

# Construction of Silk Fiber Core in Lepidoptera

František Sehnal\* and Michal Žurovec

Entomological Institute, Academy of Sciences, Branišovská 31, 370 05 České Budějovice, Czech Republic

Received October 14, 2003; Revised Manuscript Received December 2, 2003

The formation and properties of lepidopteran silk fibers depend on amino acid repeats in the principal protein, heavy chain fibroin (H-fibroin). In H-fibroins of the “bombycoid” type, concatenations of alanine or of the GAGAGS crystalline motifs (1st tier repeats) and adjacent sequences breaking periodicity make 2nd tier repeats. Two to six such repeats comprise a 3rd tier assembly, and 12 assemblies, linked by an amorphous sequence, constitute the repetitive H-fibroin region. Heterogeneity in the repeat length and intercalation of amorphous regions prevent excessive crystallization. In the “pyraloid” H-fibroins, iterations of simple motifs are absent and assemblies of several complex motifs constitute highly regular repeats that are organized in about 12 highest order reiterations without specific spacers. Repeat homogeneity appears crucial for the alignment and interaction of the disjunct motifs that must be registered precisely to form crystallites; repeat heterogeneity is associated with decreased fiber strength. Both H-fibroin types are typically hydrophobic, and their secretion requires disulfide linkage to light chain fibroin and participation of another protein, P25. These auxiliary proteins are absent in saturniid moths with amphiphilic H-fibroin repeats. The selection at nucleic acid and protein levels and the availability of nutrients play roles in H-fibroin evolution.

## Introduction

Silk fibers are produced from various types of ectodermal glands in the mites, spiders, and several groups of insects.<sup>1</sup> Commercial silk is obtained from the cocoons spun by certain caterpillars (larvae of moths and butterflies) before pupation. Until the discovery of nylon and other synthetic fiber polymers, the silk of domestic silkworm, *Bombyx mori*, was an economically and, at the time of war, also strategically important commodity. The mechanism of silk production in the paired labial gland of caterpillars and the structure of the spun-out filament have drawn the attention of biologists and mechanical engineers but neither of these two processes is fully understood.

From the technology of cocoon processing into silk thread, it has long been known that the raw silk filament contains a sericin coating soluble in hot alkaline water and a water insoluble fiber core, which is called fibroin. Fibroin was identified as a product of the posterior region of the gland, whereas sericin is produced in the middle region that also serves as silk reservoir.<sup>2</sup> The use of molecular biology techniques enabled detail analysis of silk composition in a handful of species.<sup>3</sup> Fibroin proved to be a polymer of a large protein that is usually associated with two small proteins. The sericin coating, which functions as glue during cocoon construction, consists of several proteins derived from two or more primary transcripts by the mechanisms of differential transcript splicing and protein glycosylation. The filament core and especially its coating are permeated with several small proteins that seem to provide protection against microbial degradation and animal digestion.

The transformation of liquid silk, which is stored in the silk glands, into a strong and elastic filament occurs during dope passage through a spinneret at ambient temperature and in aqueous environment.<sup>4</sup> Fiber polymerization under such gentle conditions is a challenge for industry and a puzzle for science. As a contribution to its unraveling, we summarize available information on the composition of the silk filament core and identify those structural features of its major component that are in our opinion indispensable for fiber formation. A brief note on silk evolution is added.

## Composition of the Silk Filament Core

Japanese researchers found that the silk core of *B. mori* (family Bombycidae, superfamily Bombycoidea) is composed of a heavy (H) and a light (L) fibroin chain that are linked by a disulfide bond.<sup>5</sup> The genes for both components were isolated and characterized.<sup>6,7</sup> Independent studies in France identified another silk core constituent that was named P25 after its deduced size.<sup>8</sup> L-Fibroin occurs in the silk as a single protein species of about 25 kDa, whereas P25 is glycosylated to 27 and 30 kDa products.<sup>9</sup> The necessity of disulfide linkage between L-fibroin and H-fibroin for their secretion was shown in mutants with defects in either of these components.<sup>10</sup> P25 was suggested to act as a chaperon that facilitates transport and secretion of the highly insoluble H-fibroin,<sup>11</sup> but conclusive evidence for this function is lacking. It was demonstrated that H-fibroin, L-fibroin, and P25 are assembled in elementary secretory units in the ratio 6:6:1.<sup>12</sup>

Identification of apparently homologous silk genes in *Galleria mellonella* (family Pyralidae, superfamily Pyraloidea) suggested that the formation of silk filament from H-fibroin, L-fibroin, and P25 is widespread in Lepidoptera.<sup>13</sup> The orthologues of *B. mori* H-fibroin gene were found in

\* To whom correspondence should be addressed. Telephone: +420-385310350. Fax: +420-385310354. E-mail: Sehnal@entu.cas.cz.



GAGAGAGAGAGTGS S GFGPYVANGGYSGY EYAWSS E SDFGTGS

GAGAGS GAGAGS GAGAGS GAGAGS GAGAGY GAGY GAGAGAGY GAGAGS GAGS  
GAGAGS GAGAGS GAGAGS GAGAGS GAGAGS GAGAGS GAGAGS GAGAGS GAGS G S GAGAGS GAGAGS GAGAGY GAGY GAGY G  
GY GAGAGAGY GAGAGS GAAS

GAGAGAGAGAGTGS S GFGPYVAHGGYSGY EYAWSS E SDFGTGS

GAGAGS GAGAGS GAGAGS GAGAGS GAGAGS GAGAGS GAGAGY GAGY GAGY GAA Y GAGAGAGY GAGAGS GAAS  
GAGAGS GAGAGS GAGAGS GAGAGS GAGAGS GAGAGS GAGAGS GAGAGS GAGAGS GAGAGS GAGAGY GAGAGAGY  
GAGAGS GAGS  
GAGAGS GAGAGS GAGAGS GAGAGS GAGAGS GAGS G S GAGAGS GAGAGS GAGAGY GAGY GAGY GAGY GAGAGAGY  
GAGAGS GAGS  
GAGAGS GAGAGY GAGAGAGY GAGY GAGAGAGY GAGAGT GAGS  
GAGAGS GAGAGS GAGAGS GAGAGS GAGAGS GAGAGS GAGS G S GAGAGS GAGAGS GAGAGS GAGAGS GAGAGS GA  
GAGY GAGAGAGY GAGY GAGAGAGY GAGAGS GAGS  
GAGAGS GAGAGS GAGAGS GAGAGY GAGY GAGAGS GAAS

GAGAGAGAGAGTGS S GFGPYVAHGGYSGY EYAWSS E SDFGTGS

GAGAGS GAGAGAGAGAGS GAGAGY GAGY GAGY GAGY GAGAGAGY GAGAGS GTGS  
GAGAGS GAGAGY GAGY GAGY GAGAGS GAAS  
GAGAGAGAGS GAGAGS GAGAGS GAGAGS GAGAGS GAGAGY GAGY GAGY GAGY GAGAGS GAAS  
GAGAGS GAGAGS GAGAGS GAGAGS GAGAGS GAGAGY GAGY GAGY GAGY GAGAGAGY GAGAGS GAAS  
GAGAGS GAGAGAGS GAGAGS GAGAGS GAGAGS GAGS GAGAGS GAGAGS GAGAGY GAGAGS GAAS

GAGAGAGAGAGTGS S GFGPYVANGGYSGY EYAWSS E SDFGTGS

**Figure 3.** Deduced amino acid sequence of *B. mori* H-fibroin. The representative sequence (between residues 1619 to 2642)<sup>22b</sup> is arranged to reveal characteristic repeat motifs that make up three highest order repeat assemblies. Uneven GAGAGS motifs are shadowed to accentuate regularity of their repetitions. Residues Tyr and Val that disturb regularity are printed in bold and underlined, and the repeat termini (typically GAAS) are double underlined. Noncrystalline (amorphous) spacers are shown in italics.

by grouping of polar residues amidst nonpolar residues, occur within this stretch. The homology is high between *G. mellonella* and *E. kuehniella*, which represent two subfamilies of Pyralidae, and low in *B. mori* and *A. pernyi*, presumably due to insertions/deletions and point mutations that occurred during evolution of Bombycoidea. The homologous part of the N-terminus ends with a short area rich in Ser and Ala. The following sequence of nearly 60 residues in *G. mellonella*, 11 in *E. kuehniella*, 25 in *B. mori*, and 75 in *A. pernyi* is diversified and seem to represent degenerate species specific repeats.

Regular repeats in the major central part of H-fibroin comprise 2400 residues in *A. pernyi*,<sup>14</sup> 5054 residues in *B. mori*,<sup>22</sup> and probably around 6000 residues in *G. mellonella*.<sup>17</sup> The nonrepetitive C-terminus includes 69 residues in *G. mellonella*, 64 in *E. kuehniella*, and 58 in *B. mori* (Figure 2B). About half of this sequence appears as degenerate repeats and is terminated with a Ser-rich area. The remainder is similar in all three species. It is rich in polar, primarily basic residues and contains three Cys in conserved positions. For *B. mori*, it was shown that the last two cysteines form an intramolecular disulfide bond, whereas the most upstream Cys makes a linkage to L-fibroin.<sup>22</sup> A very different carboxy terminus of just 34 residues was found in the H-fibroin of *A. pernyi* and other saturniids, in which the repetitive region ends with two modified repeats followed by a short and unique nonrepetitive sequence (Figure 2B). The content of neutral and basic polar residues is also high but the positions of Cys are different from those described above. The modified carboxy-end of saturniid H-fibroin apparently entails conformation changes consistent with the absence of L-fibroin and P25.

### Species Diversity of H-Fibroin Repeats

Each of the examined lepidopteran families is characterized by specific arrangement of H-fibroin repeats. The repetitive

nature of *B. mori* H-fibroin was deciphered as catenations of crystalline motif GAGAGS.<sup>24</sup> The length of such reiterations varies in different inbred silkworm lines.<sup>22,25</sup> In all cases, however, strings of several GAGAGS hexamers followed by a modified sequence form a second-order repeat, and several such repeats assemble in a third-tier repetition.<sup>26</sup> Sequencing of the whole gene<sup>22</sup> confirmed that the 2nd tier repeats contain a varying number of the GAGAGS copies. Their periodicity is disturbed by stretches of 10–28 residues where 2–6 Ala or Ser are replaced with Tyr (rarely Val). Such a stretch is followed by a single GAGAGS motif and by GAAS or a similar tetrapeptide that conclude each repeat (Figure 3). Two to six 2nd tier repeats and an amorphous sequence of 43 residues make up the third-order iteration. The amorphous sequences, each consisting of a chain of GA dipeptide followed by a unique sequence with charged residues, function as spacers breaking the repetitive region into 12 blocks (Figure 1). Both the length and the composition of spacers are conserved through the H-fibroin, except for the first and the last copies that contain a short insertion.

In Saturniidae such as *A. pernyi*, the repetitive region consists of four types of repeats that are composed of a relatively hydrophilic part (11–22 residues) and a hydrophobic string of 12–14 (exceptionally 5 and 15) alanines (Figure 4).<sup>14</sup> The repeats are designated here as types Y, W, R, and H to emphasize their characteristic residues. The Y and W repeats are similar in that both include a sequence GGYGSDS, which is preceded by GGY in the Y-type and by WGD in the W-type. There are slight variations in the composition and length of these repeats. Random combinations of 3–6 copies of these repeats, followed by one R- and one H-type repeat make one of the 12 highest-order repetitions (Figure 1). The fairly conserved tandem of R and H type repeats may be regarded as a spacer.

185 SGAGGSGGYGGYGS SAAAAAAAAAAAAA Y  
 GSGAGGSGGYGGYGGYGS SAAAAAAAAAAAAA Y  
 GSSAGGAGGGYGWGDGGYGS SAAAAAAAAAAAAA W  
 GSGAGGSGGYGGYGS SAAAAAAAAAAAAA Y  
 GSSAGGAGGGYGWGDGGYGS SAAAAAAAAAAAAA W  
 SSGAGGRDGGYGS SAAAAAAAAAAAAA R  
 RRAGHDRAAGSAAAAAAAAAAAAA H  
 SGAGGSGGYGWGDGGYGS SAAAAAAAAAAAAA Y  
 GSGAGGAGGGYGWGDGGYGS SAAAAAAAAAAAAA W  
 SGAGGSGGYGGYGS SAAAAAAAAAAAAA Y  
 GAGAGGAGGSYGWGDGGYGS SAAAAAAAAAAAAA W  
 GSGAGGRDGGYGS SAAAAAAAAAAAAA R  
 RRAGHDSAAGSAAAAAAAAAAAAA H  
 SGAGGSGGYGWGDGGYGS SAAAAAAAAAAAAA Y  
 GSGAGGAGGGYGWGDGGYGS SAAAAAAAAAAAAA W  
 SGARGSGGYGGYGS SAAAAAAAAAAAAA Y  
 GSGAGGVGGYGWGDGGYGS SAAAAAAAAAAAAA W  
 GSGAGGRDGGYGS SAAAAAAAAAAAAA R  
 RRAGHDSAAGSAAAAAAAAAAAAA H  
 SGAGGSGGYGWGDGGYGS SAAAAAAAAAAAAA W  
 GSGAGGAGGGYGWGDGGYGS SAAAAAAAAAAAAA W  
 SGARGSGGYGGYGS SAAAAAAAAAAAAA Y  
 GSGAGGVGGYGWGDGGYGS SAAAAAAAAAAAAA W  
 GSGAGGRDGGYGS SAAAAAAAAAAAAA R  
 RRAGHDRAAGSAAAAAAAAAAAAA H  
 SGAGGSGGYGWGDGGYGS SAAAAAAAAAAAAA W  
 SGAGGSGGYGGYGS SAAAAAAAAAAAAA Y

**Figure 4.** Deduced amino acid sequence of *A. pernyi* H-fibroin.<sup>14</sup> The representative sequence between residues 185 to 1453 is arranged to align the crystal and amorphous regions. Four types of repeats are marked as Y, W, R, and H, respectively (right column, italics), to emphasize their characteristic residues. The fairly conserved RH repeat tandem may be regarded as a spacer of higher order assemblies.

The repetitive region of H-fibroin in the Pyralidae family consists of long repeats that do not contain concatenations of simple motifs, have relatively low contents of Gly and Ala, and are rich in bulky residues.<sup>17,18</sup> Two to five types of repeats can be recognized in the examined species (Figure 5). The repeats designated as A-type include a conserved zone of 42–43 residues. The zone has a core identical or similar to PVIVIEEN that is bilaterally flanked by strings of Ser and Ala, and a terminal part where Gly alternates with bulky residues. Strings of tripeptides containing one or two Gly residues, e.g., GLG, GLN, GPY, GVS, etc., and short sequences composed of Ser and Ala make up the remaining part of the A-repeats and build up the other repeat types. In *G. mellonella*, H-fibroin includes highly conserved repeats A (63 amino acid residues), B<sub>1</sub> (43 residues), and B<sub>2</sub> (18 residues). About twelve assemblies AB<sub>1</sub>AB<sub>1</sub>AB<sub>1</sub>AB<sub>2</sub>-AB<sub>2</sub>(AB<sub>2</sub>) constitute the repetitive region (Figure 1). The assemblies are not flanked by special spacers (as is the case in *B. mori* and *A. pernyi*). Regular A<sub>1</sub>B repeat arrangements probably occupy a considerable part of the repetitive region in *E. kuehniella*, but the central and terminal parts of the region include combinations of diversified A<sub>1</sub> and A<sub>2</sub> repeats. A rather erratic arrangement of five repeat types was found in the H-fibroin of *Plodia interpunctella* (Figure 5).

#### Hydrophobic H-Fibroins Require Association with Auxiliary Proteins

The conversion of jelly dope into a solid fiber during silk spinning obviously requires change in H-fibroin conformation

**Table 1.** Percent Representation and Actual Numbers (in parentheses) of Hydrophilic Residues in the Nonrepetitive N- and C-Termini and the Repetitive Regions<sup>a</sup> of Secreted H-Fibroins

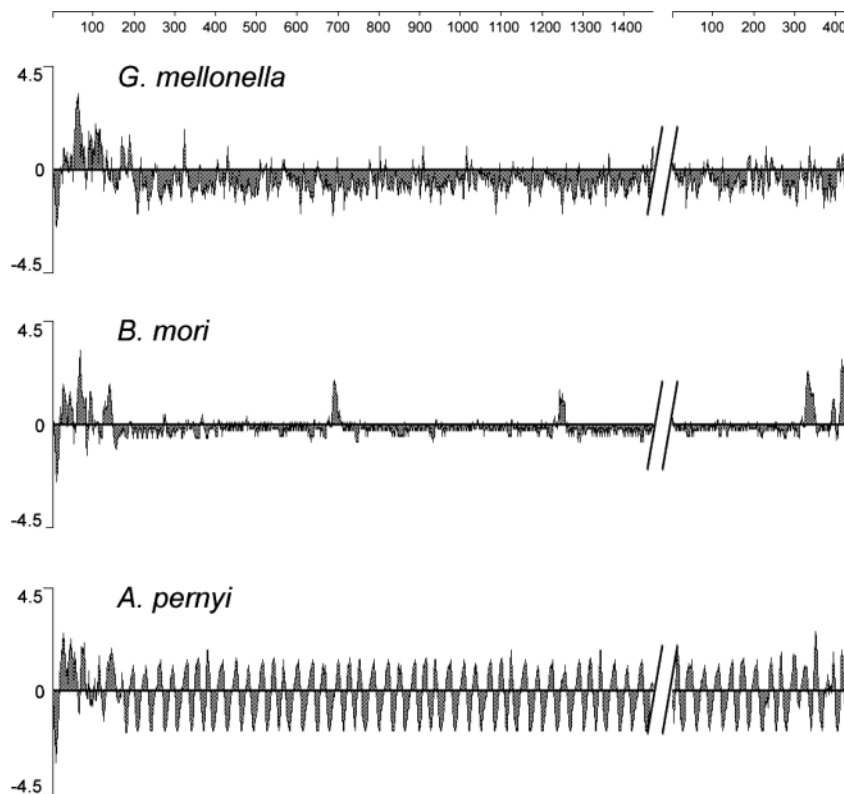
residues	N-terminus	repetitive region	C-terminus
<i>Galleria mellonella</i>			
acidic: D,E	14.5% (25)	2.6% (39)	1.4% (1)
basic: K,H,R	8.1% (14)	0.6% (9)	13.0% (9)
neutral: N,Q,S,T,Y,C	30.1% (52)	2.9% (43)	30.0% (20)
<i>Bombyx mori</i>			
acidic: D,E	15.8% (21)	0.6% (34)	1.7% (1)
basic: K,H,R	8.2% (11)	0.2% (8)	17.2% (10)
neutral: N,Q,S,T,Y,C	35.3% (47)	18.3% (920)	37.9% (22)
<i>Antheraea pernyi</i>			
acidic: D,E	11.5% (17)	4.6% (111)	5.9% (2)
basic: K,H,R	12.8% (19)	2.8% (69)	20.6% (7)
neutral: N,Q,S,T,Y,C	25.7% (38)	12.5% (300)	44.1% (15)

<sup>a</sup> Only about 20% of the repetitive region was sequenced in *G. mellonella*.

from a highly hydrated gel to a hydrophobic polymer. Acquisition of differently hydrated conformations probably depends on amphiphilicity that may be increased through association with auxiliary proteins. The N- and C-termini of H-fibroin contain a high proportion of hydrophilic residues, and many of them carry a charge, whereas the repetitive region includes a much lower proportion of polar residues (Table 1). The termini are short and hardly influence general solubility of H-fibroin but their conservation indicates that they play some role. One possibility is that they represent sites of H-fibroin interaction with ions. For *B. mori*, it was shown that hydrated fibroin molecules combine into large protein complexes (elementary units) which aggregate into yet larger assemblies connected by divalent ions. As H-fibroin moves through the gland lumen, the elementary units elongate, possibly due to changes in the ionic environment.<sup>27</sup> Ion interactions could occur via acidic residues that are conserved in the N-terminus of all H-fibroins analyzed so far.

Decisive for H-fibroin solubility is the distribution of hydrophilicity in its huge repetitive region. The region was examined in sufficient detail in just three species and the results reveal that there are two principal types: one with prevailing hydrophobicity and the other with alternating hydrophobic and hydrophilic areas (Figure 6). The former type was found in the 500 kDa H-fibroin of *G. mellonella*<sup>17</sup> and 391 kDa H-fibroin of *B. mori*.<sup>22</sup> Virtual lack of hydrophilic regions precludes peptide chain folding that would expose a surface prone to hydration. These H-fibroins may probably be hydrated only with the aid of L-fibroin and P25 that are indispensable components of the silk fiber core in the respective species. L-fibroin is predominantly hydrophobic,<sup>7,19</sup> but formation of an internal disulfide loop and the disulfide bond to H-fibroin are likely to expose the hydrophilic N-terminus of the molecule. Glycosylation of this part of L-fibroin in *P. xuthus*<sup>20</sup> further increases hydrophilicity. P25 consists of alternating hydrophobic and hydrophilic areas with considerable number of charged residues. The solubility of mature protein is enhanced by glycosylation.<sup>28</sup> For *B. mori*, it was demonstrated that association of P25 with H-fibroin is based on hydrophobic interactions,<sup>9</sup> probably allowing exposure of the hydrophilic





**Figure 6.** Hydrophobicity/hydrophilicity plots of the representative parts of H-fibroins generated by Protean (Lasergene) software. The first 194 (*G. mellonella*),<sup>17</sup> 168 (*B. mori*),<sup>22b</sup> and 137 (*A. pernyi*)<sup>14</sup> residues and the last 69, 58, and 34 residues, respectively, represent the nonrepetitive terminal regions. The scale bar depicts the number of amino acid residues.

$\beta$ -sheets,<sup>33</sup> but their presence causes local disorders in crystallinity.<sup>35</sup> This may be expected to increase fiber extensibility because the altered  $\beta$ -sheet distance in the crystallites, which is maintained by a low number of interacting Tyr residues, allows a split-up at a force much smaller than needed for disintegration of the standard Gly/Ala  $\beta$ -sheet stacks. Fiber extensibility probably depends on the amorphous spacers that break the repetitive region into blocks of different sizes. The high content (39.5%) of polar residues in the spacers makes their interactions with the GAGAGS motifs unlikely. The spacers thus function as flexible linkers connecting the crystalline regions.

In the *Antheraea* species, H-fibroin  $\beta$ -sheets and crystallites are primarily built by the strings of alanines (Figure 4). The Gly rich sequence in the remaining part of each repeat can possibly also make crystallites but their formation is distorted by the variability of residues alternating with Gly. Some of the Ser and most of the Tyr residues present at transition between the Gly rich and the poly-Ala regions have a random coil conformation<sup>36</sup> and disrupt crystallization. The interaction of repeats during crystallization is also impaired by variations in their lengths and their numbers in the highest order repetitive blocks (Figure 1). The repeat H, which occurs at the end of each block, consists of a unique sequence of 11 residues and a string of 13 Ala. The unique sequence includes 1–2 neutral and 4–5 charged polar residues. This noncrystalline amorphous region apparently provides fiber flexibility.

In the H-fibroin of pyralid moths, regions rich in Ala and Ser at the beginning of repeat A and elsewhere (Figure 5) are likely to form crystallites that were discovered in the

silk of *G. mellonella* with X-ray diffraction.<sup>31</sup> The involvement of Ser in crystallite building is likely in view of the finding that the replacement of some Ala with Ser residues in a synthetic GlyAla polypeptide chain has little impact on the energy minimum of peptide crystallization.<sup>37</sup> On the other hand, amino acids with bulky side chains disrupt formation of crystalline structures.<sup>38</sup> It is noteworthy that the crystalline Ser-rich chain in the A repeats is broken with a noncrystalline sequence of four hydrophobic and several charged residues, such as EVIVIDDR in *G. mellonella*, PVVIEENQ in *E. kuehniella*, and PVVVVEEN or PVIIEED in *P. interpunctella*. This sequence may be regarded as an amorphous motif and is probably important for fiber flexibility. The strings of GGX and GXX motifs, in which Gly alternates with bulky residues (Figure 5), might also form  $\beta$ -sheets, but the diversity and large size of X residues impedes tripeptide registration into crystallites. If formed, these crystallites are likely to split at a low force. Other arrangements are possible; in a spider silk, regular GGX iterations are believed to form a  $3_{10}$ -helix.<sup>39</sup> All possible interactions of dispersed motifs seem to depend on the regularity of their distribution in the reiterated repeats (see below).

#### Structural Provisions for Fiber Strength and Flexibility

The tensile strength of silk is attributed to the presence of stiff crystals linked together by amorphous flexible chains.<sup>40,41</sup> Crystalline motifs identified in the H-fibroins of *B. mori*, *A. pernyi*, and *G. mellonella* occupy differently large areas and should therefore generate fibers of different strength. For the sake of simplicity, we postulate that four adjacent Ala and

Ser residues or two dipeptides GA or GS are minimal requirements for the crystallization of  $\beta$ -sheets. The percentage of residues in such groupings is about 85% in the repeats of *B. mori*, 41% in *A. pernyi*, and only 12% in *G. mellonella*. This would indicate that *B. mori* silk contains a higher proportion of crystallites and is therefore much stronger than the other silks but the strength of all silks is similar.<sup>40,42</sup> Hence, it is likely that crystallites are formed by additional peptide motifs than those considered in our analysis and that other conformations than crystallites contribute to the fiber strength.

More research is needed to unravel conformations of diverse H-fibroin motifs in the gellike and the fibrous forms of silk. The diversity of H-fibroin repeats in just three lepidopteran families (Figures 3-5) shows that fiber polymerization may be based on surprisingly different sequences. Analysis of known H-fibroins indicates that the nature of interacting motifs, which determine the secondary protein conformation, is crucial, but the mode of motif dispersion in higher order repeats and the regularity of both short and long reiterations are also of importance. The distribution of interacting motifs, which is related to the length of higher order repeats, controls the likelihood of their interactive matching. All repeat types in the H-fibroin of *A. pernyi* are 24 to about 30 residues long and include a polyalanine motif. The shortest 2nd tier repeats in *B. mori* are of analogous size, whereas the longest ones consist of 108 residues. Iterations of simple motifs are lacking in the pyralid H-fibroin repeats. The shortest repeats in *P. interpunctella* include less than 20 residues but a length exceeding 50 residues is more typical. The AB<sub>1</sub> repeats in *G. mellonella* comprise 105, and the A<sub>1</sub>B repeats in *E. kuehniella* 114 residues. The spans of crystalline sequences within each repeat are species specific but the extents of the intervening, less crystalline regions are similar. In *B. mori*, a string of GAGAGS motifs occupies most of the repeat length, but close to the repeat end, it is interrupted by a Tyr-rich sequence that is mostly 21 residues long. Similarly, the polyalanine chains in the H-fibroin of *A. pernyi* are 11–21 residues apart, and the poorly defined AlaSer-rich motifs in the H-fibroin of pyralids are separated by 6–30 residues.

Structural analysis of H-fibroins suggests that two repeat arrangements are compatible with proper crystallite formation. The repeats made of strings of simple motifs must contain sequences preventing excessive formation of the  $\beta$ -sheets and crystallites, whereas polymerization of H-fibroins with widely distributed, short, and structurally complex motifs is possible only when high conservation of the repeats facilitates registration of the interacting motifs. The first type of H-fibroin occurs in *B. mori* and *A. pernyi* and is therefore called “bombycoid”, whereas the second “pyraloid” type is characteristic for the pyralid moths. If all GAGAGS motifs in *B. mori* and all alanines in *A. pernyi* H-fibroins built crystallites, the silk would become a solid rod rather than a flexible fiber. However, excessive H-fibroin crystallization in *B. mori* is prevented by the unequal repeat length and by the disturbance of the GAGAGS monotony with the insertions of Tyr or Val. In the H-fibroin of *A. pernyi*, the combination of several repeat types seems to play

a similar role. Fiber flexibility depends on spacers of 54 residues in *B. mori* and 43 (sum of the R- and H-type repeats) in *A. pernyi*, which break the monotonic repeat catenations into 12 blocks. The “pyraloid” H-fibroins are characterized by motifs that are unlikely to build large compact crystallites. They are orderly distributed in long repeats without long intercalated spacers; “crystalline” motifs imparting fiber strength are indistinct and intermingled with “amorphous” charged regions conferring flexibility. The repeats are extremely uniform indicating that motif interactions providing for H-fibroin polymerization depend on their precise alignment. Indirect evidence from the gene mapping in *G. mellonella* suggests that the highest order repeat assembly in about 12 blocks is conserved.<sup>17</sup>

All putatively interacting sequences in the “pyraloid” H-fibroin are short, and their alignment for polymerization is facilitated by high conservation of the repeats. The measurements of tensile strength demonstrated that the regularity of repeats is crucial for the silk properties.<sup>18</sup> The strong silks of *G. mellonella* and *E. kuehniella* contain H-fibroins with 2–3 types of highly conserved repeats of about 50–114 residues, whereas the feeble silk of *P. interpunctella* is based on H-fibroin with at least five types of erratic repeats 18–49 residues long (Figure 5).

The presence of about 12 highest order reiterations in all examined species suggests that it is of some significance for gene stability or for fiber formation. For example, such an arrangement may foster polymerization by facilitating superposition of the motifs located in distant protein regions. The noncrystalline spacer sequences (also called amorphous regions) in the “bombycoid” H-fibroin restrain crystallization and thereby control fiber rigidity. Their conformation is not known, but it is likely that it confers fiber elasticity. Interestingly, the only fully analyzed spider silk also includes an ensemble of 10 repeats, each of them about 440 residues long<sup>43</sup> (highest order repeats in the examined Lepidoptera include from about 180 to nearly 400 residues).

### Evolutionary Aspects

The aforementioned differences in the H-fibroin repeats among relatively closely related species demonstrate that the repetitive region of *H-fibroin* genes undergoes rapid evolution. Early X-ray diffraction studies of various silks led to a conclusion that the extent and speed of molecular silk diversification are greater than in any other fibrous protein,<sup>44</sup> possibly because functional requirements placed on silk have been reduced to the formation of “a satisfactory mechanical fiber” and the range of structures is not limited by other extracellular or intracellular constraints.<sup>32</sup> Fiber formation may be based on diverse peptide sequences, but the selection for mechanical properties and other factors, such as energetic cost of different amino acids, often drive convergent evolution of similar sequences. For example, although silk production evolved independently in the labial glands of caterpillars and the epidermal abdominal glands of spiders,<sup>45</sup> convergent evolution led to polyalanine motifs in the H-fibroins of *Antheraea* moths (Figure 4) as well as in a number of spider silks.<sup>46</sup>

The repetitive nature of *H-fibroin* genes makes them prone to crossing-over recombination that results in repeat homogenization within the gene and divergence when a species splits into two or more genetically isolated lines. The homogenization of repeats, which is known from various types of repetitive DNA (microsatellites, repetitive regions of mitochondrial DNA, tandem gene arrays etc), including genes encoding spider silk proteins,<sup>47</sup> is also referred to as concerted evolution or gene conversion.<sup>48</sup> It ensures that an individual member of a repeat family does not evolve independently of the other family members. The mechanisms of gene conversion are responsible for fast sequence changes by promoting spread of a variant repeat to all family members.<sup>49</sup> The interspecies divergence at the level of basic motifs apparently starts as a point mutation that is followed by propagation of the altered motif throughout the repetitive sequence. Variations in the length of higher order repeats are probably caused by polymerase slippage during DNA replication or by misalignment during the crossing-over.<sup>50</sup> The presence of internally nonrepetitive amorphous “spacers” restricts both the replication slippage and the misalignment and thereby stabilizes the gene. The extent and the velocity of changes in the repeat structure seem to vary in different fibroin types for reasons that are poorly understood.

The “bombycoid” type of H-fibroin is characterized by simple concatenations of alanines or oligopeptides (Figure 3), in contrast to the complex repeats in the “pyraloid” H-fibroin of *G. mellonella* and *E. kuehniella* (Figure 5). The 2nd and 3rd tier repeats in *B. mori* H-fibroin vary substantially in length. Similar length variation occurs in the incomplex repeats of *A. pernyi* H-fibroin<sup>14</sup> and in the silk of the *Dolomedes*<sup>46</sup> and *Nephila clavipes*<sup>39,51</sup> spiders. As mentioned earlier in this article, length irregularities in the repeats composed of crystalline motifs limit the extent of crystallization and thereby reduce rigidity of the silk fiber. By contrast, repeats composed of several motif types in the “pyraloid” H-fibroins in *G. mellonella* and *E. kuehniella*, as well as the similarly complex repeats encoded by the silk cDNA1 and cDNA4 of the spider *Plectreurys*,<sup>50</sup> are characterized by remarkable length conservation. The maintenance of the uniform repeat length might be at least partially explained by existence of large DNA units that undergo crossing-over as discrete entities. DNA blocks without intrinsic repeats are less likely to be internally misaligned during the process of concerted evolution.

There is no doubt that silk production is an ancestral feature of Lepidoptera. However, fast diversification of the H-fibroin gene impedes tracking down its early evolutionary history and the origins of the “bombycoid” and “pyraloid” repeat types in the species analyzed so far. The two H-fibroin types certainly do not represent rigidly and irreversibly separated lineages. Type conversion and formation of intermediate types probably occurred repeatedly in Lepidoptera phylogeny due to diversification of the tandem arrays of simple motifs in the “bombycoid” type and shortening of the long unique repeats in the “pyraloid” type.

Lepidopteran and spider silk proteins are composed primarily of three amino acids: Gly, Ala, and Ser. The dominance of Gly and Ala, which are encoded by nucleotide

triplets GGX and GCX, results in high GC content in the *H-fibroin* gene and mRNA. In all examined lepidopteran species, this GC abundance is partly compensated for by preferred use of A and T nucleotides in the third codon position. Maintenance of a balance between GC and AT nucleotides probably restricts codon usage in all *H-fibroin* genes. It was proposed that the restriction is dictated by the most stable conformations of the DNA or its transcript.<sup>52</sup> Codon usage restrictions restrain gene diversification and foster inclusion in H-fibroin of amino acids encoded by suitable codons. Hence, the presence of large amino acids in the “pyraloid” H-fibroin (Figure 5) might be in part an outcome of selection for gene stability. It cannot be excluded that convergent evolution of the motifs such as GPG(X)<sub>n</sub>, which occur in the pyralid moths and certain spider silks,<sup>46</sup> was also initiated by forces stabilizing the corresponding genes.

Intraspecies homogenization and interspecies diversification of the repeats occur and are partly selected for at the nucleic acid level, but the final and decisive sorting between the promoted and the condemned modifications is based on the amino acid sequence of the protein product. H-fibroin must meet functional requirements such as to provide certain strength and elasticity to the silk fiber.<sup>17,18</sup> The selection for fiber properties, however, occurs under certain nutrient supply conditions that are likely to affect silk composition as has been proposed for the spiders.<sup>53</sup> Lepidoptera must receive in their food amino acids Arg, His, Ile, Leu, Lys, Met, Thr, and possibly also Phe, Tyr, and Trp,<sup>54</sup> and the synthesis of the remaining amino acids requires different energy inputs. The impacts of diet and spinning habits on the H-fibroin structure can be illustrated on a comparison of *B. mori* with *G. mellonella*. The first species feeds exclusively on a relatively poor diet of the mulberry leaves and makes a large biomass investment into the cocoon. It is not surprising that *B. mori* H-fibroin consists mostly of the “inexpensive” residues Gly, Ala, and Ser, which are for the most part synthesized in the silk glands, and Tyr, which is derived from other larval tissues.<sup>55</sup> By contrast, the caterpillars of *G. mellonella*, which produce silk with considerable proportion of bulky residues, feed on the bee comb that may be short in certain amino acids but is always very rich energetically. The caterpillars normally spin large amounts of silk to construct protective tubes and much less silk for the cocoons. Tube spinning is abandoned under conditions of limited protein supply.<sup>56</sup> Selection pressure for the use of simple amino acids was probably much lower in the evolution of *G. mellonella* than in *B. mori*.

**Acknowledgment.** We thank Dr. Catherine L. Craig of Harvard University for critical reading of the manuscript. Our research was supported by Grant A5007402 from Grant Agency of the Czech Academy of Sciences.

## References and Notes

- (1) Craig, C. L. *Annu. Rev. Entomol.* **1997**, *42*, 231.
- (2) Sehnal, F.; Akai, H. *Int. J. Insect Morphol. Embryol.* **1990**, *19*, 79.
- (3) Fedič, R.; Žurovec, M.; Sehnal, F. *J. Insect Biotech. Sericol.* **2002**, *71*, 1.
- (4) Vollrath, F.; Knight, D. P. *Nature* **2001**, *410*, 541.

- (5) Shimura, K.; Kikuchi, A.; Ohtomo, K.; Katagata, Y.; Hyodo, A. *J. Biochem.* **1976**, *80*, 693.
- (6) Suzuki, Y.; Brown, D. D. *J. Mol. Biol.* **1972**, *63*, 409.
- (7) Yamaguchi, K.; Kikuchi, Y.; Takagi, T.; Kikuchi, A.; Oyama, F.; Shimura, K.; Mizuno, S. *J. Mol. Biol.* **1989**, *210*, 127.
- (8) (a) Couble, P.; Moine, A.; Garel, A.; Prudhomme, J. C. *Dev. Biol.* **1983**, *97*, 398. (b) Chevillard, M.; Couble, P.; Prudhomme, J. C. *Nucleic Acids Res.* **1986**, *14*, 6341.
- (9) Tanaka, K.; Mori, K.; Mizuno, S. *J. Biochem.* **1993**, *114*, 1.
- (10) Takei, F.; Kikuchi, Y.; Kikuchi, A.; Mizuno, S.; Shimura, K. *J. Cell Biol.* **1987**, *105*, 175.
- (11) Tanaka, K.; Inoue, S.; Mizuno, S. *Insect Biochem. Mol. Biol.* **1999**, *9*, 269.
- (12) Inoue, S.; Tanaka, K.; Arisaka, F.; Kimura, S.; Ohtomo, K.; Mizuno, S. *J. Biol. Chem.* **2000**, *275*, 40517.
- (13) Žurovec, M.; Sehnal, F.; Scheller, K.; Kumaran, A. K. *Insect Biochem. Mol. Biol.* **1992**, *22*, 55.
- (14) Sezutsu, H.; Yukuhiro, K. *J. Mol. Evol.* **2000**, *51*, 329.
- (15) Hwang, J. S.; Lee, J. S.; Goo, T. W.; Yun, E. Y.; Lee, K. S.; Kim, Y. S.; Jin, B. R.; Lee, S. M.; Kim, K. Y.; Kang, S. W.; Suh, D. S. *Biotechnol. Lett.* **2001**, *23*, 1321.
- (16) Datta, A.; Ghosh, A. K.; Kundu, S. C. *Insect Biochem. Mol. Biol.* **2001**, *31*, 1013.
- (17) Žurovec, M.; Sehnal, F. *J. Biol. Chem.* **2002**, *277*, 22639.
- (18) Fedič, R.; Žurovec, M.; Sehnal, F. *J. Biol. Chem.* **2003**, *278*, in press.
- (19) (a) Žurovec, M.; Vašková, M.; Kodrlik, D.; Sehnal, F.; Kumaran, A. K. *Mol. Gen. Genet.* **1995**, *247*, 1. (b) Yang, C.; Teng, X.; Žurovec, M.; Scheller, K.; Sehnal, F. *Gene* **1998**, *209*, 157.
- (20) Tanaka, K.; Mizuno, S. *Insect Biochem. Mol. Biol.* **2001**, *31*, 665.
- (21) Tamura, T.; Inoue, H.; Suzuki, Y. *Mol. Gen. Genet.* **1987**, *206*, 189.
- (22) (a) Zhou, C. Z.; Confalonieri, F.; Medina, N.; Zivanovic, Y.; Esnault, C.; Yang, T.; Jacquet, M.; Janin, J.; Duguet, M.; Perasso, R. *Nucleic Acid Res.* **2000**, *28*, 2413. (b)
- (23) Tanaka, K.; Kajiyama, N.; Ishikura, K.; Waga, S.; Kikuchi, A.; Ohtomo, K.; Takagi, T.; Mizuno, S. *Biochim. Biophys. Acta* **1999**, *1432*, 92.
- (24) (a) Fraser, R. D. B.; MacRae, T. P.; Steward, F. H. C. *J. Mol. Biol.* **1966**, *19*, 580. (b) Gage, L. P.; Manning, R. F. *J. Biol. Chem.* **1980**, *255*, 9444.
- (25) Manning, R. F.; Gage, L. P. *J. Biol. Chem.* **1980**, *255*, 9451.
- (26) Mita, K.; Ichimura, S.; James, T. C. *J. Mol. Evol.* **1994**, *38*, 583.
- (27) Zhou, C. Z.; Confalonieri, F.; Jacquet, M.; Perasso, R.; Li, Z. G.; Janin, J. *Proteins Struct., Funct., Genet.* **2001**, *448*, 119.
- (28) Hossain, K. S.; Ochi, A.; Ooyama, E.; Magoshi, J.; Nemoto, N. *Biomacromolecules* **2003**, *4*, 350.
- (28) (a) Chevillard, M.; Deleage, G.; Couble, P. *Sericologia* **1986**, *26*, 435. (b) Žurovec, M.; Kodrlik, D.; Yang, C.; Sehnal, F.; Scheller, K. *Mol. Gen. Genet.* **1998**, *257*, 264.
- (29) Marsh, R. E.; Corey, R. B.; Pauling, L. *Biochim. Biophys. Acta* **1955**, *16*, 1.
- (30) Marsh, R. E.; Corey, R. B.; Pauling, L. *Acta Cryst.* **1955**, *8*, 710.
- (31) Warwicker, J. O. *J. Mol. Biol.* **1960**, *2*, 350.
- (32) Lucas, F.; Rudall, K. M. In *Comprehensive Biochemistry*; Florin, M., Stotz, E. H., Eds.; Elsevier: Amsterdam, 1968; 26B.
- (33) Asakura, T.; Yao, J. *Protein Sci.* **2002**, *11*, 2706.
- (34) Iisuka, E. *Biorheology* **1965**, *3*, 1.
- (35) Asakura, T.; Sugino, R.; Okumura, T.; Nakazawa, Y. *Protein Sci.* **2002**, *11*, 1873.
- (36) Nakasawa, Y.; Asakura, T. *Macromolecules* **2002**, *35*, 2393.
- (37) Kaplan, D. L.; Mell, C. M.; Arcidiacono, S.; Fossey, S.; Senecal, K.; Muller, W. *Protein-Based Materials*; McGrath, K., Kaplan, D., Eds.; Birkhäuser: Boston, 1997.
- (38) Simmons, A.; Ray, E.; Jelinski, L. *Macromolecules* **1994**, *27*, 235.
- (39) Xu, M.; Lewis, R. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 7120.
- (40) Denny, M. W. *The Mechanical Properties of Biological Materials. Symposia of Society for Experimental Biology*; Cambridge University Press: Cambridge, U.K., 1980; Vol. 34.
- (41) Gosline, J. M.; DeMont, M. E.; Denny, M. W. *Endeavour* **1986**, *10*, 37.
- (42) Hepburn, H. R.; Chandler, H. D.; Davidoff, M. R. *Insect Biochem.* **1979**, *9*, 69.
- (43) Hayashi, C. Y.; Lewis, R. V. *J. Mol. Biol.* **1998**, *275*, 773.
- (44) Rudall, J. C.; Kenchington, W. *Annu. Rev. Entomol.* **1971**, *16*, 73.
- (45) Schultz, J. *Biol. Rev.* **1987**, *62*, 89.
- (46) Gatesy, J.; Hayashi, C.; Motriuk, D.; Woods, J.; Lewis, R. *Science* **2001**, *291*, 2603.
- (47) Hayashi, C. Y.; Lewis, R. V. *Science* **2000**, *287*, 1477.
- (48) Zimmer, E. A.; Martin, S. L.; Beverley, S. M.; Kan, Y. W.; Wilson, A. C. *Proc. Natl. Acad. Sci. U.S.A.* **1980**, *77*, 2158.
- (49) Dover, G. A. *Nature* **1982**, *299*, 111.
- (50) Ishimura, S.; Mita, K. *J. Mol. Evol.* **1992**, *35*, 123.
- (51) Jensen, L. M.; Zhang, Y.; Shooter, E. M. *J. Biol. Chem.* **1992**, *267*, 19325.
- (52) Mita, K.; Ichimura, S.; Zama, M.; James, T. C. *J. Mol. Evol.* **1988**, *203*, 917.
- (53) Craig, C. L.; Riekel, C.; Herberstein, M. E.; Weber, R. S.; Kaplan, D.; Pierce, N. E. *Mol. Biol. Evol.* **2000**, *17*, 1904.
- (54) Nation, J. L. *Insect Physiology and Biochemistry*; CRC Press: Boca Raton, FL, 2002.
- (55) Prudhomme, J. C.; Couble, P.; Garel, J. P.; Daillie, J. In *Comprehensive Insect Physiology, Biochemistry and Pharmacology*; Kerkut, G. A., Gilbert, L. I., Eds.; Pergamon Press: Oxford, U.K., 1985; Vol. 10.
- (56) Jindra, M.; Sehnal, F. *J. Insect Physiol.* **1989**, *35*, 719.

BM0344046