The fiber properties of lepidopteran silk depend on the amino acid repeats that interact during H-fibroin polymerization. The aim of our research was to relate repeat composition to insect biology and fiber strength. Representative regions of the H-fibroin genes were sequenced and analyzed in three pyralid species: wax moth (Galleria mellonella), European flour moth (Ephestia kuehniella), and Indian meal moth (Plodia interpunctella). The amino acid repeats are species-specific, evidently a diversification of an ancestral region of 43 residues, and include three types of regularly dispersed motifs: modifications of GSSAASAA sequence, stretches of tripeptides GXZ where X and Z represent bulky residues, and sequences similar to PVIVIEE. No concatennations of GX dipeptide or alanine, which are typical for Bombyx silkworms and Antheraea silk moths, respectively, were found. Despite different repeat structure, the silks of G. mellonella and E. kuehniella exhibit similar tensile strength as the Bombyx and Antheraea silks. We suggest that in these latter two species, variations in the repeat length obstruct repeat alignment, so sufficiently long stretches of iterated residues get superposed to interact. In the pyralid H-fibroins, interactions of the widely separated and diverse motifs depend on the precision of repeat matching; silk is strong in 

Silks have long attracted attention as biodegradable fibers of considerable strength, elasticity, and durability. X-ray diffraction studies revealed that the fibers contain crystallites, interpreted to be stacks of β-sheets formed by regular amino acid repeats in the core protein called fibroin (1). The highly organized crystalline regions were suggested to confer the toughness and the interspersed “amorphous” regions the elasticity of the silk fiber. The distance between β-sheets in the silk crystallites of Bombyx mori (family Bombycidae) corresponded to hydrogen bridges between side chains of glycine and alanine residues whereas in the silk of Antheraea pernyi (family Saturniidae) to interactions between alanine-alanine residues (2, 3). These conclusions were consistent with the total amino acid composition of the respective silk.

Modern protein and gene analyses characterized the structure of silk in detail (4). It was confirmed that the fiber core consists primarily of a large (200–500 kDa) protein, designated heavy chain fibroin (H-fibroin),1 that is associated with two small proteins, light chain fibroin and P25, in most Lepidoptera. Several sericins enveloping the fiber core serve as glue for the cocoon construction, and a few additional peptides seem to confer protective functions. Most of the physical fiber properties appear to be determined by the amino acid sequence of the H-fibroin, which is for the most part composed from repeats. The repeats in B. mori H-fibroin are dominated by iterations of a GAGAS motif (5, 6), and a string of alanines occurs at the end of each of the four types of repeats in Antheraea H-fibroin (7, 8).

On the basis of x-ray diffraction data (1), we expected to find polyalanine repeats in the H-fibroin of the waxmoth Galleria mellonella (family Pyralidae), but gene analysis revealed a more complex amino acid arrangement (9). The overall gene structure and the nucleotide sequences at the 5′ and 3′ gene ends were similar to the H-fibroin genes of B. mori (6), A. pernyi (7), and Antheraea yamamai (8), but the major internal sequence was unique. About 95% of the gene encoded amino acid repeats arranged hierarchically into 10–12 large assemblies. The remarkable homogeneity of the repeat length and composition contrasted with repeat variability in the H-fibroins of Bombyx and Antheraea (6–8). Repeat regularity, absence of concatenations of short motifs, and relatively high content of amino acid residues with bulky side chains indicated that organization of the H-fibroin polymer in G. mellonella differs from that in Bombyx and Antheraea (9).

The silk of G. mellonella was reported to be considerably more extensible and somewhat stronger than the silks of B. mori and A. pernyi (10, 11). These properties are probably related to the mode of silk use in G. mellonella. The caterpillars of B. mori and A. pernyi spin silk only for cocoon construction at the end of larval life, and high tensile strength is the major functional requirement of the fiber. From an early larval stage, the caterpillars of G. mellonella, which develop in bee colonies, spin silky tubes that protect them against the detection and killing by the bees. They continuously renew and enlarge the

1 The abbreviations used are: H-fibroin, heavy-chain fibroin; RT, reverse transcription; nt, nucleotide.
tubes, devouring some of the old silk to utilize amino acids that are in short supply in their comb diet. Before pupation, the larva can descend on a newly spun fiber to leave the bee colony in search for a suitable place for cocoon construction. The silk produced by G. mellonella must obviously be extensible to allow widening of the tube when the larva turns around and at the same time strong enough to hold the larval weight. It was deduced that the uniformity of H-fibroin repeats underlies the excellent mechanical properties of G. mellonella silk (9).

G. mellonella is a member of Galleriinae subfamily of the Pyralidae moth family. In the present study we examine correlation between the regularity of H-fibroin repeats and the silk properties in two pyralids that belong to subfamily Phyctinae (12). Larvae of the Mediterranean flour moth, Ephesia kuehniella, which typically attack ground cereal products, spend most of their life in silk tubes that probably provide some protection against parasitoids and reduce water loss. Larvae of the Indian meal moth, Plodia interpunctella, seem to be less dependent on protective silk production because they typically develop inside nuts, seeds, and dried fruits. From the comparison of the three species we conclude that silk strength in pyralids is correlated with the regularity of H-fibroin repeats. We also conjecture that the partially recycled silk of pyralids serves as a temporal and readily accessible depot of essential and energy-rich amino acids.

**EXPERIMENTAL PROCEDURES**

**Insect Culture and Silk Gland Dissection—** Cultures of the greater waxmoth, G. mellonella L. (Pyralidae: Galleriinae), and the Mediterranean flour moth, E. kuehniella Zeller (Pyralidae: Phyctinae), were maintained in our Institute as described previously (13, 14). The colony of P. interpunctella (Hubner) (Pyralidae: Phyctinae) was raised on a cereal diet supplemented with ground hazel nuts, sucrose, milk powder, and brewer’s yeast (45.1:1:1:1:1), and the colony was kindly provided by Dr. B. Darvas of the Plant Protection Institute, Budapest, Hungary. To obtain silk glands for molecular analyses, fully grown larvae were anesthetized in water and dissected under saline. About 20 larvae were obtain silk glands for molecular analyses, fully grown larvae were anesthetized in water and dissected under saline. About 20 larvae were used for each silk gland preparation in G. mellonella (body weight ~180 mg) and 50 larvae in E. kuehniella (body weight ~35 mg) and P. interpunctella (body weight ~20 mg). Dissected glands were frozen in liquid nitrogen and stored at ~80 °C.

**PCR Primers—**Conserved regions of known H-fibroin proteins and H-fibroin genes (4) were used to design primers. Sequence similarities between G. mellonella H-fibroin and several tags of E. kuehniella H-fibroin cDNA allowed us to use semi-degenerate primers in an early stage of our research. Commonly used overlapping forward primers E1 and E2 correspond to the MRVTPV and FVILCCA parts of the signal peptide, and the forward primer F1 and reverse primer R1 correspond to the YEED motif in the N-terminal non-repetitive fibroin sequence in G. mellonella H-fibroin (9). The VIVI motif, which is typical for the internal repeats, was used to design reverse and forward primers F2 and R2. The forward primer F3 was based on the GAGNI motif present in some of the E. kuehniella tags, and reverse primer R3 was based on the RRRQF(L)VK sequence conserved in the non-repetitive C-terminal region of G. mellonella and B. mori (4). Final sequence verification was carried out with specific primers.

**Amplification of H-fibroin Sequences from the Genomic DNA—** Larval carcasses left after silk gland dissection were used as a source of genomic DNA. About 1 g of tissue was crushed in a glass mortar under liquid nitrogen and homogenized in 10 ml of lysis buffer (0.1 M NaCl, 0.05 M EDTA, 0.5% SDS, and 0.01 M Tris, pH 7.5). Genomic DNA was extracted in a standard way with phenol/ethanol precipitation, treated with RNase A and proteinase K for 30 min at 37 °C, purified with chloroform, and stored as an ethanol precipitate. PCR was performed with a 25-μl volume. Typically, initial denaturation at 94 °C for 1 min was followed by amplification in 35 cycles of 30 s at 94 °C, 25 s at 51 °C, and 90 s at 72 °C, and termination with a final extension at 72 °C for 10 min.

**Reverse Transcription-PCR (RT-PCR)—** Total RNA was isolated from the whole silk glands of the last instar larvae using the LiCl/urea method. First strand cDNA was prepared from 3 μg of total RNA in a 20-μl PCR mixture containing 200 units of SuperScript II, 1× First-strand buffer, 0.01% dithiothreitol, 500 μM dNTP, 40 units of RNase-OUT ribonuclease inhibitor (all reagents from Invitrogen), and 20 pmol of oligo(dT) adapter primer designed in our laboratory and referred to as Trikant (5). Five microliters of cDNA was added to a reaction mixture that contained 1× high-fidelity PCR buffer, 80 μM dNTP, 0.5 μM forward and reverse primers, and 1 unit of enzyme (Phusion, New England Biolab), and amplification was conducted for 35 cycles: 98 °C for 30 s, 72 °C for 30 s, and 72 °C for 30 s.

**RESULTS**

**Identification of H-fibroin Genes—** To obtain H-fibroin gene sequences from E. kuehniella and P. interpunctella, initial RT-PCR and genomic DNA PCR were performed with forward primers E1 and E2 in all combinations with reverse primers R1 and R2. The dominant RT-PCR products obtained with the E1/R1 pair were about 1000 nt long in E. kuehniella and 750 nt in P. interpunctella. The similarity of their sequences with the G. mellonella H-fibroin gene (9) left no doubt that they represented 5′ ends of H-fibroin cDNAs of the respective species. The PCR with genomic DNA template and primers E1 and R2 resulted in amplification of multiple products ranging in size from about 500 to 2500 nt. Most of them were cloned and sequenced. The results showed that the first 42 nt at the 5′ end of the genomic DNA fragments were identical with the 5′ area of the RT-PCR products in both species. The following 660 nt were used for the further RT-PCR products.
an intron, and the remainder was part of the second exon.

The non-coding upstream region of the H-fibroin gene was identified in E. kuehniella by means of inverse PCR. A ClaI restriction site, which was found in the intron (in a position later identified as 474–479 nt from the transcription start), was taken as the 3' limit of the amplified sequence. A fragment of 800 nt was amplified by nested PCR, and its sequence proved that it contained the 5' end and upstream flanking region of the H-fibroin gene. Sequence homology with H-fibroin genes of other species (4) allowed recognition of a transcription start site 28 nt upstream from the initiation codon. The 3' end of the fragment matched the first exon and part of the intron. As expected, the fragment was terminated with ClaI sites at both ends.

The major aim of our research was to reconcile DNA sequences encoding amino acid repeats that are involved in H-fibroin polymerization. To this end, PCR on genomic templates was performed with the semi-degenerate reverse primer R2 and specific forward primers based on the already known non-repetitive region of the second exon. A "ladder" of PCR products ranging from 600 to 2400 nucleotides was obtained in both E. kuehniella and P. interpunctella, reflecting the presence of repeats containing sequence complementary to the R2 primer. Five products of different sizes were isolated from each species and sequenced from both ends. They proved to be identical at their 5' ends and similar at the 3' ends, but differed by the number of repeats that made up the major central part of each product. In addition, species-specific patterns of repeats could be recognized. The repeats were extremely uniform in both length (342 nt) and structure in E. kuehniella (GenBank™ accession number AY253534), whereas rather random assemblies of several types of short motifs were found in P. interpunctella (accession number AY253533). We regarded the identified sequences as typical for the repetitive region of the respective H-fibroin gene and refrained from searching for longer PCR fragments, especially as the extreme homogeneity of nucleotide repeats and lack of suitable restriction sites would make cloning of long repetitive fragments very difficult.

The arrangement and length of repeats found in E. kuehniella were slightly different from the previously disclosed cDNA tags.² To attest if the differences were due to repeat variation in different gene sequences, we set out to examine cDNAs derived from the 3' end of the gene. The first strand cDNA was synthesized from total silk gland RNA with the aid of the F2 forward primer and the Trikant reverse primer. The PCR product contained 762 nt of which the last 19 nt represented a poly(A) tail. In other PCR runs with the F3 primer, we obtained an overlapping fragment nearly 750 nt long that also terminated with a poly(A) tail.

The Structure of H-fibroin Gene in Pyralid Moths—Alignment of the presently obtained DNA sequences with the H-fibroin gene of G. mellonella (9) facilitated interpretation of the new data. Of 3148 nt identified at the 5' end of the H-fibroin gene in E. kuehniella (GenBank™ accession number AY253534), the first 255 nt represent the non-transcribed upstream region which is followed by a 28-nt leader and 42-nt coding sequence of the first exon, a 660-nt intron, and 2163-nt from the proximal part of the second exon. The sequenced 3' end (GenBank™ accession number AY253535) includes 733 nt of which 471 nt encode amino acid repeats, 177 nt the non-repetitive H-fibroin C terminus, and 85 nt represent the 3' non-translated region terminated with a polyadenine chain. Analysis of the H-fibroin 5' end beginning with the translation start site in P. interpunctella revealed an exon sequence of 42 and 721 nt intron and 1668 nt of a second exon (GenBank™ accession number AY253533).

The sequence alignment disclosed several homologous regions in the fibroin genes of pyralid moths (Fig. 1). Sequence similarities in the 5' gene end and adjacent 5'-flanking region extend to position −55. Distinct homology occurs between positions −12 to −55 and is particularly high in the area of the TATA box (−26 to −31 in E. kuehniella and −27 to −32 in G. mellonella). An alternative TATA box, which is present in G. mellonella at −47 to −51, is also associated with a small area of homology. Homology of the leader sequence between G. mellonella and E. kuehniella surpasses 70%.

The first exon and the initial 6 nt of the intron are identical in all three pyralid species (Fig. 1A). The introns are alike by having high contents of adenine and thymidine, but homology is restricted to the exon/intron and intron/exon boundaries demarcated by the classical GT/AG borders. The intron in G. mellonella H-fibroin is unusually long due to insertion of a repetitive DNA element, Gm1 (9). The beginning of the second exon exhibits a high degree of interspecies homology; the first 326 nt in G. mellonella, 323 nt in E. kuehniella, and 320 nt in P. interpunctella can be aligned with just a few gaps that represent occasional deletions or insertions (Fig. 1). The homology, which is particularly high between E. kuehniella and P. interpunctella, is interrupted by an insertion of 201 nt in G. mellonella and 69 nt in E. kuehniella just before the start of the repetitive sequence. The initial stretch of 90 nt of the repetitive region is clearly similar in all three species, then the G. mellonella sequence diversifies and interspecies similarity is lost. The partly conserved region of 139 nt, which encodes a portion of an amino acid repeat designated A-type (see below), reappears in the repetitive nucleotide sequence in the subsequent parts of the gene (data not shown) in all species. Most of the repetitive sequence, however, is species-specific.

DNA repeats were shown to be highly homogenous throughout H-fibroin gene in G. mellonella (9). Repeat homogeneity was now found in E. kuehniella (GenBank™ accession numbers AY253534 and AY253535) and to a lesser extent in P. interpunctella (GenBank™ accession number AY253533). For G. mellonella and E. kuehniella it has been demonstrated, and for P. interpunctella it is assumed that virtually identical repetitive DNA blocks occur both at the 5' and 3' ends of the gene, suggesting that they occupy the entire central gene region.

At the non-repetitive 3' end, about 150 nt of the coding region and 80 nt of the non-translated tail contain DNA blocks that are the same in G. mellonella and E. kuehniella. Very high homology between the two compared species occurs in the last 84 nt before the stop codon (Fig. 1B). The polyadenylation signal is localized closer to the stop codon in E. kuehniella than in G. mellonella. The poly(A) tail is added to the transcript about 70 nt after the polyadenylation signal in the former, and more than 110 nt in the latter species.

Deduced Amino Acid Sequence of the Analyzed H-fibroins—Analyzed gene sequences allowed us to deduce 1468 amino acid residues from the N terminus of G. mellonella H-fibroin, 735 in E. kuehniella, and 508 in P. interpunctella. The initial stretch of 18 residues, which complies to the rules characterizing the signal peptide (17), is nearly identical in all three species (Fig. 2A). The following sequence is diversified. Residues 19–28 diverge in E. kuehniella, but the following region of 95 residues is 97% identical in this species and P. interpunctella. The region includes seven areas that are also conserved in G. mellonella. These areas contain acidic (Asp and Glu) and basic (Arg and Lys) residues in a neutral environment of amino acids with bulky side chains, such as Val, Leu, Ile, Asn, Pro, and Tyr (Fig. 2A). The high homology between E. kuehniella and P. interpunctella is disrupted by an insertion of 23 residues just before the start of the repetitive sequence in the former species. An
insertion of 67 residues is present before the start of repeats in *G. mellonella*. The insertions have relatively high Pro content in both species but are dissimilar otherwise. Ratios between the charged, neutral polar, and hydrophobic residues in the non-repetitive part of the H-fibroin are similar in the three species (Table I). The physiological significance of the high representation (22.8–27.2%) of charged residues is unknown. We speculate that it may control the process of polymerization by preventing adherence of hydrated H-fibroin molecules during their synthesis in the endoplasmic reticulum, transport in secretory vesicles, and storage in the silk gland lumen.

The major central part of the H-fibroin consists of repeats that differ from the non-repetitive N terminus by a lower proportion of the charged residues and increased content of the neutral and hydrophobic residues that account for about 60–75% of the sequence (Table I). The repetitive region begins with a repeat designated A type, which includes a homologous zone encompassing 42 (E. kuehniella and *P. interpunctella*) or 43 (*G. mellonella*) residues and a species-specific terminal part (see below). The zone contains a central core composed of Pro, Val, Ile, and Glu, such as the PVIVIEE sequence in E. kuehniella, which is flanked by crystalline-like regions made of Ala, Ser, and Gly (Fig. 2A). At the end of the zone, Gly alternates with bulky residues such as Leu, Pro, Asn, and Tyr. The degree of zone homology between the three species is shown in the alignment of sequences at the beginning of the repetitive regions (Fig. 2A). The match ranges from 47% between *G. mellonella* and *P. interpunctella* and 52% between

---

**Fig. 1. Parts of the H-fibroin gene of *G. mellonella* (G.m., GenBank™ accession numbers AF095239 and AF095240), *E. kuehniella* (E.k., GenBank™ accession numbers AY253534 that AY253535), and *P. interpunctella* (P.i., GenBank™ accession number AY253533).** Sequences are aligned to maximize homologies that are indicated with a gray background. Non-coding parts are printed in lowercase and coding parts in uppercase letters. A, H-fibroin 5' ends with the upstream flanking regions; TATA boxes are framed. The sequences are numbered from the commonly used transcription start site in *G. mellonella*. The first and last two nucleotides of the introns are printed in boldface. The beginning of the repetitive region is marked with an arrowhead. The sequence encoding the homologous zone of the first A-type amino acid repeat is underlined. B, non-repetitive sequences encoding the last 53 amino acid residues and the non-coding 3' tails. Stop codons are printed in boldface; polyadenylation signals are boxed, and the poly(A) tail is underlined.
E. kuehniella and G. mellonella, to 69% between E. kuehniella and G. mellonella. Amino acid residues in matching positions are printed on a gray background. A, H-fibroin N termini with signal peptides printed in italics. The beginning of the repetitive regions is marked with an "arrow," and sequences regarded as the homologous part of the A-type repeat are underlined. B, sequences of C termini numbered backward from the last amino acid residue. Trailing ends of the repetitive region (underlined) contain similar amino acid motifs, and the blocks of the last 28 residues are nearly identical in the two species. Cysteine residues are double-underlined; for B. mori it has been shown that Cys in position 22 provides linkage to light chain fibroin, whereas the penultimate and last cysteine residues form an intramolecular bridge (18).

**Table I**

| Type of residues | G. mellonella | E. kuehniella | P. interpunctella |
|------------------|--------------|--------------|------------------|
|                  | N terminus   | Repeats      | N terminus       | Repeats      | N terminus       | Repeats      |
| Acidic           | 14.5         | 2.7          | 13.4             | 1.9          | 16.5             | 2.6          |
| Basic            | 8.7          | 0.7          | 9.4              | 0            | 10.7             | 1.2          |
| Polar            | 31.2         | 21.0         | 32.3             | 36.3         | 30.1             | 36.0         |
| Neutral          | 16.8         | 34.5         | 17.3             | 27.3         | 13.6             | 23.7         |
| Hydrophobic      | 28.3         | 41.0         | 26.0             | 34.6         | 26.2             | 36.2         |
| Isoelectric point| 4.44         | 3.63         | 4.90             | 3.08         | 4.84             | 4.30         |

**Fig. 2.** Comparison of the terminal sequences of H-fibroin proteins in G. mellonella (G.m.), E. kuehniella (E.k.), and P. interpunctella (P.I.). The sequences are aligned to maximize their mutual matching and to reflect homologies in the corresponding nucleotide sequences (Fig. 1). Amino acid residues in matching positions are printed on a gray background. A, H-fibroin N termini with signal peptides printed in italics. The beginning of the repetitive regions is marked with an "arrow," and sequences regarded as the homologous part of the A-type repeat are underlined. B, sequences of C termini numbered backward from the last amino acid residue. Trailing ends of the repetitive region (underlined) contain similar amino acid motifs, and the blocks of the last 28 residues are nearly identical in the two species. Cysteine residues are double-underlined; for B. mori it has been shown that Cys in position 22 provides linkage to light chain fibroin, whereas the penultimate and last cysteine residues form an intramolecular bridge (18).

**Characteristics of Amino Acid Repeats**—Several types of repeats can be recognized, but all of them can be viewed as species-specific extensions of a zone of 42–43 residues that is conserved, albeit with some modifications, in the three species. Recognition of several repeat types accentuates conservation and reiteration of this zone that alternates with other amino acid repeats. Repeats designated type A include the homologous zone and a distal species-specific section. Interspecies sequence homology in this section and in the other repeats is restricted to short motifs that probably represent vestiges of an ancestral H-fibroin molecule. The sequence of B repeats resembles and is probably derived by duplication from the distal portion of the A type repeat of the given species (Fig. 3). Additional types of repeats were found only in P. interpunctella and represent combinations of shorter motifs.

In the H-fibroin of G. mellonella, A repeats of 63 residues alternate regularly with shorter repeats, designated B₁ (43 residues) and B₂ (18 residues). There are actually two subtypes of A differing in a few residues. One is associated with B₁, and the other with B₂. Because each A repeat is always followed by a B type, the repetitions (AB₁) and (AB₂), which are 106 and 81 residues long, can be regarded as two versions of one long repeat. Repeat uniformity rests on the highly conserved length and composition of the encoding DNA blocks. About 12 assem-
bles AB₁AB₁AB₂AB₂(AB₂)AB₂ make up about 95% of the wax-moth H-fibroin molecule (9).

Two types of A repeats can be distinguished in E. kuehniella. In A₁, the species-specific sequence succeeding the homologous zone includes a PYGPN motif, a crystalline region such as GSAASSAAASGAAGAGG, and a stretch of bulky residues alternating with Gly. The sequence is typically a 36-mer, but in a few cases it is truncated (in one case to 17 residues) or extended (to 39 residues). A₂ differs from A₁ by a few residue replacements in the homologous zone and by a longer unique sequence that consists of 49 residues. The first half and the last quarter of this extension contain doublets of bulky residues interspaced with Gly, and the third quarter is “crystalline-like,” being composed of Ala, Ser, and Gly. Each B repeat begins with a nonapeptide GALLSGAAG and continues with a 42-mer matching the second half of the A₁ repeat.

The regularity of repeats varies within the repetitive realm of E. kuehniella H-fibroin. The beginning of repetitions is characterized by regular combinations of highly conserved repeats, A₁ and B, which make a repetitive unit of 72 residues (Fig. 3B). There are only a few regular variations in the sequence of nearly 600 residues that were identified. Pro, Ala, Ala, and Leu in positions 37, 50, 56, and 68 from the start of A₁, and Val and Leu in positions 10 and 13 from the start of A₂, replace Ala and Gly in A₂. Residues in the exactly known positions are numbered from the translation start site. For E. kuehniella (B), internal sequences 1, 2, 4, 5, and 6 (their order is arbitrary) were deduced from GenBank™ entries BG695831, BG695832, AY034944, BG695835, and BG695828, respectively; residues identified after correction of the original nucleotide sequence (see “Experimental Procedures”) are printed in lowercase.

Repeats close to the C terminus (marked with t) span positions 379 and 56 upstream from the H-fibroin end.
Table II

| Species     | Profile surface | Tensile strength | Tensile strength |
|-------------|-----------------|------------------|------------------|
|             | μm²             | F [N·m⁻²] x 10⁸  | F [N·m⁻²]        |
| B. mori     | 155 ± 11        | 1.81 ± 0.12      | 7.4 x 10⁸        |
| G. mellonella| 22.9 ± 5.5      | 1.10 ± 0.11      | 7.5 x 10⁸        |
| E. kuehniella| 18.2 ± 5.7      | 1.82 ± 0.44      |                  |
| P. interpunctella| 6.9 ± 0.3 | 0.29 ± 0.005 |                  |

B, alternate with Gln, Ser, Ser, and Ile, and with Ala and Val, respectively (Fig. 3A). Previously reported sequences of randomly amplified H-fibroin cDNAs and the newly established sequence of the H-fibroin 3’ end (GenBank accession number AY253535) indicate that the regularity of repeats is disturbed in the central and terminal parts of H-fibroin. The A₁ repeat was found in doublets or in conjunction with A₂, and neither the sequence nor length of either of them was fully conserved. One highly conserved B re-repeat occurs close to the end of the repetitive region that has an A₂B₁A₃B₁ arrangement.

The H-fibroin of P. interpunctella is characterized by more repeat types and lower repeat regularity in comparison with the two other species (Fig. 3C). Three types of A repeats were defined based on the amino acid composition of the homologous zone; only one of these types was found in more than one copy. The terminal unique section of all A repeats is short and for the most part composed of bulky residues and Gly. Concatenations of triplets consisting of Gly, a hydrophobic and a polar residue, are also typical for the terminal part of repeat C, the major parts of repeats B and D, and the entire repeat E. “Crystalline” regions of 7–12 small residues (Ala and some Ser and Gly) are present in the N-terminal parts of repeats B and C.

Physical Properties of the Silk Fibers—Silk fibers spun by fully grown larvae at the stage of wandering were used to measure physical silk properties. Microscopic observations confirmed that the fibers were composed of two threads, each secreted by one of the paired glands and glued together by a layer of sericins. Diameters of the two threads were measured in native silks and the cross-section surfaces calculated under the approximation that the threads are cylinders. The sum of other measuring equipment and different origin of the silk fibers. The published data are based on measurements of well dried fibers obtained from cocoons, which had been stored for an unknown length of time, whereas we used freshly produced fibers. In addition, the fiber pulling strain exerted by a cocoon-spinning larva is probably less than the pulling tension imposed by the body weight of a larva slipping on the smooth slanted plane in our experiments. However, because we used identical methods for fiber acquisition and measurements, the data in Table II provide reliable comparative values that can be correlated with the H-fibroin amino acid sequence.

DISCUSSION

Physiological Demands on H-fibroin Structure—Construction of a cocoon in which the larva pupates is a general function of lepidopteran silk. The cocoon must be firm to resist mechanical damage, durable to withstand harsh weather conditions, non-palatable to persist attacks by animals, and immune to microbial degradation. High tensile strength is the major mechanical requirement imposed on the cocoon silk. In the case of Pyralidae and some other groups of Lepidoptera, silk is also used to protect larvae that live in resilient silky tubes. Such silk must be strong to hold larval weight when it suspends on a silk fiber and extensible to allow expansion of the tube when the larva turns around. The results of physical measurements are consistent with these deductions. The silks of B. mori and A. pernyi, which are used exclusively for cocoons, were reported to have a tensile strength of 7.4 x 10⁸ and 5.8 x 10⁸ newton·m⁻², and extensibility of 24 and 35%, respectively (11). By contrast, the silk of G. mellonella, which is spun both for larval tubes and cocoon, is 7.5 x 10⁸ newton·m⁻² strong and 101% extensible (10).

Selection for physical properties has not been the only driving force in the evolution of H-fibroin. Insects are unable to synthesize the amino acids Arg, His, Ile, Leu, Lys, Met, and Thr and possibly also Phe, Tyr, and Trp (20). These essential amino acids must be obtained in the diet or supplied by symbiotic microorganisms. Because both of these resources may be limited, it is not surprising that the content of essential amino acids is low in the H-fibroin of species employing silk solely for spinning cocoons, which are eventually discarded. The deduced content of essential amino acids in the H-fibroin of B. mori is less than 5% and in A. pernyi less than 13%. The latter value would be considerably lower if Tyr were not regarded as essential.

Silk used for the construction of larval tubes in the pyralid moths is partly recycled. A larva typically extends the tube to a wider dimension on one end and devours the opposite narrow end. Under these circumstances, good digestibility of the silk is desirable and selection pressure against the use of essential amino acids is reduced. The content of essential amino acids in the analyzed H-fibroins of pyralids actually amounts to about 17%, and the representation of non-essential but bulky residues (Pro, Val, Glu, Gln, Asp, and Asn), which are a rich energy source, is equally high. Our previous investigations revealed that silk digestibility is controlled by proteinase inhibitors, and possibly other specific proteins, which are absent in the tube silk but are added to the silk synthesized for cocoon spinning (21, 22).

Characteristic Structural Features of Pyralid H-fibroin—The non-repetitive N and C termini resemble the corresponding parts of H-fibroins in other Lepidoptera (4). The composition and arrangement of the repetitive region, however, is unique. The repeats are characterized by a high representation of Ser and bulky residues, whereas the contents of Gly and Ala are low in comparison to B. mori (Table III). The presence of a homologous zone in the repetitive sequence is the most distinct feature of pyralid H-fibroins. It must be emphasized that G. mellonella represents a subfamily that is remote from the Phycitinae to which E. kuehniella and P. interpunctella belong. These taxonomic relations are consistent with the degree of similarity in the non-repetitive N-terminal H-fibroin region (Fig. 2A) and justify generalizations of our data for the whole Pyralidae family.

Homologous zones of 42–43 residues are mutually separated by 61 or 36 residues in G. mellonella, usually 71 residues but...
Fibroin Evolution and Silk Properties

Table III
Dominating amino acids in the sequenced parts of the repetitive H-fibroin region

| Amino acid | G. mellonella | E. kuehniella | P. interpunctella | B. mori |
|------------|--------------|---------------|------------------|--------|
| Gly        | 522          | 203           | 69               | 2524   |
| %          | 31.6         | 10.3          | 17.8             | 47.4   |
| Ala        | 393          | 180           | 92               | 1662   |
| %          | 23.8         | 21.5          | 23.7             | 31.2   |
| Ser        | 298          | 214           | 81               | 644    |
| %          | 18.1         | 25.8          | 20.1             | 12.1   |
| Ile, Leu, Val | 264      | 71            | 48               | 85     |
| %          | 16.0         | 8.5           | 12.4             | 1.6    |
| Pro        | 52           | 43            | 23               | 11     |
| %          | 3.2          | 5.1           | 5.9              | 0.2    |
| Asn        | 9            | 19            | 10               | 11     |
| %          | 0.6          | 10.3          | 10.6             | 0.2    |
| Glu, Asp   | 45           | 19            | 10               | 43     |
| %          | 2.7          | 2.3           | 2.6              | 0.8    |
| Tyr        | 4            | 25            | 10               | 277    |
| %          | 0.2          | 3.0           | 2.6              | 5.2    |

Values are given as total number and per cent ratio. B. mori is included for comparison.

The absence of concatenations of simple motifs and the short length of polyalanine chains distinguish H-fibroin of pyralid moths from the H-fibroins of Saturniidae (7, 8) and Bombycidae (24), the only other lepidopteran families for which data are available. Fibroins of many spider silks also contain polyalanine chains (25–27) or reiteration of simple Gly-rich motifs (28). The situation in Pyralidae is exceptional. Distant similar nine chains (25–27) or reiteration of simple Gly-rich motifs are less frequent in this species but dominate certain repeat sections in E. kuehniella and P. interpunctella. Reiterations of the same triplet are rare; GLN, GLG, and GPY occasionally occur in doublets, and a string of four GLN was found in an extended region in E. kuehniella, mostly 10 but rarely up to 30 residues in G. mellonella, around 53% in E. kuehniella, and around 53% in P. interpunctella. Various crystalline motifs occupy about 39% of the repetitive H-fibroin region in G. mellonella and P. interpunctella, and around 53% in E. kuehniella. They are distributed within sequences rich in bulky residues, being 9, 20, and 39 residues apart in G. mellonella, mostly 10 but rarely up to 30 residues in E. kuehniella, and 9–30 residues in P. interpunctella.

The absence of concatenations of simple motifs and the short length of polyalanine chains distinguish H-fibroin of pyralid moths from the H-fibroins of Saturniidae (7, 8) and Bombycidae (24), the only other lepidopteran families for which data are available. Fibroins of many spider silks also contain polyalanine chains (25–27) or reiteration of simple Gly-rich motifs (28). The situation in Pyralidae is exceptional. Distant similarity to the structure and distribution of their crystalline motifs indicate a possible course of evolution of the repetitive fibroin region.

A striking feature of the non-repetitive N-terminal part of pyralid H-fibroins is the spacing of charged residues, especially the distribution of six glutamates, each of them close to a valine or a similar residue and 10 or 15 positions apart from the next glutamate (Fig. 2). Conservation of five such groupings in the H-fibroins of Antheraea and Bombyx (32, 33) indicates their functional importance. We suggest that they represent vestiges of an unknown molecule ancestral to the H-fibroin and that a sequence similar to the stretch SGEDKLYRTFVIETDAS-GNEVIY, which is present in positions 129–151 in E. kuehniella H-fibroin (Fig. 2), might have been at the root of repeat evolution. We imagine that mutations and selection for polymerization fostered extension of the initial SG dipeptide to the crystalline region where the homologous zones begin (Fig. 3). Changes in single nucleotides would convert codons for Arg, Glu, and Asp to those encoding Gly and Ala. The central part of the sequence stretch, with adjacent charged and hydrophobic residues, is conserved in the homologous zone of the examined pyralids in several variations, such as GEVIYD and PAPVIVIED in G. mellonella, SAPVIVIEEN and GPVVIEEDN in E. kuehniella, and GPVVIEEDN in P. interpunctella. The subsequent crystalline motif, which is rich in Gly, Ala, and Ser, and the tail end of the homologous zone, which contains bulky residues, can also be traced to the non-repetitive sequence under discussion.

The homologous zone must have evolved before the family Pyralidae split into subfamilies. Further evolution apparently involved multiplication of the terminal tripeptides to full-length A repeats. Their partial duplication gave rise to the B repeats. Concerted evolution driven by unequal crossing-over and gene conversion (34) homogenized repeat sequences within the gene. Homogeneity of the relatively long and complex repeats, which lack concatenations of single residues or oligopeptides, distinguishes pyralid H-fibroins.

In contrast to Pyralidae, the evolution of H-fibroin in the superfamilly Bombycoidea, represented by genera Antheraea (family Saturniidae) and Bombyx (Bombycidae), was driven by tandem duplication of short motifs (35), leading to concatenations of alanine or GlyX dipeptides. Most of Antheraea H-fibroin is built of four types of alternating repeats, which are 20–40 residues long and include chains of 13 alanine residues (7, 8). The repeats in B. mori H-fibroin consist of GlyX iterations in which X is Ala in 65% of cases, Ser in 23%, and Tyr in 9% (24). The evolution of H-fibroin in Bombycoidea was convergent to that of some spider silks that also contain runs of polyalanines or poly(GlyX) motifs (23).
Two Mechanisms of H-fibroin Tallying for Polymerization—

Despite profound differences in the structure of H-fibroins, the caterpillars of Pyralidae, Bombyciidae, and Saturniidae produce silk fibers of similar properties. It is generally accepted that H-fibroin polymerization depends on proper alignment of amino acid residues that cross-link peptidic chains with hydrogen bridges. We propose that there are two structural mechanisms facilitating apposition of the interacting amino acid motifs. The “bombycoïd” type (Bombycoidea include Saturniidae and Bombyciidae) is characterized by strings of alanines or oligopeptidic built from 2 to 6 simple residues. The length of these crystalline alanine chains and oligopeptidic assemblies varies along the H-fibroin molecule (38), apparently due to mis-matching recombination and DNA polymerase slippage that is likely to occur in long repetitive DNA sequences (37). This is associated with great allelic variation in H-fibroin that is likely to occur in long repetitive DNA sequences (37).

The expansion of such crystalline units in pyralid evolution varies along the H-fibroin molecule (36), apparently due to the center of the homologous zones suggest that they are also involved in polymerization; however, the nature of their interactions is unknown. We assume that they confer to the silk additional properties than strength. For example, silk elasticity has been traditionally ascribed to the non-crystalline, amorphous protein regions (15), such as those occupying most of pyralid H-fibroins.

Correlations between Repeat Regularity and the Silk Strength—Silk fiber tensile strength is believed to depend largely on the presence of β-pleated crystallites, but nothing is known about the control of their size. The H-fibroin of B. mori with more than 10 repetitions of the GAGAGS motif in each of about 50 reiterated units should contain huge crystals rendering the silk extremely rigid and much stronger that the silk of pyralids. However, presumably small and isolated crystallites in the H-fibroins of G. mellonella and E. kuehniella seem to provide the same tensile strength. A plausible explanation is that the amorphous regions in pyralid H-fibroin contribute to the fiber strength and that there are innate limits to the crystal growth in B. mori H-fibroin.

Examination of the pyralid H-fibroins confirms that fiber strength is not strictly correlated with the abundance of crystalline units. Their proportion in the H-fibroin repeats is about 39% in both G. mellonella and P. interpunctella, but the tensile strength of H-fibroin is much higher in the former species. It appears that strength is correlated with repeat regularity, which is high in G. mellonella and degraded in P. interpunctella. Repeat heterogeneity is associated with irregular spacing and thereby disturbed alignment of the motifs whose interactions underlie H-fibroin polymerization. Imprecise apposition of short motifs reduces the number of possible interactions, and the resulting polymer is flimsy. We propose that fiber strength in pyralid H-fibroin depends on the uniformity of H-fibroin repeats.

Acknowledgments—We thank Drs. B. Darvas and F. Szentkíralyi of the Plant Protection Institute, Budapest, Hungary, for the culture of P. interpunctella and Drs. Marian Goldsmith, University of Rhode Island, and Judith H. Willis, University of Georgia, for critical comments on the manuscript.

REFERENCES

1. Lucas, F., and Rudall, K. M. (1968) in Comprehensive Biochemistry (Florkin, M., and Stotz, E. H., eds) Vol. 26B, pp. 475–558, Elsevier Science Publishers B.V., Amsterdam.
2. Marsh, R. E., Corey, R. B., and Pauling, L. (1955) Biochim. Biophys. Acta 16, 1–34.
3. Marsh, R. E., Corey, R. B., and Pauling, L. (1955) Acta Crystallogr. 8, 710–715.
4. Feizi, T., Zurovec, M., and Sehnal, F. (2002) J. Insect Biochem. Seric. 32, 1–15.
5. Suzuki, Y., and Brown, D. D. (1972) J. Mol. Biol. 63, 409–429.
6. Zhou, C.-Z., Confalonieri, F., Medina, N., Zeravnic, V., Enaush, C., Yang, T., Jacquet, M., Janin, J., Duguet, M., and Persano, R. (2000) Nucleic Acids Res. 28, 4269–4278.
7. Sieradzki, K., and Yukhurov, K. (2000) J. Mol. Biol. 301, 329–338.
8. Hwang, J. S., Lee, J. S., Goo, T. W., Yun, E. Y., Lee, K. S., Kim, Y. S., Jin, B. R., Lee, S. M., Lee, S. W., Kang, S. Y., and Suh, D. S. (2001) Biotechnol. Lett. 23, 3231–3236.
9. Zurovec, M., and Sehnal, F. (2002) J. Biol. Chem. 277, 22639–22647.
10. Hedges, D. R., Hedges, D. R., and Davidoff, M. R. (1979) Insect Biochem. 9, 69–77.
11. Denney, M. W. (1980) Symp. Soc. Exp. Biol. 34, 247–272.
12. Monroe, E., and Solis, A. M. (1999) in Handbuch der Zoologie IV (Kristensen, N. P., ed) Vol. 35, 491 pp. Walther de Gruyter, Berlin.

Table IV

| Residue | Codon | G. mellonella | E. kuehniella | P. interpunctella | B. mori Ref. 30 | A. pernyi Ref. 31 |
|---------|------|--------------|--------------|-----------------|----------------|-----------------|
| Ala     | gca  | 59.7         | 44.3         | 47.0            | 18.1           | 58.2            |
|         | gcc  | 11.9         | 8.9          | 9.0             | 10.4           | 6.7             |
|         | ggc  | 3.8          | 0.5          | 6.0             | 3.0            | 20.0            |
|         | gcu  | 24.4         | 46.3         | 38.0            | 71.3           | 14.9            |
| Gly     | gga  | 50.1         | 69.6         | 86.4            | 39.4           | 26.6            |
|         | ggc  | 21.6         | 11.2         | 7.4             | 4.1            | 27.5            |
|         | ggg  | 2.1          | 0.9          | 3.7             | 1.0            | 4.3             |
|         | ggu  | 26.0         | 18.3         | 2.5             | 55.5           | 41.5            |
| Ser     | age  | 6.0          | 9.0          | 20.7            | 13.4           | 1.7             |
|         | agu  | 3.0          | 12.9         | 12.6            | 1.9            | 11.4            |
|         | uca  | 40.2         | 47.6         | 44.8            | 68.0           | 57.5            |
|         | ucc  | 9.0          | 15.5         | 8.1             | 0.6            | 5.1             |
|         | ucg  | 13.1         | 5.15         | 2.3             | 1.7            | 9.8             |
|         | ucu  | 28.2         | 20.2         | 11.5            | 14.3           | 14.5            |

Codon usage (in %) for the most common amino acids in the repeated region of the H-fibroins.
13. Sehnal, F. (1966) Z. Wiss. Zool. 174, 53–82
14. Marec, F. (1990) Acta Entomol. Bohemoslov. 87, 445–458
15. Wainwright, S., Biggs, W., Currey, J., and Gosline, J. (1982) Mechanical Design in Organisms, 423 pp. Princeton University Press, Princeton, NJ
16. Willis, T. G., Judayel, D. M., Coignet, L. J. A., AbdullaHaf, M., Treleaven, J. G., Catovsky, D., and Dyer, M. J. S. (1997) Blood 90, 2456–2464
17. Nielsen, H., Engelbrecht, J., Brunak, S., and van Heijne, G. (1997) Protein Eng. 10, 1–6
18. Tanaka, K., Kajiyama, N., Ishikura, K., Waga, S., Kukuchi, A., Ohtomo, K., Takagi, T., and Mizuno, S. (1999) Biochim. Biophys. Acta 1432, 92–103
19. Inoue, S., Tanaka, K., Arisaka, F., Kimura, S., Ohtomo, K., and Mizuno, S. (2000) J. Biol. Chem. 275, 40517–40528
20. Nation J. L. (2002) Insect Physiology and Biochemistry, 485 pp. CRC Press, Inc., Boca Raton, FL
21. Nirmala, X., Kodrik, D., Žurovec, M., and Sehnal, F. (2001) Eur. J. Biochem. 268, 1–10
22. Nirmala, X., Mita, K., Vanisree, V., Žurovec, M., and Sehnal, F. (2001) Insect Mol. Biol. 10, 437–445
23. Gatesy, J., Hayashi, C., Motriuk, D., Woods, J., and Lewis, R. (2001) Science 291, 2663–2665
24. Zhou, C.-Z., Confalonieri, F., Jacquet, M., Perasso, R., Li, Z.-G., and Janin, J. (2001) Proteins Struct. Funct. Genet. 448, 119–122
25. Xu, M., and Lewis, R. (1990) Proc. Natl. Acad. Sci. U. S. A. 87, 7129–7124
26. Hinman, M., and Lewis, R. (1992) J. Biol. Chem. 267, 19320–19324
27. Guerette, F., Ginzinger, D., Weyer, B., and Gosline, J. M. (1996) Science 272, 112–115
28. Hayashi, C. Y., and Lewis, R. V. (1997) J. Mol. Biol. 275, 773–784
29. Colgin, M. A., and Lewis, R. (1998) Protein Sci. 7, 667–672
30. Mita, K., Ichimura, S., Zama, M., and James, T. C. (1988) J. Mol. Biol. 203, 917–925
31. Yukhiro, K., Kanda, T., and Tamura, T. (1997) Insect Mol. Biol. 6, 39–55
32. Kusuda, J., Tazima, Y., Onimaru, K., Nisaki, O., and Suzuki, Y. (1984) Mol. Gen. Genet. 203, 359–364
33. Tamura, T., Inoue, H., and Suzuki, Y. (1987) Mol. Gen. Genet. 206, 189–195
34. Smith, G. P. (1976) Science 191, 528–535
35. Ichimura, S., and Mita, K. (1992) J. Mol. Evol. 35, 123–130
36. Mita, K., Ichimura, S., and James, T. C. (1994) J. Mol. Evol. 39, 583–592
37. Armour, J. A. L., Monckton, D. G., Neil, D. L., Tamaki, K., MacLeod, A., Allen, M., Croutier, M., and Jeffreys, A. J. (1993) Genome Anal. 7, 43–47
38. Manning, R. F., and Gage, L. P. (1980) J. Biol. Chem. 255, 9451–9457
39. Beckwitt, R., Arcidiacono, S., and Stute, R. (1998) Insect Biochem. Mol. Biol. 28, 121–130