# Protein Composition of Silk Filaments Spun under Water by Caddisfly Larvae

Naoyuki Yonemura,[†,‡] František Sehnal,*[,†] Kazuei Mita,[‡] and Toshiki Tamura[‡]

*Biology Centre, Academy of Sciences, Institute of Entomology, Branišovská 31, 370 05 České Budějovice, Czech Republic, and National Institute of Agrobiological Sciences, Tsukuba, Ibaraki 305-8634, Japan*

*Received July 10, 2006; Revised Manuscript Received September 6, 2006*

Silk fiber produced by the larvae of Trichoptera (caddisflies) and Lepidoptera (moths and butterflies) is composed of two filaments embedded in a layer of glue proteins. In an aerial environment Lepidoptera spin silk filaments assembled from heavy chain fibroin (H-fibroin), light chain fibroin (L-fibroin), and the glycoprotein P25. The silk filament of caddisflies, which is produced and persists in water, contained homologues of H-fibroin (>500 kDa) and L-fibroin (25 kDa) but not of P25. The amphiphilic nature of H-fibroin and its high content of charged amino acids probably facilitate the secretion and storage of a covalently linked L-fibroin/H-fibroin dimer in the absence of P25. Several types of short amino acid motifs were arranged in orderly fashion in the regularly reiterated repeats that made up more than 95% of the length of H-fibroin. The H-fibroins of *Hydropsyche angustipennis* and *Limnephilus decipiens* from different caddisfly suborders contained GPXGX, SXSXSXSX, and GGX motifs such as the lepidopteran and spider silks but differed from them by a lack of poly(A) and poly(GA) motifs. H-fibroins of both caddisfly species harbored a conserved repeat of 31 residues but were distinguished by a few species-specific motifs and their organization in higher order repeats. Structural differences may be related to the silk function as a catching net in *H. angustipennis* and a stitching fiber in *L. decipiens.*

## Introduction

The silk used in textiles is manufactured from cocoons that are spun by caterpillars (larvae of Lepidoptera) as a shelter for pupae. The silk fiber consists of two filaments which are derived from a pair of labial glands and sealed into a single fiber by coating proteins called sericins.[1] Each filament is basically a polymer of heavy-chain fibroin (H-fibroin, 220−500 kDa) which contains short hydrophilic ends and a long repetitive central region. Reiterations of simple motifs such as GAGAGS in *Bombyx mori*[2] or strings of 10−12 Ala in *Antheraea* species[3] are organized in higher order repeats. Motifs containing alternations or strings of Gly, Ala, and Ser seem to be present in the H-fibroin of all Lepidoptera[4,5] and form $\beta$-sheets and three-dimensional crystallites that reinforce the filament.[6] The formation of noncovalent inter- and intramolecular links is essential for H-fibroin polymerization, and the nature of motifs and their arrangement into higher order repeats determine physical properties of the silk filament.[7]

Studies on the domestic silkworm, *B. mori*, reveal that H-fibroin is associated with light-chain fibroin (L-fibroin) and P25 (also called fibrohexamerin).[8] Disulfide linkage between H-fibroin and L-fibroin (about 25 kDa) is indispensable for the secretion of both components,[9] and weak links between L-fibroin and P25 seem to protect the latter against deglycosylation.[10] P25 occurs in two forms that differ by the degree of glycosylation and apparently act as chaperones, facilitating transport and secretion of the highly insoluble H-fibroin/L-fibroin heterodimers.[8] The assembly of H-fibroin, L-fibroin, and P25 into elementary secretory units[11] is probably essential for the long-term storage of hydrated silk dope in the gland lumen and its rapid polymerization into hydrophobic filaments during spinning. The construction of the silk filament from the three protein components has been conserved for 150 million years of lepidopteran evolution.[5] L-fibroin and P25 have been lost from the silkmoths of the Saturniidae family, which constructs silk filaments from dimers of amphiphilic H-fibroins.[12,13]

Silk components are stored in the gland lumen for days or weeks as a highly concentrated gel that is converted to a solid filament when passing out of the gland. The mechanism of polymerization is not well understood, but shearing and loss of water apparently play important roles. The process of silk polymerization may be different in caddisfly (Trichoptera) larvae, which spin under water. Histological studies distinguish secretions from the posterior and the middle sections of the caddisfly silk glands,[14,15] and by analogy with Lepidoptera the secretions are called "fibroin" and "sericin", respectively. However, no protein homologies between the silk proteins of Trichoptera and Lepidoptera have been demonstrated. Amino acid composition is established only for the whole silk of one trichopteran species[16] without any information on the number of individual silk components. A silk protein contig, which was identified in an EST library based on the silk glands of another caddisfly species, exhibited no homology with known lepidopteran silk proteins.[17] This observation is intriguing because Trichoptera and Lepidoptera evolved from a common ancestor and similarity in their silk composition might be expected. We launched a study to verify whether homologues of H-fibroin, L-fibroin, and P25 occur in caddisfly silk and, if so, what is the structure of repetitive motifs in the H-fibroin. The aim of our work was to find out whether the design of lepidopteran silk is compatible with spinning in an aqueous environment.

We chose to study two species of caddisflies that differ dramatically in their use of silk. The larvae of *Hydropsyche angustipennis* spin hiding tubes and catching nets (Figure 1A). Until pupation the larva lives in a flexible tubular retreat that

* Corresponding author. Tel. +420-38-7775220; Fax +420-38-5310354; E-mail: sehnal@entu.cas.cz.

[†] Biology Centre, Academy of Sciences.

[‡] National Institute of Agrobiological Sciences.

10.1021/bm060663u CCC: $33.50 
Published on Web 11/07/2006

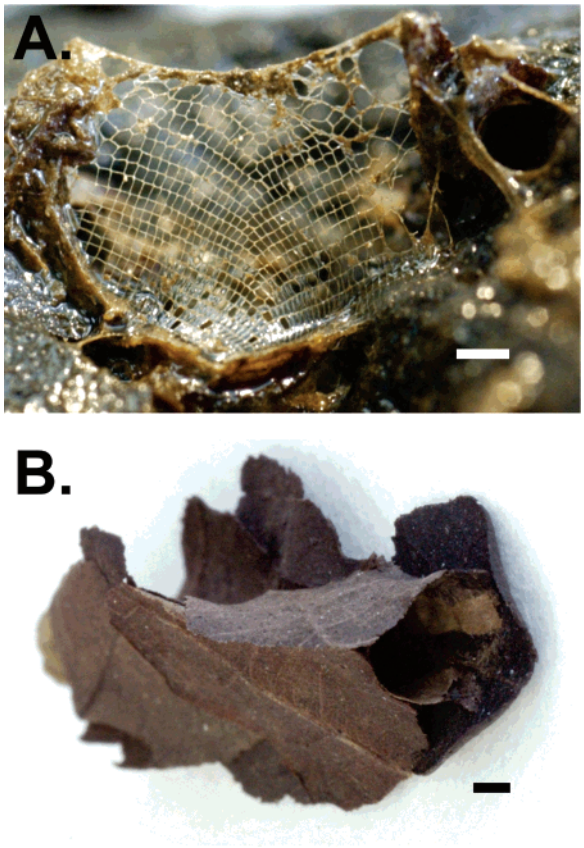


**Figure 1.** (A) A catching net with the opening of the retreat tube spun by the larva of *Hydropsyche angustipennis*. (B) A portable case of a *Limnephilus decipiens* larva stitched by from pieces of grass leaves with a silk fiber. Bars indicate length of 1 mm.

opens toward the water current and is camouflaged and strengthened by bits of decaying plant material attached to the walls. A fine cobweb spun at the tube entrance and opened against the water current is used to collect food brought in by the water. By contrast, the larvae of *Limnephilus decipiens* use silk fiber for stitching fragments of grass into portable cases in which they hide and eventually pupate (Figure 1B).

## Materials and Methods

**Insects and Sample Preparation.** Last instar larvae of caddisflies were collected in the vicinity of České Budějovice (Czech Republic) in April and May. The larvae of *H. angustipennis* (Curtis) (suborder Annulipalpia, family Hydropsychidae) were found under stones in a small brook and those of *L. decipiens* (Kolenati) (Integripalpia, Limnephilidae) on submerged vegetation in a large pond. The larvae of *H. angustipennis* were collected without their hides, whereas those of *L. decipiens* were taken with their cases. Collected larvae were brought to the laboratory in shallow containers cooled with ice. Most larvae were promptly dissected and their silk glands frozen in liquid nitrogen. To obtain material for the silk protein analysis, freshly dissected silk glands of *H. angustipennis* were ruptured in a small volume of chilled distilled water (50 $\mu$L per pair of glands), and liquid silk flowing out of the glands was collected for 5 min. The silk-containing water was mixed with an equal volume of the urea buffer (8 M urea, 2% SDS, and 5% $\beta$-mercaptoethanol in 10 mM Tris, pH 7.0), incubated overnight at 4 °C, and centrifuged, and the supernate was used for protein analysis. To gather spun silk, about 20 larvae were kept for a couple of days at 10 °C in a glass jar containing 200 mL of tap water and silk was collected from the jar wall.

**Silk Protein Analysis.** Silk extracts in 4 M urea were centrifuged, heated for 5 min in a boiling water bath, and loaded onto polyacrylamide gels. Electrophoresis was performed under denaturing conditions. Gels were either stained with Coomassie brilliant blue R-250 or used for semidry protein blotting onto Immobilon-P membranes (Millipore). Selected protein fractions were sent to the Protein and Nucleic Acid Core Facility of the Medical College of Wisconsin for N-terminal peptide sequencing.

**cDNA Libraries**. Total RNA was prepared from 25 pairs of silk glands from *H. angustipennis* and 20 pairs from *L. decipiens*, respectively. The glands were pulverized in a mortar under liquid nitrogen, and the powder was extracted with 750 $\mu$L of ISOGEN-LS reagent (Nippon Gene Co. Ltd, Tokyo). Genomic DNA was removed by partitioning with 200 $\mu$L of chloroform, and RNA was precipitated from the remaining aqueous phase with an equal volume of 2-propanol. The precipitate was rinsed with 70% ethanol, dried briefly at room temperature, dissolved in 0.5% SDS with 20 mM sodium acetate (pH 5.3), and stored at −80 °C. Silk-gland specific cDNA libraries were constructed commercially by Takara Bio Inc. (Ohtsu City, Japan) with a ZAP-cDNA Synthesis Kit (Stratagene). An ABI PRISM 3700 Genetic Analyzer (Applied Biosystems) was used to sequence 1056 randomly chosen clones of *H. angustipennis* and 1056 clones of *L. decipiens.* Inserts with open reading frame (ORF) longer than 300 bp were subjected to BLASTX homology search against protein databases available at the ExPASy Proteomics server (http:// www.expasy.org). Matching sequences (over 95% identity in more than 100 base pairs) were aligned into contigs with the aid of the Clustal W program. The number of separate ESTs belonging to a contig was taken as a measure of the expression level of the respective gene. Most contigs contained the entire ORF and the 3′ untranslated region of the respective cDNA.

**PCR.** RNA was extracted with the ISOGEN-LS reagent from newly prepared silk glands and used with the SMART RACE cDNA Amplification kit (BD Sciences) to prepare the 1st strand cDNA and to perform standard PCR and RACE PCR. Specific primers (shown in Results) were applied to confirm and extend sequences identified in the contigs, and degenerate primers based on lepidopteran P25 were used in attempts to detect P25 homologues. PCR products were purified by electrophoresis in agarose gels, extracted with a QIAquick Gel Extraction Kit (QIAGEN), and either taken for direct sequencing or cloned into a pCR4-TOPO vector (Invitrogen). Plasmid DNA inserts were sequenced with the aid of gene specific primers or with the T3, M13 forward, and M13 reverse primers matching the vector. Prism 3100 or Prism 310 Genetic Analyzers were used with the PRISM Sequencing Analysis and the AutoAssembler software (ABI) for sequencing, sequence alignments, and conceptual translations.

**Northern Analysis.** The expression of identified genes in *H. angustipennis* was verified with Northern analysis. Total RNA was extracted separately from the posterior section or the distal portion of the middle sections of 20 silk glands. In other experiments, total RNA was prepared from the entire silk gland of individual larvae. RNA was fractionated by electrophoresis and blotted onto Hybond N+ membranes (Amersham Biosciences). The L-fibroin probe was obtained by amplifying inserts of appropriate cDNA clones with the T7 and M13 reverse primers; the products were trimmed with *Eco*RI and *Xho*I and isolated with a QIAquick gel extraction kit. H-fibroin probes were amplified from appropriate cDNA clones with T7 and T3 primers and purified by filtration through a Montage PCR centrifugal filter (Millipore) without restriction enzyme treatment. Probes were labeled with the ECL Direct Nucleic Acid Labeling and Detection System (Amersham Biosciences), and hybridization was performed as recommended by the manufacturer.

## Results

**Genes Encoding Proteins of the Silk Filament Core.** Sequences corresponding to genes encoding structural silk proteins were searched for in silk gland specific cDNA libraries. Out of 724 ESTs analyzed in the first screen in *H.* angustipennis, 135 could be assembled into a single contig with a 3′ region

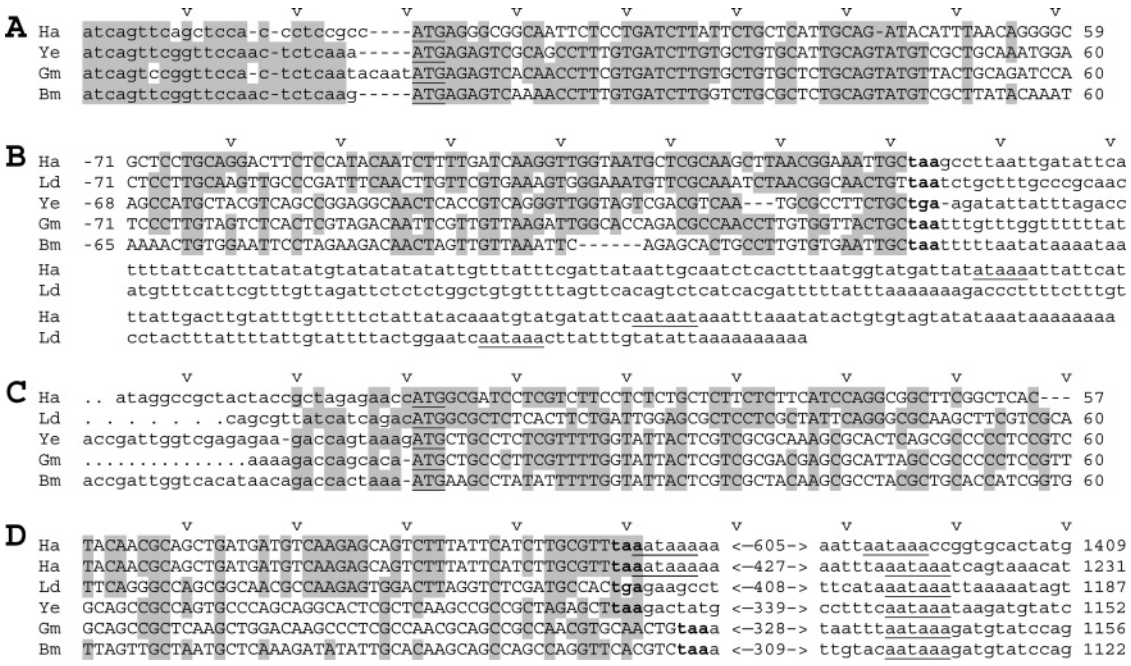


**Figure 2.** Terminal H-fibroin and L-fibroin sequences in the caddisflies *H. angustipennis* (Ha) and *L. decipiens* (Ld) and the moths *Yponomeuta evonymella* (Ye), *Galleria mellonella* (Gm), and *Bombyx mori* (Bm). Nucleotides are numbered from the translation initiation codon (underlined), except for the 3′ H-fibroin end (B) which is numbered backward from the stop codon. Coding nucleotides are in upper case and the noncoding ones in lower case. Stop codons are in boldface, and polyadenylation signals are underlined. Highlighted nucleotides indicate homology between the two caddisfly species and at least one moth species or between at least one caddisfly and two moth species. (A) H-fibroin 5′ region (GenBank entries: Ha, AB214506; Ye, AB195979; Gm, AF095239; Bm, V00094). (B) H-fibroin 3′ region aligned around the stop codon (GenBank entries: Ha, AB214507; Ld, AB214509; Ye, AB195978; Gm, AF095240; Bm, AF226688). Full-length of the untranslated tail is shown only for the caddisflies. (C), (D) Portions of the 5′ and 3′ L-fibroin regions, respectively; two cDNAs with different lengths for the 3′ untranslated region were found in the Ha L-fibroin (GenBank entries: Ha L-fibroin long version, AB214508; Ld, AB214510; Ye, AB195977; Gm, S77817; Bm, ×17291).

homologous to the 3′ end of the lepidopteran H-fibroin gene. A similar contig was represented in *L. decipiens* by 36 among 316 analyzed ESTs. Further analysis confirmed our conjecture that the contigs represented H-fibroin cDNAs. Another abundant EST type, which was present in 36 copies in the databank of *H. angustipennis* and 17 copies in that of *L. decipiens*, encoded proteins of 261 and 254 amino acid residues, respectively, which proved homologous to lepidopteran L-fibroins (see below).

On the basis of the situation in Lepidoptera, cDNAs encoding a homologue of P25 should be represented in the libraries with similar frequencies as the L-fibroin cDNAs. However, none of the 167 and 133 contigs, which were assembled from the complete set of sequenced ESTs in *H. angustipennis* and *L. decipiens*, respectively, encoded a protein resembling P25. Further search for P25 homologues was carried out with 3′ RACE PCR employing the degenerate primer NYP25F01 (5′- ACA ACC TVG CBG CBA ACT CGA AHT GC -3′). This primer corresponded to nucleotides (nt) 155−180 of the conserved region of *Galleria mellonella P25*[18] and was used successfully for *P25* identification in an evolutionarily distant moth species.[5] Silk gland specific first strand cDNA of *H. angustipennis* was used as template for numerous PCR reactions performed under various conditions, but no relevant product was obtained. Additional evidence for the absence of a P25 homologue in caddisflies came from N-terminal sequencing (13−20 residues were identified) of all 20−40 kDa proteins that were separated by SDS PAGE from the silk extract of *H. angustipennis*. Two proteins proved to be L-fibroin (see below), and six were structurally different from all known lepidopteran silk proteins. Analysis of their role in the silk is pending.

**Characterization of H-Fibroin cDNAs.** The length of the 3′ untranslated region of the H-fibroin contig of *H. angustipennis* varied. Three ESTs could be assembled into a 2574 nt sequence ("long version") and 50 others fited a contig of 2492−2503 nt, with the length depending on the number of nt between the poly(A) signal and the start of the poly(A) chain (GenBank Accession No. AB214507 for the long version registered as the 3′ region of H-fibroin cDNA). A reverse primer, HYHFBR02 (5′- TCC GTC TCC GAT AAT GTC GGA GTC CAA -3′), corresponding to nt 184−210 from the 5′ end of the long contig, was used for 5′ RACE, and several PCR products of different length were amplified and cloned. Clones 303 (1662 nt) and 318 (2318 nt) were sequenced (GenBank Accession numbers AB241627 and AB241628, respectively). The sequences were similar and, in a region of 231 nt and some shorter areas, identical, but differences in other portions left no doubt that they represented separate internal regions of the H-fibroin gene. Additional 5′ RACE was performed with reverse primer HYHFBR31 (5′- CTC CGA GGC CGC GTG GTC CGT AGC AA -3′) which matched a DNA stretch that encoded the amino acid motif GCYGPRGLG in clone 303 instead of the regularly occurring GWYGPRGLG. A smeared product was cloned and sequenced. One of the clones consisted of 561 nt and included a 5′ nonrepetitive sequence followed by the repeats: HYHFBR31 hybridized to a region 375 nt downstream from the beginning of the repetitive blocks. Finally, 5′ RACE performed with the reverse primer HYHFBR41 (5′- GCG CCG CAA CAC CCT TAC CG ATC TTC -3′), derived from the end of the nonrepetitive region, yielded a product (HYHFBR41RACE) that apparently began with the transcription start site (position 1 in Figure 2A). The sequence downstream from the nonrepetitive 5′ end was extended with the primer pair HYHFBF41 (forward, 5′- TCC TAA ACA TGG CAA GGG TAA AGG CC -3′, nt 342−367) and HYHFBR31, and the second (1087 nt) and the third (1603 nt) shortest of several amplified products were cloned and sequenced. A combination of these sequences with

the HYHFBR41RACE sequence was defined as a 1943 nt 5′ region of *H. angustipennis* H-fibroin (GenBank Accession No. AB214506).

Analysis of all established sequences yielded the following information about the H-fibroin gene of *H.* angustipennis. The leader sequence and about 50 nt of the ORF, which began in position 25 of the cDNA, were clearly homologous to the lepidopteran H-fibroin gene (Figure 2A). The ORF continued with about 350 nt of a unique sequence followed by repetitive blocks (not shown here). Similar blocks were found in the internal gene sequences (clones 033 and 318) and in most of the 3′ contig. The terminal nonrepetitive part of the 3′ ORF included 216 nt prior to the stop codon. A stretch of about 70 nt immediately preceding the stop codon was similar to the corresponding part of the lepidopteran H-fibroin gene (Figure 2B). The following untranslated terminus contained two putative polyadenylation signals in positions 2470 and 2552 from the start of the 3′ contig. Use of the first signal was associated with the common cDNA version that typically included 2503 nt, but rare species lacking 4, 8, or 10 nt just before the poly(A) tail insertion also occurred. The second polyadenylation signal was used in the longer cDNA.

The H-fibroin contig of *L. decipiens* revealed the 3′ region of the gene. It contained a 1548 nt coding sequence, a presumptive polyadenylation signal in position 1692, and a poly-(A) tail following nucleotide 1711 (GenBank Accession No. AB214509) Repetitive blocks occupied the first 1364 nt, nonrepetitive coding sequence composed the subsequent 189 nt, and the noncoding tail made up the last 163 nt (without the poly(A) chain). The region spanning about 150 nt prior to the stop codon was similar to the H-fibroin of *H. angustipennis*, and the last 70 nt of ORF showed homology to the lepidopteran H-fibroin genes (Figure 2B). The high degree of sequence similarity in the nonrepetitive ORF contrasted with the diversity of the noncoding 3′ tail. No attempt was made to amplify the 5′ end of the H-fibroin cDNA in *L.* decipiens.

**L-Fibroin cDNAs.** Contigs encoding L-fibroin homologues exhibited similarities to the 5′ and 3′ ends of the L-fibroin gene of the moths (Figures 2C, 2D). Two contig versions were identified in *H. angustipennis*, and both were used for 5′ RACE with reverse primer HYLFBR01 (5′- GCG GCT TGA GCG AGA ACT TGA GCA TA -3′). The identical 18 nt of the extreme 5′ region were amplified in both cases. Hence, the full-length cDNAs proved to have identical leader sequences (27 nt) and open reading frames (771 nt), but the 3′ untranslated region of one cDNA version included 204 nt (without the poly-(A) tail) and the other contained 638 nt (GeneBank Accession No. AB214508). The overlapping parts of the two cDNAs were identical except for positions 324 and 858, where cytosines occurred in the longer, and thymines in the shorter version. The replacement in position 324 was in the coding region but did not change the codon sense; the second mutation was in the 3′ noncoding region. The long cDNA version included four distinct polyadenylation signals, of which the first was used in the short (Figure 2D) and the last one in the long cDNA version (not shown). The presence of two L-fibroin mRNAs was confirmed by 3′ RACE with the HYLFBF01 primer (5′- CTC AAG TTC TCG CTC AAG CCG CTG AA -3′).

A single contig encoding the L-fibroin homologue in *L. decipiens* consisted of a portion of the leader sequence, 747 nt of the coding region, 441−448 nt of the untranslated 3′ region, and a poly(A) tail (GeneBank Accession No. AB214510). The extreme 5′ end showed some similarity to the L-fibroin gene of the moths (Figure 2C). The 3′ untranslated region contained

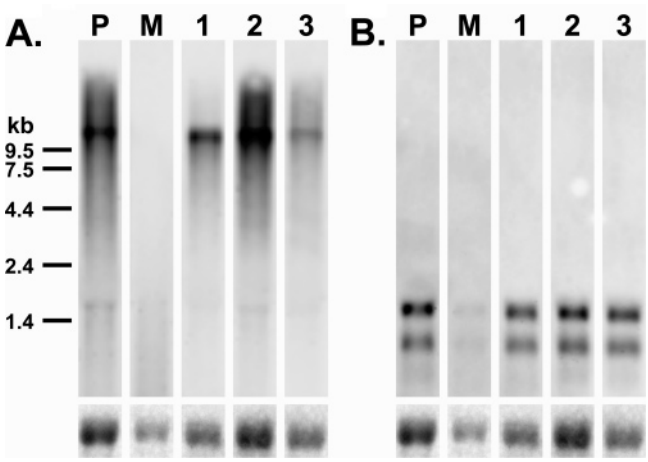


**Figure 3.** Identification of H-fibroin (A) and L-fibroin (B) mRNAs on a Northern blot of total RNA prepared from the middle (lanes M) and posterior (P) sections of 20 silk glands and from the whole glands of individual last instar larvae (lanes 1, 2, and 3) of *H. angustipennis.* Positions of DNA size markers are shown on the left.

only one polyadenylation signal, but the site of poly(A) attachment varied slightly. Both the 5′ and 3′ ends (Figures 2C, 2D) and the internal sequence contained regions of high homology with the L-fibroin of *H. angustipennis*. For example, areas encompassing 547−580 and 434−479 nt in the *L. decipiens* sequence were 94% and 86% identical, respectively, to the 558−591 and 439−484 nt regions of the *H. angustipennis* cDNA. However, the overall homology of the nucleotide sequences was low.

**H-Fibroin and L-Fibroin Expression in Silk Glands.** Expression of the H-fibroin and L-fibroin genes in different silk gland regions was examined in *H. angustipennis.* Northern blots of total RNA were prepared from the posterior and the middle-distal sections of the glands dissected from last instar larvae. The H-fibroin probe was derived from nt 1003-2142 of the 3′ contig (GenBank Accession No. AB214507) and corresponded to the repetitive ORF region. The L-fibroin probe matched the L-fibroin cDNA (GenBank Accession No. AB214508) from position 16 to the end. Both probes detected specific mRNAs in the extracts from the posterior silk gland section (Figure 3, panels A and B, lane P) and did not react with RNA extracted from the middle section (lane M). The H-fibroin probe hybridized with a single transcript of about 15 kb (Figure 3A), revealing that the identified sequence of 8497 nt represented more than half of the mRNA. The L-fibroin probe detected two bands corresponding to about 1.0 and 1.4 kb mRNAs (Figure 3B), consistent with the cDNA sizes described above. Identical hybridizing bands occurred in the analysis of samples prepared from the glands of individual larvae (Figure 3A and B, lanes l-3). The detection of H-fibroin and L-fibroin mRNAs exclusively in the posterior silk gland section was consistent with the data on Lepidoptera. Similar to *H. angustipennis*, some Lepidoptera produce two L-fibroin mRNAs differing in the length of the noncoding 3′ tail.[13,19]

**Deduced H-Fibroin Protein.** The identified 5′ end of the H-fibroin gene of *H. angustipennis* encoded 482 amino acid residues. The alignment of the deduced peptide with the N-terminus of lepidopteran H-fibroins showed clear homology in the predicted signal region[20] which covered 20 residues (Figure 4A). The following nonrepetitive sequence of 115 residues was characterized by frequent grouping of two nonpolar residues (Ile, Leu, Val, Phe, Pro) flanked by one or two charged residues (Asp, Glu, Lys, Arg). These motifs, the overall negative charge, and the similar distribution of charged residues re-

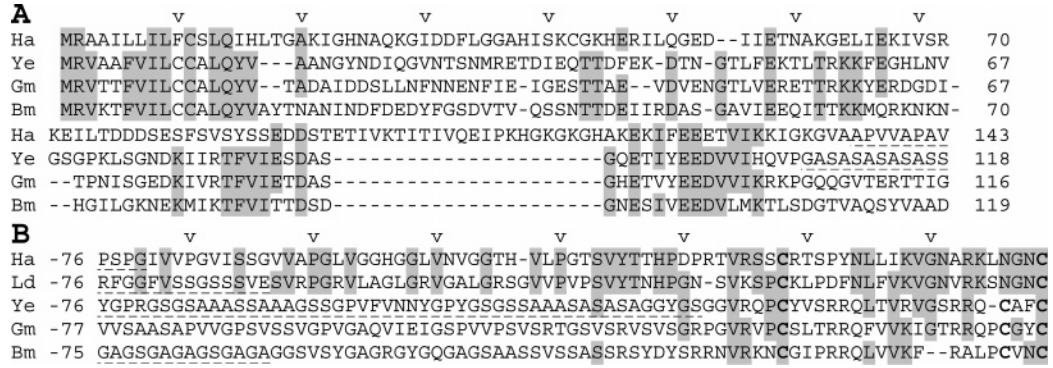


**Figure 4.** Amino acid sequences deduced from the H-fibroin cDNAs of *H. angustipennis* (Ha), *L. decipiens* (Ld), *Y. evonymella* (Ye), *G. mellonella* (Ga), and *B. mori* (Bm) (for full insect names and GenBank Accession Numbers see Figure 1). The sequences are aligned to maximize homologies; residue positions conserved in two caddisfly species and at least one moth species or between at least one caddisfly and two moth species are highlighted. Repetitive sequences adjacent to the nonrepetitive termini are dash-underlined. (A) N-Termini numbered from the initial Met. (B) C-Termini numbered backward from the last Cys residue.

sembled the N-terminus of lepidopteran H-fibroin. Seven residues of the *H. angustipennis* motif AKEKIFEEETVIKK were identical, and the remaining ones were replaced with residues of properties similar to the H-fibroin of Lepidoptera (Figure 4A). Certain similarity to lepidopteran sequences also occurred in the nonrepetitive C-terminus that included 72 residues in *H. angustipennis* and 62 in *L. decipiens* (Figure 4B). The final stretch of 38 residues was homologous to the H-fibroin of Lepidoptera, including a conserved position of $Cys_{-22}$ which in *B. mori* provides covalent linkage to the L-fibroin.[21] $Cys_{-1}$ was also conserved, but its disulfide bridge partner in *B. mori*, $Cys_{-4}$, was lacking in the H-fibroin of caddisflies.

Major portions of the identified H-fibroin cDNAs encoded amino acid repeats composed of several types of short motifs, which are shown in different colors in Figures 5 and 6. In *H. angustipennis*, the repeats began with residue 136 of the deduced H-fibroin amino acid sequence and continued until the end of the identified N-terminus in position 639 (Figure 5). Similar repeats made up the peptide sequences of clones 303 and 318 as well as the sequence of 697 residues that preceded the nonrepetitive C-terminus. Repeat types HA, HB, HC, and HD could be distinguished by the arrangement of short amino acid motifs. The truncated versions of repeats HC and HD are referred to as $HC_1$, $HC_2$, $HC_3$, and $HD_1$, and the slightly extended versions of the HB and HD repeats are labeled $HB_E$ and $HD_E$, respectively. Most kinds of the short motifs occurred in several repeat types. The pentapeptide GPRGL (green in Figure 5) and its modifications (e.g., an Arg replacement by another polar residue) occurred in all repeats except $HC_3$. A similar motif with Pro in the second position but without Gly, e.g., APVVY, APIIR, and QPIYY, was typical for the HD repeats. These repeats further included a stretch of 31 residues (underlined) that was partly conserved in the H-fibroin of *L. decipiens* (see below). The stretch contained a central Pro flanked by symmetrical sets of 15 residues; residues at the same distance from Pro were either identical or similar. Positions 6−14 on either side from Pro were occupied by a Ser-rich motif (blue). A similar motif also occurred in the HA and HC repeats. The HA and HB repeats were characterized by a motif consisting of two hydrophobic residues flanked by 3−4 predominantly hydrophilic residues (gray). Finally, the HA and HC repeats contained a few GGX triplets (shown in yellow, X = Leu, Tyr, or Trp).

The conservation of limited repeat types suggests that they represent motif arrangements essential for H-fibroin function. This may also apply to the repeat assemblies in higher tiers. The assembly was irregular at the start of the repetitive region that included unique repeats $HB_E$, $HD_E$, $HC_1$, and $HC_2$, and at the end of the repetitive region that harbored two copies of $HC_1$. However, the major part of H-fibroin seemed to be composed of the repeat blocks [($HD_1$ $HD_1$) HD HA HB (HB HB) HC $HC_3$] (the repeats in parentheses were not always present).

The analyzed portion of the H-fibroin repetitive region in *L. decipiens* included 454 residues (Figure 6). The representation of individual amino acids differed from the H-fibroin of *H. angustipennis* by the absence of Ala, a lower content of Pro, and a higher representation of basic residues. Short motifs, which resembled those of *H. angustipennis* H-fibroin, composed higher order repeats LD, LE, and LF, which all terminated with a GRRRGPWGRG sequence but were different otherwise. The LD repeat with a stretch of 31 amino acid residues (underlined) containing a central Pro and two Ser-rich lateral motifs was very similar to the major part of the HD repeat, whereas the LE and LF repeats were specific to *L. decipiens*. The LE repeats were characterized by several reiterations of the GGX motif and the LF repeats by four copies of the SV dipeptide. Repeats of the same type did not follow after one another, but the identified sequence was too short to recognize a possible higher order arrangement.

**L-Fibroin Sequence.** The L-fibroin contigs were putatively translated into a protein of 257 amino acid residues (deduced mass 26951 Da) in *H. angustipennis* and 249 residues (25980 Da) in *L. decipiens* (Figure 7). Signal peptide cleavage was predicted to occur between residues 18 and 19, and 21 and 22, respectively.[20] The secreted L-fibroin of *H. angustipennis* began with a sequence HCNTAGLVQATWGLIEDGEI which was found in two proteins purified from the silk by polyacrylamide electrophoresis. Their mobility corresponded to the 26 and 28 kDa standards and their N-terminal sequences read HXNTAGLVQATXGLIEDGEI and HXNTAGLVQATXGLIXD, respectively. We assume that both proteins are encoded by the identified L-fibroin cDNA. The size discrepancy between the 28 kDa protein and the deduced L-fibroin mass (25089 Da after the signal peptide removal) may be due to postranslational modifications. The deduced L-fibroin sequence contained three N-myristoylation and two prenylation sites that may be involved in attaching a nonpeptide moiety to the peptide chain.

Sequence alignment (Figure 7) showed a clear homology between the L-fibroins of the two caddisfly species and their relationship to the lepidopteran L-fibroins, which were characterized by conserved positions of 3 Cys residues. The investigations on *B. mori* revealed that the first two residues form an intramolecular disulfide bridge, while the most distal Cys provides a linkage to the H-fibroin.[22] The positions of these

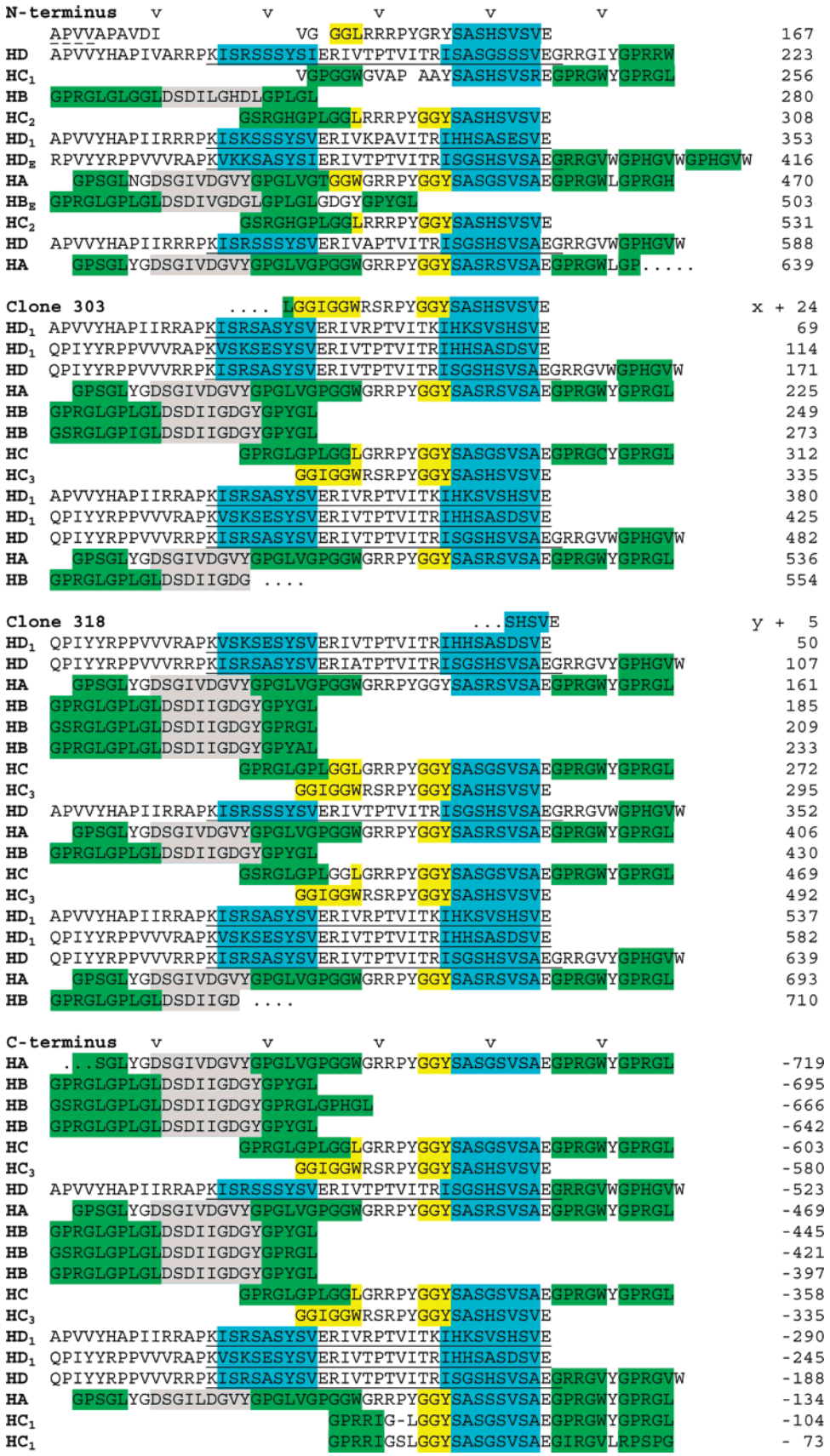


**Figure 5.** Repeats in the amino end (N-terminus), internal regions (clones 303 and 318) and carboxy end (C-terminus) of *H. angustipennis* H-fibroin. The N-terminal sequence is numbered from the initial Met (cf. Figure 4A), internal sequences from their first residues, designated x and y, respectively, and the C-terminus is numbered backward from the terminal Cys (cf. Figure 4B). Distinct motifs are highlighted with different colors and aligned in repeats designated HA, HB, etc. A sequence conserved in the H-fibroin of *L. decipiens* (see Figure 6) is underlined.

cysteines, short stretches of amino acids in their vicinity, and the spacing of certain other residues were conserved in the caddisfly L-fibroins. However, the presence of 2−3 gaps in the protein alignments, C-terminal extensions, and additional Cys


**Figure 6.** Repeats (LD, LE, and LF) in the sequenced carboxy end of *L. decipiens* H-fibroin. The sequence is numbered backward from the terminal Cys (cf. Figure 4B). Distinct motifs are highlighted in colors. The region conserved in *H. angustipennis* is underlined. The relatedness of *L. decipiens* and *H. angustipennis* H-fibroins is demonstrated by a comparison of randomly chosen HD and LD repeats.


**Figure 7.** Deduced L-fibroin amino acid sequences in *H. angustipennis* (Ha), *L. decipiens* (Ld), *Y. evonymella* (Ye), *G. mellonella* (Ga), and *B. mori* (Bm) (for GenBank Accession numbers see Figure 2). Residues in comparable positions in both caddisfly species and at least one moth species or in a caddisfly and at least two moths are highlighted; Cys residues (shown in boldface) conserved in all compared species are underlined.

residues distinguished caddisfly L-fibroins from those of the moths. The similarity of L-fibroins between *H. angustipennis* and *L. decipiens* amounted to 46% (119 out of 257 residues) and extended throughout the sequence. The positions of residues with functional groups in the side chain were preferentially conserved. This included two cysteines that do not have counterparts in the lepidopteran L-fibroins.

## Discussion

**Assembly of the Silk Filament.** Data presented in this paper demonstrate that the silk filament of caddisflies is largely composed of H-fibroin and L-fibroin. The distribution patterns of specific amino acid residues in the nonrepetitive H-fibroin termini and throughout the L-fibroin sequence indicated that these proteins interact in a similar way to the homologous silk components in Lepidoptera. General hydrophilicity and conserved spacing of certain amino acid residues are probably essential for the interaction of H-fibroin ends during the gradual aggregation of H-fibroin molecules in micelles and globules.[23] Divalent ions, which have been shown to connect large micelle assemblies in *B. mori*,[24] possibly interact with acidic residues present at the amino end of both lepidopteran and trichopteran

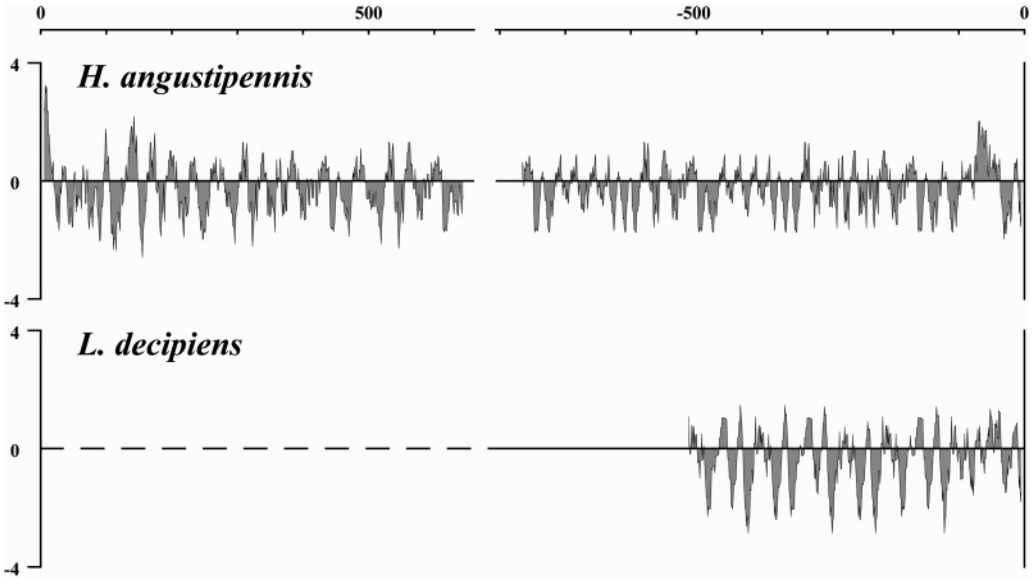


**Figure 8.** Kyte-Doolittle hydropathy plots of the terminal H-fibroin regions. Regions shown consist of repeats except for the first 135 residues in the N-terminus of *H. angustipennis* and the last 72 and 62 residues in the C-terminus of *H. angustipennis* and *L. decipiens*, respectively. Positive values indicate hydrophobicity.

**Table 1.** Per Cent Representation of Selected Amino Acids in the Deduced Repetitive Sequences of the H-Fibroins in Two Caddisfly and Four Moth Species[a]

| residue | *H. angustipennis* | *L. decipiens* | *Y. evonymella* | *G. mellonella* | *B. mori* | *A. pernyi* |
|---|---|---|---|---|---|---|
| Gly | 21.7 | 24.6 | 27.7 | 31.6 | 45.9 | 27.3 |
| Ala | 4.8 | 0.0 | 26.3 | 23.8 | 30.2 | 43.1 |
| Ser | 17.8 | 17.3 | 17.8 | 18.1 | 12.1 | 11.3 |
| Ile, Leu, Val | 20.5 | 21.5 | 8.4 | 16.0 | 2.2 | 1.6 |
| Thr, Tyr | 6.1 | 5.0 | 6.1 | 2.3 | 6.2 | 5.6 |
| Pro | 10.1 | 4.8 | 2.4 | 3.2 | 0.3 | 0.2 |
| Asp, Glu | 5.9 | 4.7 | 2.3 | 2.7 | 1.0 | 5.1 |
| Arg, His, Lys | 14.0 | 16.7 | 0.6 | 0.8 | 0.2 | 3.0 |

[a] Analysis is based on 1042 residues in *H. angustipennis*, 516 in *L. decipiens*, 955 in *Y. evonymella*, 1277 in the waxmoth, *G. mellonella*, 5263 (full sequence) in the silkworm, *B. mori*, and 2639 (full sequence) in the oak silkmoth *A. pernyi*.

H-fibroin. In *B. mori* it has been demonstrated that disulfide linkage between $Cys_{-22}$ of the H-fibroin[20] and $Cys_{170}$ of the L-fibroin[22] is indispensable for the secretion of both proteins.[9] We showed that caddisfly H-fibroin and L-fibroin contain Cys residues in conserved positions, indicating that they also form a heterodimer. Several prolines in the H-fibroin C-terminus of both caddisfly species enforce a particular conformation that may facilitate H-fibroin interactions with other molecules.

We failed to detect a homologue of P25, which is noncovalently attached to the H-fibroin N-terminus in *B.* mori and facilitates the transport and secretion of the insoluble H-fibroin/ L-fibroin dimer.[8,11] An intensive search among 4000 sequences of an EST library of *Rhyacophila obliterata*, a species from the third caddisfly suborder, Spicipalpia, was also not successful (unpublished), suggesting that the absence of P25 is a general feature of Trichoptera. The processing of H-fibroin/ L-fibroin from a hydrated gel to a solid fiber in the absence of P25 may be facilitated by the amphiphilicity of both molecules. Whereas the repetitive region of lepidopteran H-fibroins is largely hydrophobic,[7] hydrophilicity prevailed in the caddisfly H-fibroins (Figure 8). The Grand average of hydropathicity in the repetitive H-fibroin region was −0.330 in *H. angustpennis* and −0.438 in *L. decipiens.* The L-fibroin was also considerably more hydrophilic in Trichoptera than in Lepidoptera and contained regularly alternating hydrophobic and hydrophilic regions. Key residues providing for the hydropathy profile of L-fibroin were conserved in *H. angustipennis* and *L. decipiens*, accentuating their functional importance.

We do not know how the silk filament of caddisflies is stabilized to endure the long exposure to water. Covalent cross-linkage of the H-fibroin/L-fibroin dimer to additional protein-(s) is a possibility. A 37 kDa protein composed of 5 repeats, each 63−65 amino acids long and including 4−5 Cys residues, has been found in the silk gland lumen of an unidentified *Hydropsyche* species.[17] Abundant expression of the corresponding gene (10.3% of all analyzed ESTs in a silk gland specific library) throughout the silk gland but maximally in the posterior region indicates that it is an essential component of the silk filament. We confirmed the presence of a homologous protein in the silk of *H. angustipennis* (unpublished). Elucidation of its possible linkage to some of the cysteines of H-fibroin or L-fibroin requires further research. Other cross-links are also possible. For example, the sequence of 31 residues conserved in the H-fibroin D-type repeats of both *H. angustipennis* and *L. decipiens* included 1−2 lysines, a residue absent from the aerial silks but essential in some structural proteins of the extracellular matrix. In the insoluble network of elastin, Lys is converted to allysine which forms specific cross-links.[25]

**Unique Structural Features of Caddisfly H-fibroin.** The most striking feature of trichopteran H-fibroin is the paucity of Ala and the high content of basic amino acid residues (Table 1). A similar representation of amino acid residues with a high

content of Gly, Ser, and Arg and only traces of Ala was also found in the whole silk of *Pycnopsyche guttifer* from the Limnephilidae family.[16] By contrast, Ala dominates or is second in abundance and charged residues are rare in the H-fibroin repeats of Lepidoptera. Caddisfly H-fibroins also contained high amounts of large hydrophobic residues (Ile, Leu, Val) which are relatively common in the H-fibroins of the pyralid moths[4,26] but are scarce in other Lepidoptera. A high representation of Pro in *H. angustipennis* contrasts with the negligible content of this residue in the H-fibroins of *B. mori* and *A. pernyi.*

The short amino acid motifs poly(A), poly(GA), GPGXX, and GGX apparently evolved independently in the major silk proteins of caterpillars and spiders.[27,28] Typical poly(A) and poly(GA) motifs, which form $\beta$-sheets and three-dimensional crystallites[6] reinforcing the silk filaments, were absent from the H-fibroin of caddisflies. They seem to be functionally replaced by Ser-rich motifs that occurred in most repeat types in both caddisfly species. Experiments with short synthetic peptides containing strings of Ser or alternations of Ser with another residue proved that these motifs confer crystallinity.[28] Caddisfly H-fibroins were also rich in a pentapeptide with Pro in the second position, similar to the GPGXX array that occurs in the strong and elastic silks of spiders. For example, reiterations of GPGXX are present in spidroin 2 (MaSp2) of *Nephila clavipes*[30] and in both components of the dragline silk in *Araneus diadematus*.[31] It is assumed that GPGXX forms a $\beta$-spiral that functions such as a spring, conferring elasticity to the protein polymer.[32] The H-fibroin of *H. angustipennis* contained a few GPGXX sequences, and both species harbored a GPXGX motif where one X is a hydrophobic residue and the other one hydrophilic (usually Arg). A third motif known from spider and lepidopteran silks and occurring in the caddisfly H-fibroins is the GGX triplet. Such alternation of Gly doublets with another residue provides for a stable spiral conformation.[33]

Several H-fibroin motifs seem to be unique to the caddisflies and may be related to the use of silk in an aqueous environment. Both species contained a stretch of 31 residues in which amino acids of similar properties were located at an equal distance from a centrally located proline. On the other hand, the species differed by the arrangement of motifs in higher order repeats and by the presence of a few species-specific motifs. The sequence DSDIIGD and the Pro-containing motifs APVVY, APIIR, and QPIYY characterized the H-fibroin of *H. angustipennis*, whereas the highly charged motif EEGRRR was specific for *L. decipiens.* The nature of motifs and the mode of their assembly into higher order repeats probably specify physical properties of the silk fiber. Tensile strain has been measured in *Hydropsyche siltalai* and found to be low by comparison with the silks of caterpillars and spiders.[34] The silk of *H. siltalai* is nevertheless at least 2 orders of magnitude stronger than the maximum force to which it can be exposed in the catching net. The fiber can extend more than twice in length before it breaks. These properties are to a great extent determined by the H-fibroin sequence which is very likely similar to that shown here for *H. angustipennis.* No data are available on the mechanical properties of the silk fiber of *L. decipiens* which functions as a stitching thread.

**Acknowledgment.** We thank Dr. Karel Novák for advice on the selection of caddisfly species and Drs. Michal Žurovec and Robert Fedič for useful technical comments in course of our work. We acknowledge critical reading of the manuscript by Dr. Marian Goldsmith of the University of Rhode Island. Our work was supported by a cooperative research fellowship awarded to N. Yonemura by the Japan Society for the Promotion of Science and by Grant A5007402 to F. Sehnal from the Grant Agency of the Academy of Sciences of the Czech Republic. EST sequencing was partially supported by the Insect Technology Program of the Ministry of Agriculture, Forestry and Fisheries, Japan.

## Abbreviations

EST, expressed sequence tag; H-fibroin, heavy chain fibroin; L-fibroin, light chain fibroin; nt, nucleotides; ORF, open reading frame.

## References and Notes

(1) Fedič, R.; Žurovec, M.; Sehnal, F. J. *Insect Biotechnol. Sericol.* **2002**, *71*, 1−15.
(2) Mita, K.; Ichimura, S.; James, T. C. *J. Mol. Evol.* **1994**, *38*, 584−592.
(3) Sezutsu, H.; Yukuhiro, K. *J. Mol. Evol.* **2000**, *51*, 329−338.
(4) Fedič, R.; Žurovec, M.; Sehnal, F. *J. Biol. Chem.* **2003**, *278*, 35255−35264.
(5) Yonemura, N.; Sehnal F. **2006** *J. Mol. Evol.* **2005**, *63*, 42−53.
(6) Warwicker, J. O. *J. Mol. Biol.* **1960**, *2*, 350−362.
(7) Sehnal, F.; Žurovec, M. *Biomacromolecules* **2004**, *5*, 666−674.
(8) Tanaka, K.; Inoue, S.; Mizuno, S. *Insect Biochem. Mol. Biol.*. **1999**, *29*, 269−276.
(9) Takei, F.; Kikuchi, Y.; Kikuchi, A.; Mizuno, S.; Shimura, K. *J. Cell Biol.* **1987**, *105*, 175−180.
(10) Inoue, S.; Tanaka, K.; Tanaka, H.; Ohtomo, K.; Kanda, T.; Imamura, M.; Quan, G.-X.; Kojima, K.; Yamashita, T.; Nakajima, T.; Taira, H.; Tamura, T.; Mizuno, S. *Eur. J. Biochem.* **2004**, *271*, 356−366.
(11) Inoue, S.; Tanaka, K.; Arisaka, F.; Kimura, S.; Ohtomo, K.; Mizuno, S. *J. Biol. Chem.* **2000**, *275*, 40517−40528.
(12) Tamura, T.; Inoue, H.; Suzuki, Y. *Mol. Gen. Genet.* **1987**, *206*, 189−195.
(13) Tanaka, K.; Mizuno, S. *Insect Biochem. Mol. Biol.* **2001**, *31*, 665−677.
(14) Zaretschnaya, S. N. *Proc. Acad. Sci. U.S.S.R.* **1965**, *12*, 293−303.
(15) Engster, M. *J. Morphol.* **1976**, *150*, 183-211.
(16) Engster, M. *Cell Tissue Res.* **1976**, *169*, 77−92.
(17) Eum, J. H.; Yoe, S. M.; Seo, Y. R.; Kang, S. W.; Han, S. S. *Insect Biochem. Mol. Biol.* **2005**, *35*, 435−441.
(18) Žurovec, M.; Kodrik, D.; Yang, C.; Sehnal, F.; Scheller, K. *Mol. Gen. Genet.* **1998**, *257*, 264−270.
(19) Žurovec, M.; Vašková, M.; Kodrík, D.; Sehnal, F.; Kumaran, A. K. *Mol. Gen. Genet.* **1995**, *247*, 1−6.
(20) Nielsen, H.; Engelbrecht, J.; Brunak, S.; von Heijne, G. *Protein Eng.* **1997**, *10*, 1−6.
(21) Tanaka, K.; Kajiyama, N.; Ishikura, K.; Waga, S.; Kikuchi, A.; Ohtomo, K.; Takagi, T.; Mizuno, S. *Biochim. Biophys. Acta* **1999**, *1432*, 92−103.
(22) Yamaguchi, K.; Kikuchi, Y.; Takagi, T.; Kikuchi, A.; Oyama, F.; Shimura, K.; Mizuno, S. *J. Mol. Biol.* **1989**, *210*, 127−139.
(23) Jin, H. J.; Kaplan, D. L. *Nature (London)* **2003**, *424*, 1057−1061.
(24) Hossain, K. S.; Ochi, A.; Ooyama, E.; Magoshi, J.; Nemoto, N. *Biomacromolecules* **2003**, *4*, 350−359.
(25) Brown-Augsburger, P.; Tisdale, C.; Broekelmann, T.; Sloan, C.; Mecham, R. P. *J. Biol. Chem.* **1995**, *270*, 17778−17783.
(26) Žurovec, M.; Sehnal, F. *J. Biol. Chem.* **2002**, *277*, 22639−22647.
(27) Gatesy, J.; Hayashi, C.; Motriuk, D.; Woods, J.; Lewis, R. *Science* **2001**, *291*, 2603−2605.
(28) Bini, E.; Knight, D. P.; Kaplan, D. L. *J. Mol. Biol.* **2004**, *335*, 27−40.
(29) Simmons, A.; Ray, E.; Jelinski, L. *Macromolecules* **1994**, *27*, 5235−5237.
(30) Beckwitt, R.; Arcidiacono, S.; Stote, R. *Insect Biochem. Mol. Biol.* **1998**, *28*, 121−130.
(31) Guerette, P. A.; Ginzinger, D. G.; Weber, B. H. F.; Gosline, J. M. *Science* **1996**, *272*, 112−115.
(32) Hayashi, C. Y.; Lewis, R. V. *J. Mol. Biol.* **1998**, *275*, 773−784.
(33) Ashida, J.; Ohgo, K.; Komatsu, K.; Kubota, A.; Asakura, T. *J. Biomol. NMR* **2003**, *25*, 91−103.
(34) Brown, S. A.; Ruxton, G. D.; Humphries, S. *J. N. Am. Benthol. Soc.* **2003**, *23*, 771−779.

BM060663U