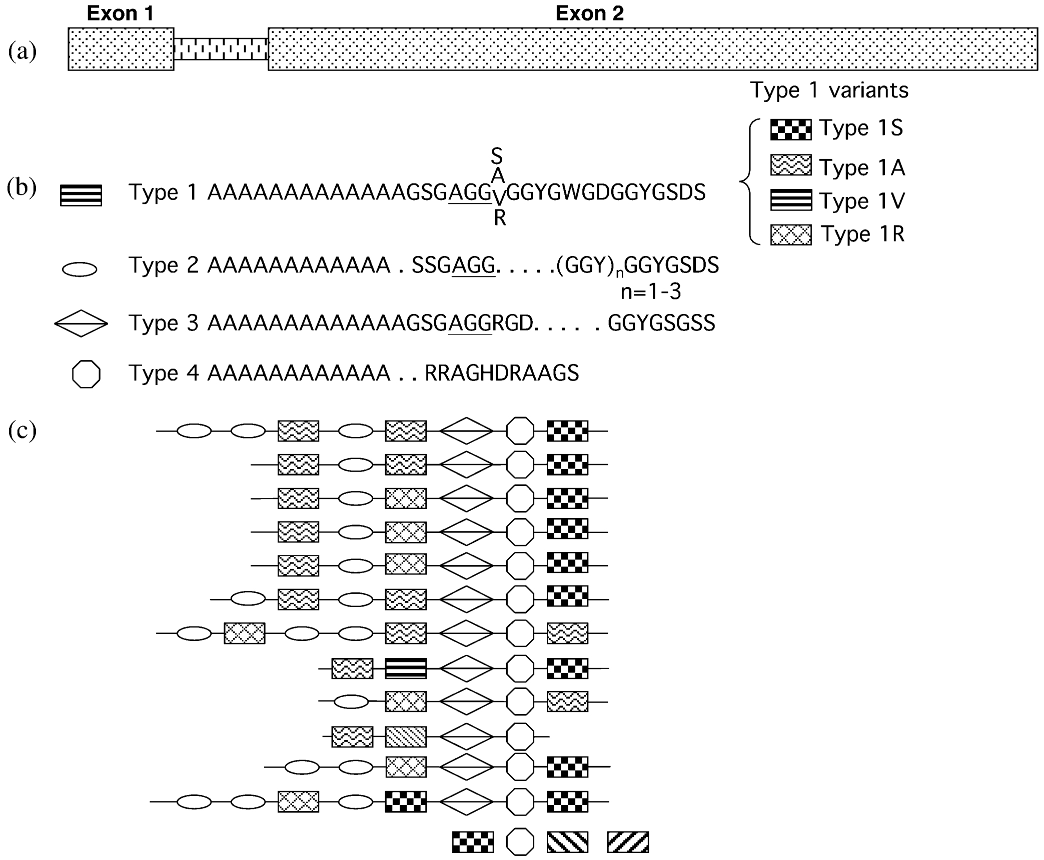
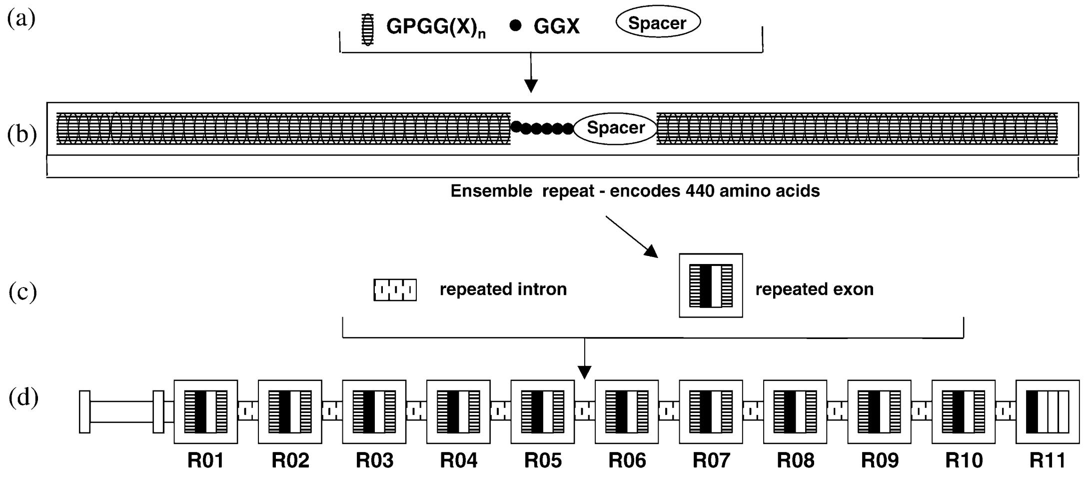


Fig. 4. Repeating sequence structure of silks spun by *A. pernyi* is hierarchically organized. Genomic DNA for cocoon silks spun by *A. pernyi* show four motifs. The figure illustrates the constant and variable domains of the gene. The Chi-like triplet, AGG is underlined. Unlike *B. mori* poly-alanine makes up the crystalline regions of the protein and the non-crystalline regions are made up of a series of motifs which contain both constant and variable domains. Figure modified from Sezutsu and Yukuhiro (2000).

2 is similar to that of exon 1 but less well conserved. In addition to the repetitive domains, are flanking, non-repetitive amino acid regions that are highly conserved and that code for the non-crystalline components of the protein (Mita et al., 1994). Finally, exon 2 contains two different types of boundary or spacer units whose sequences are non-repetitive.

The *A. pernyi-Fhc* gene is also made up of two exons and one intron, but the total length of the coding region is approximately one-half that of *Bm-Fhc* (Fig. 4). The full length of the *A. pernyi-Fhc* is approximately 8.1 kb, and it codes for a polypeptide chain of 2639 amino acids. Like *Bm-Fhc*, *Ap-Fhc* is characterized by one short exon encoding 14 amino acid residues, one intron (120 bp), and a second, larger exon encoding 2625 residues. Unlike the *Bm-Fhc* fibroin, *Ap-Fhc* contains a unique sequence of 155 residues of amino acid terminal sequence. The hierarchically organized units are predominately alanine (43%), glycine (27%), and serine (11%). The *Ap-Fhc* intron shows no nucleotide similarity to the *Bm-Fhc* intron, and exon 2 of *Ap-Fhc* and *Bm-Fhc* are dissimilar as well (Tsujimoto and Y, 1979; Mita et al., 1994; Sezutsu and Yukuhiro, 2000). The first exons of *Ap-Fhc* and *Bm-Fhc* are identical at 11 out of 14 residues (Tamura et al., 1987), which may indicate a similar function (Sezutsu and Yukuhiro, 2000).

Exon 2 encoding of the *Ap -Fhc* protein contains 80 repetitive regions made up of linked polyalanine (PAB) and non-polyalanine (NPAB) units (Fig. 4). Seventy-eight of the 80 repetitive regions can be grouped into four types of motifs based on their NPAB sequences. Type 1 motifs are made up of four subtypes. Type 2 motifs differ in GGY triplets, suggesting that replication slippage events may be



From: Hayashi and Lewis, 2000.

Fig. 5. Repeating motif and structure of the *Flag* gene of *Nephila clavipes.* (a) *Flag* silk is also hierarchically organized. Repeat motifs GPGG(X)*ₙ*, and GGX are organized into an ensemble repeat. The ensemble repeats, in turn are organized into 11 repeating exons. *Flag* silk does not contain the GAGAX or poly-A sequences that encode the β-pleated sheet domains of the *Fhc* or *MA* silk. (b) Comparison of repeat motifs for the 10 repetitive exons of two species, *N. clavipes* and *N. maculata* illustrates variation in repeat arrays. Figure modified from Hayashi and Lewis (1998).

responsible for their generation. Type 3 motifs contain an Arg-Glycine-Asp (RGD) triplet. Type 4 is highly heterologous to the other motifs (Sezutsu and Yukuhiro, 2000). *Ap-Fhc* includes variable domains that are not found in *Bm-Fhc* fibroin silks. Furthermore, the PAB sequences are highly conserved, but NPAB sequences are variable. This suggests either that the NPAB region of the molecule is evolving under less stringent selective factors than the PAB region or that the unique properties of fibroin result from positive selection on the NPAB sequences (Sezutsu and Yukuhiro, 2000).

The molecular genetic architecture of the *MA* silks spun by *Nephila clavipes* has not been determined. However, based on 1951 base pairs of genomic transcript and cDNA identified as *Ma1, Ma2,* some investigators have proposed that the *Nc-MA1 s*ilk protein is encoded by a single exon (Xu and Lewis, 1990; Hinman and Lewis, 1992; Beckwitt and Arcidiacono, 1994; Guerette et al., 1996; Beckwitt et al., 1998). It may be that both

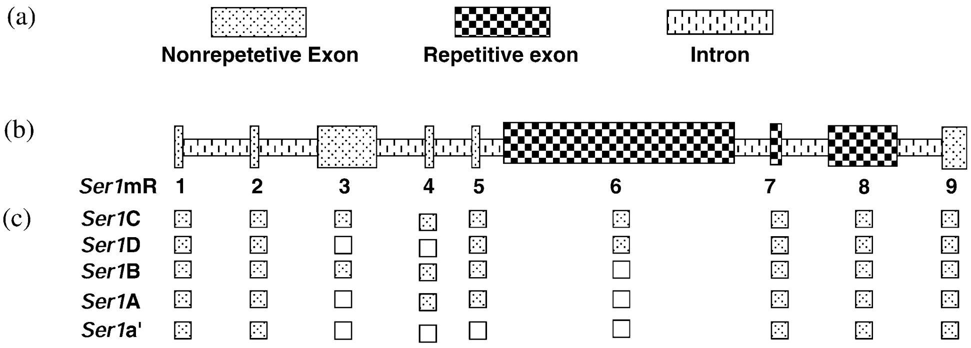


Fig. 6. Repeating motif and schematic structure of the *Ser-1*gene of *B. mori.* (a) The schematic of the *Ser-1* gene show that the irregular size of the exons and introns as well as the positions of their amorphous and crystalline regions. (b) The primary sequence of (*Ser1B*) shows the amino acid sequences of exons 1–5, 7–9. Unlike the *Fhc genes,* the crystalline and non-crystalline regions are not integrated in the Ser1B cDNA peptide. Figure modified from Garel et al. (1997).

MA proteins have an organization similar to *Ap-Fhc* (Fig. 2a,b).

The identified portions of *MA* sequence are hierarchically organized and contain characteristic, repetitive motifs that are common across ancestral and derived species (Gatsey et al., 2001). As in the *Fhc* genes, the number of amino acid, tandem repeats that make up *MA* is not rigorously conserved but retains a hierarchical and structural identity. This contrasts with the nearly exact repeats that characterize the *Bm-fhc* gene. Although it is not possible to analyze *MA* sequences in detail due to the relatively small portion of cDNA and gene that have been sequenced, the pieces identified so far are strikingly similar in organization to those observed in *Ap-fhc* gene (Sezutsu and Yukuhiro, 2000). Like *Ap-fhc*, both *Ma2* and *Ma1* repetitive units are organized into poly-alanine blocks (6–10 residues) and non-poly-alanine blocks (16–43 residues). Furthermore, like *Ap-fhc*, *Ma1* and *Ma 2* seem to have ensemble repeats that are non-randomly distributed across the cDNA transcripts sequenced to date. Finally, in a similar position and preceding the variable domains of the spider NPABs, there is a nucleotide sequence associated with 'genetic hotspots'. Compared with genetic hotspots in *A. pernyi Fhc* (GGA), the spider sequence AGG is almost identical in nucleotide composition and relative position to the NPABs. These similarities suggest a possible structural organization for *MA* silk. This could also suggest that the NPAB sequences of *MA* silk, like those of *Ap-fhc,* may be subject to dynamic rearrangement.

6. The Flag, Ser, and BR genes are made up of multiple exons and introns

Flag, *Ser* and *BR* silks, made up of multiple exons and introns, have a completely different genetic architecture from the *Fhc* and *MA* genes (Figs. 5 and 6). In the cases of *Ser* and *BR*, some of their exons encode for crystalline protein regions and others for non-crystalline protein regions. Therefore, primary coding regions of the protein glues are not homogeneous and selective splicing of mRNA transcripts results in proteins with different structural domains. In the case of the *Flag*, 11 of the 13 exons are made up of similar repetitive units. None of the exons encode β-pleated sheet structures and instead encode protein domains that have been suggested to be a mix of β-helical and β-spiral structural domains (Hayashi and Lewis, 1998). Selective splicing of *Flag* mRNAs would result in different sized protein domains that, in turn, could affect the protein's structure and function.

The *Flag* gene spans a total of 36 kb, of which 15.5 kb is protein coding (Hayashi and Lewis, 2000). Hence, the size of the transcribed *Flag* and of *Bm-Fhc* proteins are similar, and this could suggest some functional or mechanical limit to fibroin size. None of the 13 exons that make up *Flag* contain poly-alanine and hence crystalline structural domains. Exons 3–12 encode the amino acid motifs GPGGX and GGX, which occur in tandem arrays of 440 amino acids with a non-repetitive, highly conserved spacer unit (Table 1). In the primary repeating motif $GPGG(X)_n$, alanine (32%), serine (26%), valine (11%), and tyrosine (20%) occupy 89% of the polymorphic sites. In the second primary repeat motif, alanine (46%), serine (39%), and tyrosine (11%) occupy 96% of the polymorphic sites of GGX. Thus, repeating motifs of the *Flag* gene contain more polymorphic sites than do the repeating motifs of the *Fhc* gene. *Flag* exons 1 and 2 encode for a non-repetitive terminal region. Exon 13 codes for the 3′ end of the molecule, which includes both non-repetitive and repetitive regions like those that characterize the previous exons (Fig. 5). Furthermore, the sizes of the *Flag* exons are much more homogeneous than the size of the repeating units of *Bm-Fhc.* The extreme similarity between some of the introns suggests homogenization due to recombination (Hayashi and Lewis, 2000).

Ser1 is the best-known sericin gene, although partial sequences for at least four others have been identified (Grzelak, 1995). *Ser1* is 23 KB-long and is made up of nine exons and eight introns (Fig. 6). These include a large, central, alternative exon, which encodes for 60 repeats of a characteristic 114 bp motif (exon 6), in addition to two other exons that are repetitive (7,8) (Garel et al., 1997; Michaille et al., 1986). Five different *Ser1* isoforms, each playing a different functional role in the cocoon, have been identified (Fig. 6; Zhou et al., 2000). As can be seen, the sericins are diverse.

There are five closely related *BR* genes, and they have been partially characterized. All of the *BR* genes have an exon/intron organization. Like the *Ser1* gene, *BR* exons code for either repetitive arrays or non-repetitive arrays but not both. While the *Ser1* gene encodes three differently sized, repetitive domains, the repetitive regions of *BR2, BR2.2, BR6* encode four proteins that contain an uninterrupted block of approximately 100 tandemly arranged repeats organized into two repeat units (Paulsson et al., 1990). The architecture of the repetitive regions of the *BR3* gene, however, is different. The repetitive array of the *BR3* is interrupted by 38 introns. The resulting exons are more variable in size and sequence than the repetitive arrays of the other genes. The presence of the introns in *BR3*, and the small sizes of some of its exons, may prevent exon homogenization that is dependent on unequal aligning of homologous sequences (Paulsson et al., 1990). In the case of *Flag,* homogeneous introns separate similar sized exons. The extreme homogeneity of the size of *Flag* exons suggests that they are the product of a recent duplication event and/or that exon size is maintained by stringent selection. The evolution of *Br3* may be more complicated or selection for function may be relaxed.

The importance of the nine exon/intron structure of *Ser1* is that it allows multiple protein variants to be produced from a differentially spliced mRNA transcript. Specifically, differential splicing of *Ser1* gives rise to the five different types of mature mRNA transcripts (Fig. 6). In contrast, variation in *Fhc* silks is the result of unequal cross-over during recombination and largely limited to variation in allele length. Furthermore, new *Fhc* variants are produced during meiosis and hence the frequency of their appearance is limited by the length of insect generation time. However, selective splicing of the exons that encode *Ser* mRNA can generate multiple protein variants at different times during the larvae's development. Therefore, proteins with different structural configurations and properties can be produced from a single gene within the same individual at each transcription event. In the case of the *Ser* gene, the transcribed sericin proteins range from 65 to 400 kDa: exon 1 (93bp, 13 codons), 2 (31 bp, 10 codons), 9 (569 bp and 369 bp with coding sequence followed by 188 bp of non-coding sequence), 7 (129 bp, repetitive), and 8 (114 bp repeats, repetitive) are constitutive and therefore found in all of the sericins. Exons 3 (1314bp, 483 codons), 4 (93 bp, 31 codons), 5 (78 bp, 26 codons), and 6 (6.6 kb, repetitive) are alternative (Prudhomme et al., 1985). Exons 6 and 8 code for a 40% serine-rich motif and result in most of the β-sheet structure (Komatsu, 1985). The peptides encoded by exons 3, 6 and 8 have specific physicochemical and structural properties that are responsible for the sliding and sticking of the sericins to the fibroin. Peptide 3 lends the property of high water solubility (Grzelak, 1995).

Ser 2 is 16200 bp and codes for two mRNA transcripts (5.4 bp and 3.1 kb) (Couble et al., 1987; Michaille et al., 1989). One is 3100 nucleotides, and the second is between 5000 and 6400 nucleotides, depending on which allele is considered (Garel et al., 1997; Michaille et al., 1986). The three other sericin genes, *MSGS-3*, *-4* and *-5*, code for 3500, 2950 and 450 nucleotide mRNAs, but their structures are unknown (Grzelak, 1995). The expression of both known genes (*Ser 1* and *Ser 2*) is spatially and developmentally regulated. Differential splicing of the primary transcript of *Ser1* is developmentally regulated as well (Ishiwaka and Suzuki, 1985; Couble et al., 1987; Michaille et al., 1989).

In summary, the multiple exon organization of the sericin gene allows considerable flexibility in the resulting protein because there is little homogeneity among the exons. In contrast, the exons that make up *Flag* are so similar that the effects of selective splicing of the transcript would largely limit variation to size differences of the transcribed regions.

7. Sequences coding for crystalline and non-crystalline protein domains are integrated in the repetitive regions of Fhc and MA exons, but not in the protein glues Ser1 and BR-1

A result of the two-exon structure of *Fhc* and *MA* silks described above is that the primary regions coding for its crystalline and non-crystalline protein domains are integrated into one large, repetitive exon. In contrast, the crystalline and non-crystalline regions of the protein glues *Ser1* and *BR-1* are encoded by unique exons. According to the introns-first theory of gene evolution (genes with intron/exon structure evolved first; genes without introns are derived via intron loss), the exon/intron structure of the proteins *Flag*, *Ser,* and *BR* is more ancient than the structure of the *Fhc* and *MA* silk genes (Souza et al., 1998). However, phylogenetic evidence shows that the *Flag* gene is the most recently evolved type of silk produced by spiders. The amino acid sequences of the *Flag* protein are most similar to the repetitive units of *MA2*. Comparison of the two proteins suggests that *Flag* may have evolved from *Ma2* through the loss of the crystalline, poly-A regions and subsequent gene duplication.

8. Genetic 'hot-spots' promote recombination errors in Fhc, MA and Flag

Glycine, alanine, and serine make up approximately 53–97% of the amino acids of *Fhc* silks spun by the Bombycoidea. Glycine, alanine, and serine compose 42–78% of *MA* silks spin by araneoid spiders (Craig, unpublished). In both cases, a high proportion of their repetitive amino acid and nucleotide motifs are made up of minisatellite sequences. Minisatellite sequences function as genetic hotspots, or recombination signals that induce cross-over events. The instability they introduce can enable gene conversion (Lam et al., 1974; Jeffreys et al., 1985; Paulson et al., 1992). However, gene instability may also generate adaptive variation. The ability of the minisatellite sequences to produce variants in the NPAB sequences of *Ap-fhc* can be viewed, in and of itself, as an evolved response to natural selection (Burch and Chao, 2000).

Genetic hot spots are thought to result in the dynamic reorganizations of four types of repetitive arrays in the NAPB sequences in *Ap-fhc.* Even though there is high sequence conservation in length of the PABs of *Ap-Fhc* (PABs sequence lengths in *MA* silk are less homogeneous), the non-crystalline portion of the NAPBs is variable (Fig. 4; Sezutsu and Yukuhiro, 2000). In the case of the *Ap-Fhc* gene, gene sequence rearrangements are attributed to the GCAGGUGGU nucleotides that result in the amino acid sequence AGG. AGG initiates all variable regions in the *Ap-Fhc* motifs in *Ap-fhc* as well as in *Antheraea yamamai* fibroin silk (Sezutsu and Yukuhiro, 2000). The NPAB rearrangements include multiple duplication events as well as a triplication event of a 558-bp sequence. Similarly, although only a limited amount of data is available, the repetitive nucleotide sequence that encodes GGA in *Ma1* and PGG in *Ma2* silks spun by *N. clavipes* may contribute to the substantial allelic variation observed in these proteins that results from cross-over during recombination (Beckwitt et al., 1998). *Flag* silks, also, contain genetic hotspots. Seventy-five percent of the protein is made up of proline, glycine, and alanine. The observed length differences in the tandem repeats of GPGG$(X)_n$ and GGX are probably due to cross-over events (Hayashi and Lewis, 2000).

9. Codon bias, structural constraint, point mutations, and shortened coding arrays are alternative means of stabilizing precursor mRNA transcripts

Codon bias, structural constraint, point mutations, and shortened repetitive arrays may result in stabilization of the repetitive regions of silk proteins during transcription. The nucleotide biases in *Bm-fhc* codons differentiate the non-crystalline and crystalline regions in the protein molecule. In the case of *Bm-fhc*, the resulting codons determine the secondary structure of the mRNA (Mita et al., 1988; Nakamura et al., 1991). The first and second codon positions show a GC bias that is approximately 80% due to the presence of glycine (GGN) and alanine (GCN) residues (Zhou et al., 2000). Zhou et al. hypothesize that the alternate dicodon, GGU-GCU for GlyAla repeats, results in an inverted stem-loop structure in the crystalline subdomains of the mRNA transcript. The resulting structural constraints, in turn, maintain the secondary structure of the fibroin mRNA(Mita et al., 1988). In contrast, no stem-loop structures are formed in the boundary elements due to high codon diversity. Stem structures are only weakly formed in the region of the transcript coding for amorphous domains (Zhou et al., 2000).

In contrast to GCU codon bias for alanine in the *Fhc* gene, the most abundant codon for alanine in *Ap-fhc* is GCA. Polyalanine in *Ap-Fhc* occurs in blocks of 13 residues. The GCA codon for alanine occurs in 662/1137 times, or across 58% of the coding region (Nakamura et al., 1991; Sezutsu and Yukuhiro, 2000). The maximum number of GCA isocodons in each PAB is seven. Therefore, the tri-nucleotide repeat sequences that encode poly-A could be subject to frequent replication slippage events (Sezutsu and Yukuhiro, 2000). Nevertheless, the variation in alanine number is small and may indicate significant constraint on PAB size. Stable alleles that retain tri-nucleotide repeats are often interrupted by point mutations that, as in the case of *Ap-fhc,* may stabilize the repeat tract and limit the size of the polyalanine region (Sezutsu and Yukuhiro, 2000).

Like *Fhc*, all of the spider silk genes sequenced to date show an extreme A/T bias (80–90%) in third position nucleotides and a preference for G/C in the first two codon positions (Xu and Lewis, 1990; Hayashi and Lewis, 1998). The polyalanine regions of the *MA* silks, however, are variable (4–10 alanine) and smaller than the PAB in *Ap-fhc.* Data on the possible PAB of *MA* silk are limited and based on cDNA alone. Therefore, one can only speculate that the poly-alanine sequences of *MA* are subject to less stringent functional constraint than are the PAB regions of *Ap-Fhc.* Alternatively, data on the entire *MA* gene and protein could reveal a larger order of organization that differs from that of *Ap-fhc*. We need to know the complete sequences of the *MA* fibroin gene as well as the 5′ flanking sequences (Sezutsu and Yukuhiro, 2000).

The coding regions of the *Ma2* gene and *Flag* exons contain multiple polymorphic sites, more than are found in the *Ma1* transcripts. Point mutations in the repetitive sequences of *Flag*, like point mutations in the polyalanine sequences in NPAB, may stabilize DNA during recombination; it may stabilize the mRNA transcript during transcription as well.

10. Differential regulation of gene expression and selective splicing may allow rapid adaptation of silk functional properties to different environments

Evolvability, the generation of variation, is a property of the genotype and distinct from the phenomenon of variation among individuals in populations (Wagner and Altenberg, 1996; Kirschner and Gerhard, 1998). Silk proteins are highly evolvable, as defined above, due to their highly repetitive amino acid and nucleotide sequences and the presence of apparent genetic hotspots. Despite their instability, yet, paradoxically, extreme conservation and convergence in spider fibroin sequences suggest that they are evolving under considerable functional constraint and conflict (Gatsey et al., 2001).

Orb-spinning spiders forage in all but the most extreme terrestrial habitats. If all silks were made up of only one protein new functional variants of the silk could result only from the slow accumulation of genetic differences across spider generations. Alternatively, because some spider silks are made up of more than one protein or characterized by an intron/exon molecular architecture, differential expression of protein type or selective splicing of mRNA transcripts could result in the rapid adaptation of silk proteins to new environments.

Recent field data seem to suggest that adult araneoid spiders may be able to adapt the composition of *MA* silk to the physical environments in which they forage through differential protein expression. For example, *Ma1* cDNA is largely composed of $(GA)_n$ and $(A)_n$ and a GGX motif. The alanine units, as discussed above, result in the crystalline, β-sheets. The repeated glycine motif and variable site is thought to form a 3_{10}-helix that links the crystalline regions together. This suggests that the *Ma1* mRNA transcript is largely responsible for silk strength. *Ma2* contains the repeating amino acid motifs GPGXX/GPGQQ, GP(SYG) (Table 1). A recent model for *Ma2* silks proposes that these motifs result in β-turns that, acting like springs, result in fiber elasticity. Cohesion of the spring-like structures is dependent on hydrogen bonding networks that form between spirals at their variable sites (Hinman and Lewis, 1992; Hayashi and Lewis, 1998). Therefore, because silk elasticity is attributed to the hydration of the *Ma2* encoded protein, the differential expression of *Ma1* and *Ma2* in different humidity environments could result in silks with different elasticity and strength that correlate with the relative humidity and the surface properties of the fiber. Differential expression of the *MA* peptides could allow an individual to produce silks with different functional properties in the same habitats or silks with the same functional properties in different physical habitats. In the extreme, *MA* silk made up solely of *Ma1* transcript would be strong but inelastic and silks made up solely of the *Ma2* transcript would be less strong but elastic. The flexibility afforded by the two-protein composition of spider *MA* silk as well as the exon/intron organization of *Flag* silk may be an important correlate of the burst of speciation that resulted with the emergence and rapid speciation of the Araneoidea.

One way to test the prediction of environmental flexibility in silk gene expression would be to compare the amino acid composition of the *MA* silks collected from spiders foraging in different sites and in particular to compare the relative proportions of serine, glycine, and proline residues in the total amino acid content of a silk. For example, the *Ma2* silk transcript for *N. clavipes* that has been sequenced to date contains 57 serine residues to 176 glycine residue (33%). *Ma1* silks contain 34 serine residues to 307 glycine residues (11%). Therefore, *MA* silk will contain three times more glycine to serine if *Ma2* and *Ma1* protein transcripts are produced in equal portions. Furthermore, the amount of proline present in an amino acid analysis would be linearly proportional to the presence of the *Ma2* transcript.

11. Summary

The review above examines the three types of molecular genetic organization of silk genes that have been found. The primitive Mesothelae and the Mygalomorphae produced silks that are fibrous, protein glues; according to a phylogenetic hypothesis, these evolved first in spiders. While the structure of the spider genes that encode their protein glues is unknown, it seems likely that they may be similar to that of the *Ser1* gene in *B. mori* and the Balbiani Ring (*BR*) genes in *Chironomus tentans* that encode protein glues produced by insects. *Ser1* and *BR-1* genes contain exons that encode only for crystal-forming regions and exons that encode only for non-crystalline material.

The second type of molecular genetic organization characterizes the gene for cocoon or fibroin silks produced by Lepidoptera larvae and probably *MA* or the dragline silk produced by spiders. The gene is made up of only two exons and one intron. The first exon is thought to be primarily regulatory in function, the second exon encodes for both crystalline and non-crystalline domains of a silk fibroin. This second exon is made up of repetitive elements that are hierarchically organized into crystalline and non-crystalline regions. The composition of the repetitive units is similar but their size is heterogeneous.

The third type of molecular genetic organization characterizes the *Flag* silk genes. Like *Fhc* and *MA* silk genes, *Flag* is hierarchically organized. Unlike the *Fhc* and *MA*, *Flag* is made up of multiple exons and introns. The introns are non-repetitive and highly similar to each other. The exons are repetitive, similar to each other, but different from the introns. The *Flag* exons do not contain the polyalanine, crystalline domains that make up the *Ap-Fhc* and *MA* silks. Nor do they contain the GAGAG(S) and GAGAGX regions that make up the crystalline regions of the *Bm-Fhc* silks. Furthermore, the exon/intron organization of *Flag* genes should allow greater opportunity for evolutionary innovation through selective splicing of the mRNA transcript.

References

Beckwitt, R., Arcidiacono, S., 1994. Sequence conservation in the C-terminal region of spider silk proteins (Spidroin) from *Nephila clavipes* (Tetragnathidae) and *Araneus bicentenarius* (Araneidae). J. Biol. Chem. 269, 6661–6663.

Beckwitt, R., Arcidiacono, S., Stote, R., 1998. Evolution of repetitive proteins: spider silks from *Nephila clavipes* (Tetragnathidae) and *Araneus bicentenarius* (Araneidae). Insect Biochem. Molec. Biol. 28, 121–130.

Burch, C.L., Chao, L., 2000. Evolvability of an RNA virus is determined by its mutational neighbourhood. Nature 406, 625–628.

Couble, P., Michaille, J.J., Garel, A., Couble, M.L., Prudhomme, J.C., 1987. Developmental switches of sericin mRNA splicing in individual cells of Bombyx mori gland. Dev. Biol. 124, 431–440.

Craig, C.L., 1997. The evolution of silks spun by arthropods. Annu. Rev. Ent. 42, 231–267.

Craig, C.L., Riekel, C., Herberstein, M.E., Weber, R.W., Kaplan, D., Pierce, N.E., 2000. Evidence for diet effects on the composition of silk proteins produced by spiders. Molec. Biol. Evol. 17, 1904–1913.

Gage, L.P., Manning, R.F., 1980. Internal structure of the silk fibroin gene of Bombyx mori II. Remarkable polymorphism of the organization of crystalline and amorphous coding sequences. J. Biol. Chem. 225, 9451–9457.

Garel, A., Deleape, G., Prudhomme, J.-C., 1997. Structure and organization of the *Bombyx mori* Sercin 1 gene and of the Sercins 1 deduced from the sequence of the Ser1B of cDNA. Insest Biochem. Molec. Biol. 27, 469–477.

Gatsey, J., Hayashi, C., Woods, J., Lewis, R., 2001. Extreme diversity, conservation and convergence of spider silk fibroin sequences. Science 291, 2603–2605.

Gosline, J.M., Guerette, P.A., Ortlepp, C.S., Savage, K.N., 1999. The mechanical design of spider silks: from fibroin sequence to mechanical function. J. Exp. Biol. 202, 3295–3303.

Grubb, D.T., Jelinski, L.W., 1997. Fiber morphology of spider silk: the effects of tensile deformation. Macromolecules 30, 2860–2867.

Grzelak, K., 1995. Control of expression of silk protein genes. Comp. Biochem. Physiol. B 110, 671–681.

Guerette, P.A., Ginzinger, D.G., Weber, B.H., Gosline, J.M., 1996. Silk properties determined by gland-specific expression of a spider fibroin gene family. Science 272, 112–114.

Hayashi, C., Lewis, R.V., 1998. Evidence from flagelliform silk cDNA for the structural basis of elasticity and modular nature of spider silks. J. Mol. Biol. 275, 773–778.

Hayashi, C.Y., Lewis, R.V., 2000. Molecular architecture and the evolution of a modular spider silk protein gene. Science 287, 1477–1479.

Hinman, M.B., Lewis, R.V., 1992. Isolation of a clone encoding a second dragline silk fibroin. J. Biol. Chem. 267, 19320–19324.

Iisuka, E., 1965. Degree of crystallinity and modulus relationships of silk threads from *Bombyx mori*. Biorheology 3, 1–8.

Ishiwaka, E., Suzuki, Y., 1985. Tissue and stage-specific expression of sericin genes in the midele silk gland of *Bombys mori*. Dev. Growth Diff. 27, 73–82.

Jeffreys, A.J., Wilson, V., Thein, S.L., 1985. Hypervariable 'minisatellite' regions in human DNA. Nature 314, 67–73.

Kirschner, M., Gerhard, J., 1998. Evolvability. Proc. Natl. Acad. Sci. 95, 8420–8427.

Komatsu, K., 1985. Chemical and structural studies on silk sericin.Proceedings of the 7th International Wool Textile Research Conferences, Tokyo, Vol. 1. pp. 373–382.

Kovoor, J., 1987. Comparative structure and histochemistry of silk-producing organs in arachnids. In: Nentwig, W., Heimer, S. (Eds.), The Ecophysiology of Spiders. Springer, New York, pp. 160–186.

Lam, S., Stahl, M., McMilin, K., Stahl, F., 1974. Rec-mediated recombination host spot activity in bacteriophage lambda II: a mutation which causes hot spot activity. Genetics 77, 425–433.

Michaille, J.J., Couble, P., Prudhomme, J.C., Garel, A., 1986. A single gene produces multiple sericin messenger RNAs in the silk gland of *Bombyx mori*. Biochimie 68, 1165–1173.

Michaille, J.-J., Garel, A., Prudhomme, J.-C., 1989. The expression of five middle silk gland specific genes is territorially regulated during the larval development of *Bombyx mori*. Insect Biochem. 19, 19–27.

Mita, K., Ichiumura, S., Zama, M., James, T.C., 1988. Specific codon usage patterns and its implications on the secondary structure of silk fibroin mRNA. J. Mol. Biol. 203, 917–925.

Mita, K., Ichimura, S., James, T.C., 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.

Paulson, G., Hoog, C., ernholm, 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.

Paulsson, G., Lendahl, U., Galli, G., Ericsson, C., Wieslander, L., 1990. The Balbiani Ring 3 gene in *Chironomus tentans* has a diverged repetitive structure split by many introns. J. Mol. Biol. 221, 331–349.

Prudhomme, J.C., Couble, P., Garel, J.P., Daillie, J., 1985. Silk synthesis. In: Kerkut, G.A., Gilbert, L.I. (Eds.), Comprehensive Insect Physiology, Biochemistry and Pharmacology, Vol. 10. Pergamon Press, Oxford, pp. 571–594.

Riekel, C., Vollrath, F., 2001. Spider silk fibre extrusion: combined wide- and small-angle X-ray microdiffraction experiments. Int. J. Biol. Macromol. 29, 203–210.

Riekel, C., Branden, C., Craig, C., Ferrero, C., Heidelbach, F., Müller, M., 1999. Aspects of X-ray diffraction on single spider fibers. Int. J. Biol. Macrom. 24, 187–195.

Riekel, C., Mueller, M., Vollrath, F., 1999. In situ X-ray diffraction during forced silking of spider silk. Macromolecules 32, 4464–4466.

Seldon, P., 1989. The orb-web weaving spiders in the early Cretaceous. Nature 340, 711–713.

Seldon, P., Gall, J., 1992. Palentology 35, 211.

Sezutsu, H., Yukuhiro, K., 2000. Dynamic rearrangement within the *Antheraea pernyi* silk fibroin gene is associated with four types of repetitive units. J. Mol. Evol. 51, 329–338.

Simmons, A.H., Michal, C.A., Jelinski, L.W., 1996. Molecular orientation and two-component nature of the crystalline fraction of spider dragline silks. Science 271, 84–87.

Souza, S.J.D., Long, M., Klein, R.J., Roy, S., Lin, S., Gilbert, W., 1998. Toward a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. Proc. Natl. Acad. Sci. USA 95, 5094–5099.

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

Thiel, G., Viney, C., 1997. J. Microsc. 185, 179–187.

Thiel, B., Kunkel, D., Viney, C., 1994. Physical and chemical microstructure of spider dragline: a study by analytical transmission electron microscopy. Biopolymers 34, 1089–1097.

Thiel, B., Guess, K., Viney, C., 1997. Non-periodic lattice crystals in the hierarchical microstructure of spider (major ampullate) silk. Biopolymers 41, 703–719.

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.

van Beek, J., Beaulieu, L., Schafer, H., Demura, M., Asakura, T., Meier, B., 2000. Solid-state NMR determination of the secondary structure of *Samic cynthia* ricini silk. Nature 405, 1077–1079.

Vollrath, F., Knight, D.P., 2001. Liquid crystalline spinning of spider silk. Nature 410, 541–548.

Wagner, G.P., Altenberg, L., 1996. Complex adaptations and the evolution of evolvability. Evolution 50, 967–976.

Xu, M., Lewis, R.V., 1990. Structure of a protein superfiber: spider dragline silk. Proc. Natl. Acad. Sci. 87, 7120–7124.

Yang, Z., Grubb, D., Jelinski, L., 1997. Small-angle X-ray scattering of spider dragline silk. Macromolecules 30, 8254–8261.

Zhou, C.-Z.F., Confalonieri, N., Medina, Y., et al., 2000. Fine organization of *Bombyx mori* fibroin heavy chain gene. Nucleic Acids Res. 28, 2413–2419.