J Mol Evol (2009) 68:641–653
DOI 10.1007/s00239-009-9234-5

Conservation of Silk Genes in Trichoptera and Lepidoptera

Naoyuki Yonemura · Kazuei Mita · Toshiki Tamura · František Sehnal

Received: 15 June 2008 / Accepted: 8 April 2009 / Published online: 16 May 2009

Abstract Larvae of the sister orders Trichoptera and Lepidoptera are characterized by silk secretion from a pair of labial glands. In both orders the silk filament consists of heavy (H)- and light (L)-chain fibroins and in Lepidoptera it also includes a P25 glycoprotein. The *L-fibroin* and *H-fibroin* genes of *Rhyacophila obliterata* and *Hydropsyche angustipennis* caddisflies have exon/intron structuring (seven exons in *L-fibroin* and two in *H-fibroin*) similar to that in their counterparts in Lepidoptera. Fibroin cDNAs are also known in *Limnephilus decipiens*, representing the third caddisfly suborder. Amino acid sequences of deduced L-fibroin proteins and of the terminal H-fibroin regions are about 50% identical among the three caddisfly species but their similarity to lepidopteran fibroins is <25%. Positions of some residues are conserved, including cysteines that were shown to link the L-fibroin and H-fibroin by a disulfide bridge in Lepidoptera. The long internal part of H-fibroins is composed of short motifs arranged in species-specific repeats. They are extremely uniform in *R. obliterata.* Motifs $(SX)_n$, GGX, and GPGXX occur in both Trichoptera and Lepidoptera. The trichopteran H-fibroins further contain charged amphiphilic motifs but lack the strings of alanines or alanine-glycine dipeptides that are typical lepidopteran motifs. On the other hand, sequences composed of a motif similar to ERIVAPTVITR surrounded by the $(SX)_{4\text{-}6}$ strings and modifications of the GRRGWGRRG motif occur in Trichoptera and not in Lepidoptera.

Keywords Silk evolution · Trichoptera · Lepidoptera · Fibroin · DNA repeats · Protein polymers · Insect genes

N. Yonemura · K. Mita · T. Tamura
National Institute of Agrobiological Sciences, Tsukuba, Ibaraki 305-8634, Japan

N. Yonemura · F. Sehnal (✉)
Biology Centre, Academy of Sciences, Institute of Entomology, Branišovská 31, 370 05 České Budějovice, Czech Republic
e-mail: sehnal@entu.cas.cz

Introduction

Several groups of terrestrial arthropods produce proteins that polymerize into fibers known as silk (Sehnal and Akai 1990; Craig 1997). Silk secretion from a pair of labial glands is characteristic for the larvae of several clades of Holometabola and reaches perfection in the supraorder Amphiesmenoptera, which includes Trichoptera (caddisflies) and Lepidoptera (moths and butterflies) (Akai et al. 2003). Studies on the major commercial silk producer, the domestic silkworm *Bombyx mori*, revealed that proteins secreted in the posterior section of each gland polymerize into a core silk filament; the pair of filaments is sealed into a single fiber by a layer of sericins derived from the middle section of the glands. The filaments consist of three proteins known as heavy- and light-chain fibroins (H- and L-fibroins, respectively) and the P25 protein or fibrohexamerin (Tanaka et al. 1999a). Disulfide linkage between the high molecular H-fibroin (>350 kDa) and the L-fibroin (~25 kDa) proved indispensable for the secretion of both components (Takei et al. 1987). P25 occurs as two differentially glycosylated moieties of ca. 27 and 31 kDa. Their interaction with the H-fibroin N-terminus is believed to facilitate storage of the highly insoluble H-fibroin/L-fibroin dimer in the form of a gel in the silk gland lumen, as well as gel conversion into the solid filament during spinning (Inoue et al. 2004). Identification of homologous silk components in other Lepidoptera (reviewed by Fedič et al.

2002), including the ancient suprafamily Yponomeutoidea (Yonemura and Sehnal 2006), indicated that this composition of the silk filament occurs in most Lepidoptera. It was probably secondarily lost in the evolutionary advanced family Saturniidae, whose silk filament is made of H-fibroin homodimers without participation of either L-fibroin or P25 (Tamura et al. 1987; Tanaka and Mizuno 2001).

The existence of distinct secretions from the posterior and the middle silk gland sections was demonstrated by histological staining in the caddisfly larvae. By analogy with Lepidoptera, the product from the posterior section was referred to as "fibroins" and that from the middle section as "sericins" (Zaretschnaya 1966; Engster 1976). A systematic search for silk components homologous to the lepidopteran silk proteins was undertaken in the caddisflies *Limnephilus decipiens* and *Hydropsyche angustipennis* (Yonemura et al. 2006). The first species was chosen as a representative of the suborder Integripalpia (the tube case-makers), which use silk rather sparsely for stitching portable larval cases from the pieces of plants, sand, or other foreign materials. Larvae of the second species (suborder Annulipalpia, the retreat-makers) produce large amounts of silk to spin food-collecting nets and retreat tunnels. Both species proved to express mRNAs encoding proteins homologous to the lepidopteran H-fibroin and L-fibroin in the posterior silk gland section. The search for a P25 homologue was unsuccessful at both the nucleic acid and the protein level. It was tentatively concluded that the *H-fibroin* and *L-fibroin* genes originated in ancestral Amphiesmenoptera, whereas *P25* evolved in Lepidoptera after their separation from Trichoptera. However, no data were available on the silk composition in the third caddisfly suborder, the cocoon-makers Spicipalpia, which use silk mainly for cocoon construction, similarly to many Lepidoptera. The present paper describes *H-fibroin* and *L-fibroin* transcripts in *Rhyacophila obliterata* from the Spicipalpia and shows that the *H-fibroin* and *L-fibroin* genes in this species and in *H. angustipennis* are homologous to those of Lepidoptera. The P25 silk component is apparently lacking in Trichoptera.

Materials and Methods

Insect Collection and Sample Preparation

Fully grown larvae of *Rhyacophila obliterata* (suborder Spicipalpia, family Rhyacophilidae) and *Hydropsyche angustipennis* (suborder Annulipalpia, family Hydropsychidae) were collected in the Czech Republic, at about 49°N, 13°E. Those of *R. obliterata* occurred in a pristine river in the Sumava Mountains (about 1000 m a.s.l.) in July to August, and larvae of *H. angustipennis* were taken from a small brook in the vicinity of the town České Budějovice (about 400 m a.s.l.) in April to June. Both species pupated in cases attached to stones as small domes made of sand grains. Inside the case the larva spun a cocoon that was a thin mesh of silk fibers in *H. angustipennis* and a more solid structure composed of cross-linked and mutually fusing fibers in *R. obliterata.*

Caddisfly larvae were brought to the laboratory in shallow containers cooled by ice. Dissected silk glands were mostly frozen in liquid nitrogen and used for nucleic acid analysis. Some freshly dissected glands were ruptured in a small volume of chilled distilled water (50 µl per a gland pair) and the silk gel flown out within 30 min was collected into a pipette. The gel and water admixture was briefly vortexed with an equal volume of 10 mM Tris buffer, pH 7.0, containing 8 M urea, 2% SDS, and 5% 2-mercaptoethanol. The solution was left overnight at 4°C, then centrifuged, and the supernatant was taken for polyacrylamide electrophoresis. The spun-out silk used for stitching the sand grains into protective domes was analyzed in *R. obliterata*. The silk fraction solubilized in 8 M guanidinium isothiocyanate was also subjected to electrophoresis.

Silk Gland-Specific cDNA Library

RNA was prepared from 34 pairs of *R. obliterata* silk glands that were pulverized in a mortar under liquid nitrogen. The powder was extracted with 750 µl Isogen-LS reagent (Nippon Gene Co. Ltd., Tokyo), proteins and genomic DNA were removed by partitioning with 250 µl chloroform, and total RNA was precipitated from the remaining watery 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. The cDNA library was constructed commercially by Takara Bio Inc. (Ohtsu City, Japan) with the ZAP-cDNA Synthesis Kit (Stratagene, La Jolla, CA, USA). The kit employs StrataScript reverse transcriptase for first-strand and *Pfu* DNA polymerase for complementary-strand cDNA synthesis; *Eco*RI and *Xho*I restriction sites are inserted for further cloning.

Randomly chosen clones were sequenced in both directions with the T3 or T7 primers on the ABI PRISM 3730 Genetic Analyzer (Applied Biosystems). Sequences (typically exceeding 700 nt) were edited with ABI PRISM Sequencing Analysis software, version 3.3, and those 95% identical over a stretch of more than 300 nt were assembled in clusters. The validity of cDNA clustering and the mutual relationships of different clusters were checked with respect to both the nucleotide sequence and the nature of

encoded proteins with software used in the *Bombyx* genome project (Mita et al. 2003). Contigs were assembled from the clustered tags with the software ABI PRISM AutoAssembler, version 2.0. Clustering might be incorrect in the case of cDNAs containing long repeats with reiterated motifs that could be aligned in more than one modality. To verify faithfulness of the assembly, each suspected cDNA clone was digested with *Eco*RI and *Xho*I, and the length of the released insert was measured and its 5′ end sequenced. Since the inserts were of different lengths, the overlapping array of their 5′ ends had to match the contig, which was based on the assembly of sequences obtained from automatic analysis of the cDNA clones.

Targeted cDNA Analysis and Sequencing

5′ RACE PCR was employed to identify the 5′ end of the long cDNAs that were incomplete in the library. Total RNA was prepared from silk glands with the Isogen-LS reagent (Nippon Gene Co.) and used for 5′ amplification with the BD SMART RACE cDNA Amplification Kit (BD Biosciences); reverse primers (Table 1) were derived from sequences identified in the cDNA library. PCR products were either sequenced directly or cloned into the T-vector pCR4-TOPO or pCR-XL-TOPO (Invitrogen). Sequencing employed BigDye Terminator version 3.1 and was done on ABI Genetic Analyzer 3100 or 3130 (Applied Biosystems).

Protein Analysis

Deduced amino acid sequences of the contigs were used as queries for the BLAST search in public protein databases and in our files of the silk gland cDNAs identified in the caddisflies *H. angustipennis* and *L. decipiens* (Yonemura et al. 2006). Proteins were regarded as similar if they were more than 30% identical over a region longer than 100 amino acid residues. The presence of expected proteins in the silk of *R. obliterata* was verified by electrophoretic analysis of the silk extracts in 8 M urea (Yonemura and Sehnal 2006). Blots of selected proteins in the Immobilon-P membrane (Millipore, Bedford, MA, USA) were sent to the commercial facility at the Medical College of Wisconsin for N-terminal sequencing.

Analysis of Genomic DNA

Last-instar larvae of *H. angustipennis* and *R. obliterata* were individually pulverized in mortars under liquid nitrogen. The powder was suspended in 2 ml G2 buffer (800 mM guanidine HCl, 30 mM Tris- HCl, pH 8.0, 30 mM EDTA, pH 8.0, 5% Tween-20, 0.5% Triton X-100) supplemented with 400 µg RNase A and 150 mU proteinase K. The suspension was incubated overnight at 50°C and centrifuged (20,000 *g* at 4°C for 10 min). The supernatant was loaded on a buffer-equilibrated genomic-tip column (Qiagen). Eluted DNA was precipitated with 2-propanol (70% of the elute volume), washed with 70% ethanol, dried briefly, and resuspended in 50 µl 10 mM Tris-HCl, pH 8.5. Aliquots containing 20 ng genomic DNA were taken for PCR with primers deduced from the cDNA and later also from the genomic sequences (Table 1). PCR analyses of genomic DNA preparations were done with Advantage 2 polymerase mix (BD Biosciences) and high-fidelity PCR with DNA polymerases KOD FX (Toyobo, Osaka, Japan) or *Pfu*Ultra II (Stratagene).[1] PCR products were separated by agarose gel electrophoresis; short products were visualized with ethidium bromide under UV light, and long products susceptible to UV damage were stained with Crystal Violet. DNA was extracted from agarose with the MiniElute Gel Extraction kit (Qiagen) and 10-µl eluates in 10 mM Tris-HCl, pH 8.5, were either taken for direct sequencing with the PCR primers or cloned into pCR4-TOPO or pCR-XL-TOPO vectors. This TA cloning required 3′ adenine overhang, which was present in PCR products amplified with the Advantage 2 polymerase mix, while DNA products obtained with polymerase KOD FX or *Pfu*Ultra II had to be incubated with 2.5 U ExTaq (Takara), 1 µM dATP, and 1 × ExBuffer for 15 min at 72°C.

Inverse PCR

Samples of 100 ng genomic DNA prepared as described above were digested with 10 U *Ban*II, *Hae*II, and *Xba*I, respectively, or with a mix of *Eco*RI and *Mun*I (all from Takara Bio Inc.) for 3 h. The digest was incubated overnight at 4°C with 100 U T4 ligase (New England Biolabs). Self-ligated circular fragments were precipitated and washed with ethanol, dried briefly, and dissolved in distilled water. PCR products amplified with the Advantage 2 polymerase mix were cloned into the pCR4-TOPO vector.

[1] Amplification and cloning of DNA sequences up to about 5000 bp were most conveniently done with the Advantage 2 polymerase mix. According to the User Manual for TA cloning, this mix introduces 25 errors per 100,000 bp after 25 PCR cycles. We found a slightly higher error rate, about 1 per 3000. Toyobo Co. reported that KOD FX DNA polymerase exhibited 19 errors per 144,535 bp, which is 11 times less than the *Taq* and almost equal to the *Pfu* DNA polymerase (Cline et al. 1996). Our assessment was about 1 error per 7000 bp. A 25×10^5 accuracy was reported for the DNA polymerase *Pfu*Ultra II Fusion HS, which we used for long-range PCRs, with an error rate of about 1 per 8000 bp. According to the technical support section of Stratagene, the 25×10^5 figure indicates the occurrence of 1 mutation in 2,500,000 colonies of the amplified *lac* gene; *Taq* accuracy in a similar test was about 1.25×10^5. The DNA polymerase *Pfu*Ultra II Fusion HS is, in our experience, about 3 times more accurate than the standard *Pfu* polymerase and 20 times better than the *Taq* polymerase.

Table 1 Key primers referred to in the text

Primer	Sequence	GenBank reference	Transcript	Use
HaH-F42	5′GCGGCAATTCTCCTGATCTTATTCTG 3′	AB354591/1465	31	PCR for intron detection
HaH-R41	5′GCGCCGCAACACCCTTACCGATCTTC 3′	- ″ - /1944	510	PCR for intron detection
HaH-F51	5′GAAGATCGGTAAGGGTGTTGCGGCGC 3′	- ″ - /1919	485	Inverse PCR of 5′ region
HaH-R51	5′GAATAAGATCAGGAGAATTGCCGCCC 3′	- ″ - /1488	54	Inverse PCR of 5′ region
HaH-F40	5′CTTCAACCAATGTGTCCTGCCCGTTC 3′	- ″ - / 65	-1370	PCR of 5′ gene end
HaH-R40	5′CGGCTGATACGCTTCCAGAGGCACT 3′	- ″ - /2910	1476	PCR of 5′ gene end
HaH-F46	5′GGTTGGTAATGCTCGCAAGCTTAACGG 3′	AB214507/2277	?	Inverse PCR of 3′ region
HaH-R46	5′AACCTCCGTGGCCACCTACAAGTCCA 3′	- ″ - /2161	?	Inverse PCR of 3′ region
HaL-F21	5′GCCGCTACTACCGCTAGAGAACCATGG 3′	AB354593/ 534	5	Most of the gene amplified
HaL-R21	5′TTATCAAAAGTCGCGCGCATATCCCCG 3′	- ″ - /3162	2633	Most of the gene amplified
HaL-R25	5′TGTCCAATACGGGCTACGAAAGGAAACC 3′	- ″ - /2375	1846	5′ genome walking
HaL-R32	5′CGGACCTTTCATGCACATTTG 3′	- ″ - /1462	933	5′ genome walking
HaL-R31	5′CCGAACTACATCATAGACAGCTACA 3′	- ″ - / 648	119	5′ genome walking,
HaL-F24	5′TCTACGCCCTCGGTGCTACCCTCAC 3′	- ″ - /2002	1473	3′ genome walking
HaL-F25	5′AACGCAGCTGATGATGTCAAGAGCAGTC 3′	- ″ - /2809	2280	3′ genome walking
HaL-F03	5′GTTGACACGGTCCGTCCGGGGATATG 3′	- ″ - /3120	2591	3′ genome walking
HaL-F33	5′GCTGGCATAGAGGAAGTGACAGGTGGA 3′	- ″ - / 22	-508	Whole gene PCR
HaL-R33	5′AAGTACACGTGCGAGATTCAGCCTTCTTG 3′	- ″ - /3907	3378	Whole gene PCR
RoH-F02	5′TYSTGATCTTRTKCTGYKCHATGCAG 3′	AB354689/3592	42	PCR of 5′ cDNA region
RoH-R01	5́CAAGATCACCTGAATCGGAG 3′	- ″ - /7205	3655	PCR of 5′ cDNA region
oH-F21	5′GGGCAATCTTTCTCCTGATCTTAGTCTGCTC 3′	- ″ - /3580	30	PCR for intron detection
RoH-R11	5′TGATGATGCCCATGCTTTGGTTTTTCTTG 3′	- ″ - /6567	3017	5′ cDNA RACE
RoH-F31	5′CGGGACTACAGTGGAACAGTGTACGG 3′	- ″ - /3806	256	Inverse PCR for 5′ end
RoH-R31	5′GAGCAGACTAAGATCAGGAGAAAGATTGCC 3′	- ″ - /3610	60	Inverse PCR for 5′ end
RoL-F11	5′ATACTCCGACCGAGATGGCGCTCC 3′	AB354690/1717	14	PCR of 5′ gene end
RoL-R05	5′CCATCCAAACTCACTCTCACTCGGGT 3′	- ″ - /8507	6804	PCR of 5′ gene end
RoL-F02	5′CGGAATTGTACGCCTTGGGAGCCACT 3′	- ″ - /8292	6589	PCR of 3′ gene end
RoL-R01	5′AGGTTA**t**GCTTGGTGAC**g**CTCGCTGC 3′	- ″ -/11642	9939	PCR of 3′ gene end
RoL-F18	5′ACTTGACGCGCATCCTCTTCGAAAT 3′	- ″ - /7907	6204	Inverse PCR, gene 5′ end
RoL-R15	5′GCAACCGTTTTGGCCACCCTC 3′	- ″ - /1835	132	Inverse PCR, gene 5′ end
RoL-F05	5′CATTTGCCATTTCAGACAGACAGACAAAG 3′	- ″ -/10844	9141	Inverse PCR, gene 3′ end
RoL-R11	5′AACTGGTGCTGCCGTCATTGAGTCCTAC 3′	- ″ - /9424	7721	Inverse PCR, gene 3′ end
RoL-F20	5′CGAAATCGTCATTCATTATCCCCGAGACTT 3′	- ″ - / 735	-969	Whole gene PCR
RoL-R57	5′TTCCACAGTCGATATTTAGAAAGGCGTTGGTT3′	- ″ -/12247	10544	Whole gene PCR

Note: GenBank reference: the accession number of the deposited sequence and the primer position in that sequence. Transcript: the primer position in relation to the transcription start

Genome Walking

Samples of 1 μg genomic DNA of *H. angustipennis* were digested with 30 U each *Kpn*I and *Sac*I (Takara Bio Inc.) for 2 h and further analyzed with the TOPO Walker Kit (Invitrogen). DNA digest was dephosphorylated, purified by phenol:chloroform extraction and ethanol precipitation, and used as a template for extension of the 5′ and 3′ gene regions with the Advantage 2 polymerase mix and several primers (Table 1).

Results

R. obliterata Light-Chain Fibroin cDNA and Protein

The silk gland-specific cDNA library of *R. obliterata* contained 114 related cDNAs (~5% of all clones), which apparently represented several alleles of one gene. A contig assembled from 32 clones was designated allele rhsg004.12 and deposited in GenBank under accession number AB354590. It began with the monamer CATAGGCCG and

contained an open reading frame extending from position 28 to a termination codon at nt 787. The region after nt 802 contained a stretch of 8–11 A's, a TA dinucleotide, and a stretch of 6–10 T's. Two polyadenylation signals occurred, at positions 1051 and 1055, respectively, and a poly(A) tail began at nt 1084.

Putative translation of the rhsg004.12 contig yielded a protein of 253 amino acid residues; positions of 57.6% and 48.4% residues were identical to those in the L-fibroins of the caddisflies *L. decipiens* and *H. angustipennis*, respectively, and comparisons with the moths *Yponomeuta evonymella*, *Galleria mellonella*, and *B. mori* disclosed 24.0%–26.8% similarities. The L-fibroins of Trichoptera were distinguished from those of Lepidoptera by a gap of 16–19 residues shortly after the signal peptide cleavage site and a shorter gap between the second and third Cys residues that were conserved in all L-fibroins (Fig. 1). Caddisfly L-fibroins contained additional Cys residues (three in *L. decipiens*, four in *R. obliterata*, and six in *H. angustipennis*).

N-terminal sequencing of a 26-kDa protein extracted from the silk used to glue sand grains into protective domes (Fig. 2a) yielded the sequence AIQPALIEAT, which matched residues 20–29 of the deduced L-fibroin sequence. Analyzed protein was obviously secreted L-fibroin; the first 19 residues of L-fibroin deduced from the cDNA represented the signal peptide. A 26-kDa fraction was also detected among proteins extracted from the silk gland lumen of larvae several weeks before pupation (Fig. 2b).

The *L-Fibroin* Gene of *R. obliterata*

Two overlapping portions of the gene were amplified from the genomic DNA of a single larva designated Ro03 by PCR with the Advantage 2 polymerase mix. A 7-kb part of the 5′ end was obtained with primers RoL-F11 and RoL-R05, and a 3-kb 3′ end with primers RoL-F02 and RoLR-01, respectively (Table 1). Several clones were sequenced with a number of primers and the resulting continuous sequence of 9925 nt was compared with the *L-fibroin* cDNA. The comparison revealed that the genomic sequence began at position 13 from the transcription start, included five introns, and extended to position 388 upstream from the polyadenylation signal at the 3′ end. Introns occupied most of the sequence (Fig. 3). The sequence of the coding region differed slightly from the rhsg004.12 cDNA contig (GenBank accession no. AB354590). Differences included transitions of C to T at position 288, A to C at 420, T to C at 459, C to T at 522, and C to T at 765. Transversions occurred at three other

```
                 v         v         v         v         v         v         v
Ro  MALLLLTAAFLATQGIASAAIQ------------------PALIEATWRLVEDGEIPPFALLLRDELIA-E  53
Ld  MALSLLIGALLAIQGASFVASSH----------------ISASLLEGTWDLVEQGEVEPYVLLLKDEVV---  56
Ha  MAILVFLSALLFIQAASAHCNT------------------AGLVQATWGLIEDGEIEPFSLVLRDSILAIE  53
Ye  MLPLVLVLLVAQSALSAPSVSVNQVAYNQAEGPRDNGNLINSYVTDAVFGLLDGAEQNIYMLTNQQIVNDMA  72
Gm  MLPFVLVLLVATSALAAPSVVISQDNINNIAPRVGNGRPISSALIDRAFEIVDGGDTNIYILTIQQILNDLA  72
Bm  MKPIFLVLLVATSAYAAPSVTINQYSDNEI-PRDIDDGKASSVI-SRAWDYVDDTDKSIAILNVQEILKDMA  70

Ro  AGPSST-ELYALGATFTAVGELAWPRAASGCGHSKLINACVGFNDGSTSYSELSDAIDSYAVVLSQAVDNLR 123
Ld  ----STGGVYGLGATLTGVGELAWPRPASGCGHSKLINANVALNDGTLAWGELEDAVDSYAVVLAQAVDNLR 121
Ha  NDN-PTSQLYALGATLTAVSELSWVRPSSACAYANLINANVGLANHNLGRAALSSAIDGYAQVLAQAAENIR 125
Ye  NSGDPTTQALALGQAINLVGEAV-GSTGDACAYANLANAY-----ASGNAAAVSQALSGYVNRLNANINAVA 141
Gm  DQPDGLSQSLAVTQAVAALGELATGVPGNSCEAAAVIDAYANSVRTGDN-SALSIAVANYINRLSSNIGLIS 144
Bm  SQGDYASQASAVAQTAGIIAHLSAGIPGDACAAANVINSYTDGVRSGNFAGFRQSLGPFFGHVG-QNLNLIN 142

Ro  ILGYCCIVPAPWPPMDNSCNDYGRIYSFEDSWDLAKG-------AGNK-ARCIARRLYTSFGARLNNIGAAA 195
Ld  ILGLNCIIPAPWPTLENSCGDWGRIYDFESSWSLSKV---------NKGVVCAARRLYTSFGARANNVGAAA 193
Ha  ILGQCCVLPSPWPVLDNCCGDYGRIYDFENSWSLATG-------CNSEGPRCAARDLYLALNARSNNVGAAA 196
Ye  RLAVDPTAAGSIVGSSGGCAGGGRSYQFEQVWDSVLANANAYTIGLLNEQYCMARRLYASYNPQNNNVAAAL 210
Gm  QLASNPDSLRYSSGPAGNCAGGGRSYQFEAAWDAVLNNANPYQIGLINEEYCAARRLYNAFNSRSNNVGAAI 215
Bm  QLVINPGQLRYSVGPALGCAGGGRIYDFEAAWDAIL---ASSDSSFLNEEYCIVKRLYNSRNSQSNNIAAYI 213

Ro  TSAATIAAREILEQIENDLVTYLNTVV--KSASGSWQCAQKKKNMLTLGGYLKSAIWKAASVTKHNLS     253
Ld  TSAATDAATSIISEIEDELVSYLEAVVS-KSAG----PKQKL--LRTLAGSLKASIFRASGNAKSGLRSRCH 249
Ha  TSAATTPALSIFKRIKGEISSLLSLATAPKSSG----CATRKKDLRTAAGVLKQAIYNAADDVKSSLYSSCV 257
Ye  SASAIPEVRQILSSVAAPLANLMRVVASGGNPAQAAASAQQALAQAAARA                        260
Gm  TAGAVVAQTQAAQIILPSLVNVLSAVAAGGNVAGAAA--QAG--QALANAAANVQL                  267
Bm  TAHLLPPVAQVFHQSAGSITDLLRGVGNGNDATGLVANAQRYIAQAASQVHV                       262
```

Fig. 1 Deduced L-fibroin sequences of the caddisflies *Rhyacophila obliterata* (Ro), *Limnephilus decipiens* (Ld), and *Hydropsyche angustipennis* (Ha) and the moths *Yponomeuta evonymella* (Ye), *Galleria mellonella* (Gm), and *Bombyx mori* (Bm). Amino acid residues are numbered from the translation start; those conserved at least in two caddisfly species are highlighted in gray. Cys residues are in boldface, and those with known function in Lepidoptera are underlined. The first residue of secreted L-fibroin (after the signal peptide cleavage) is underlined with double straight lines when confirmed by protein sequencing and with a single wavy line when based on prediction. GenBank entries: Ro, AB354590; Ld, AB214510; Ha, AB214508; Ye, AB195977; Gm, S77817; Bm, X17291

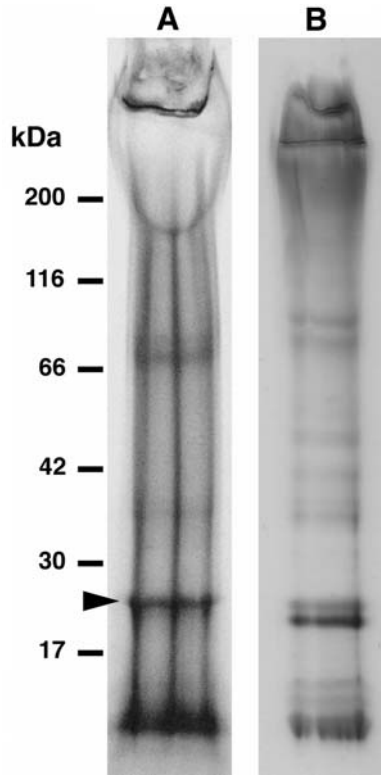

Fig. 2 Silk proteins of *Rhyacophila obliterata* separated by polyacrylamide electrophoresis from silk fibers used to construct shelters for pupation (**a**) and from dope collected from the silk gland lumen (**b**). The arrow in A indicates the fraction proven to be L-fibroin. Note that proteins of similar size occur also in the silk dope

positions: replacement of T with C at position 307 changed Phe_{103} to Leu, replacement of A with C at position 454 changed Asn_{152} to His, and replacement of A with G at position 776 changed His_{259} to Arg. The larval Ro03 obviously contained an *L-fibroin* allele different from rhsg004.12.

Inverse PCR was used to verify and extend terminal gene sequences. A 2-kb product was amplified with primers RoL-R15 and RoL-F18 (Table 1) from the self-ligated *Ban*II digest of the Ro03 larval DNA. The product contained a sequence extending upstream to –1703 from the transcription start. A 4.5-kb stretch of the 3′ gene region was obtained from a DNA sample self-ligated after digestion with *Xba*I and subjected to inverse PCR with primers RoL-R11 and RoL-F05. The sequence extended beyond the poly(A) tail attachment site and allowed design of primers for amplification of the whole *L-fibroin* gene. A product of 11,513 nt was amplified from genomic DNA using *Pfu*Ultra II polymerase with primers RoL-F20 and RoL-R57, cloned into the pCR-XL-TOPO vector, and sequenced. In combination with the overlapping products of inverse PCR the known sequence began at –1703 and included 13,142 nt. It was deposited in GenBank (accession no. AB354690) as *Rhyacophila obliterata L-fibroin* allele 03a (abbreviated *Ro03LFa).*

The *L-Fibroin* Gene of *H. angustipennis*

A considerable part of the gene was amplified from genomic DNA of a larva designated Ha23 by PCR with KOD FX polymerase. Primers HaL-F21 and HaL-R21 (Table 1), which were deduced from the cDNA analyzed earlier (Yonemura et al. 2006), amplified a 2.6-kb product that was cloned into the pCR4-TOPO vector and sequenced. As expected from the primer positions, the sequence began at +5 from the transcription start and ended 110 nt upstream from the first polyadenylation

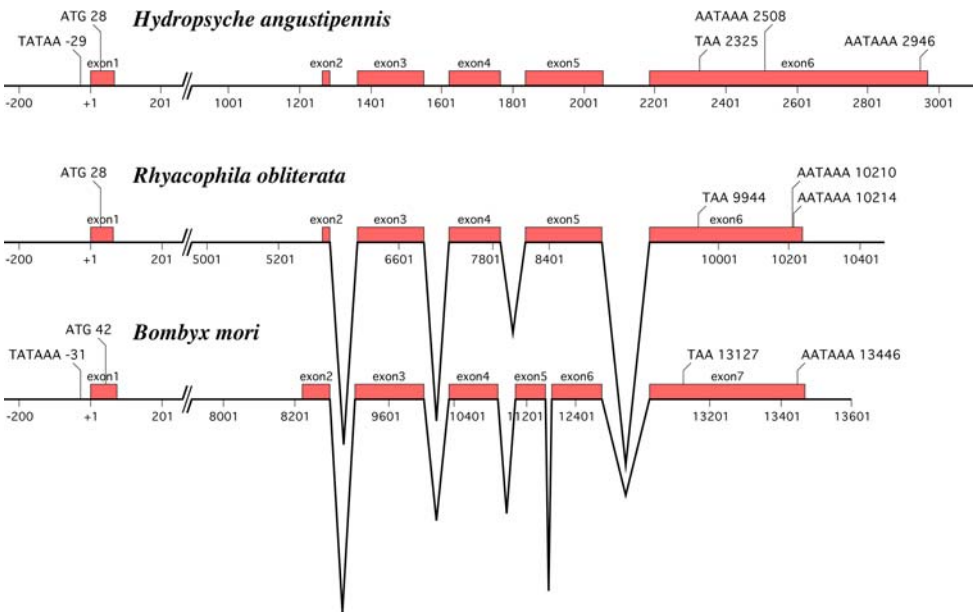

Fig. 3 *L-Fibroin* gene structure in *H. angustipennis* (GenBank accession no. AB354593), *R. obliterata* (AB354690), and *B. mori* (M76430); exons are shown as boxes, and introns as lines. Numbering of sequences begins at the transcription start. Sequences deposited in GenBank include longer upstream regions than shown here. Positions of TATA boxes, translation initiation codons, termination codons, and polyadenylation signals are indicated

signal. Comparison with the cDNA (Yonemura et al. 2006) disclosed the positions of five introns (Fig. 3).

The 5′ and 3′ parts of the gene were identified with the aid of the TOPO walker kit. Products obtained with primers HaL-R25 and HaL-F24, respectively (Table 1), were ligated and taken for PCR with the LinkAmp primer 1 paired with primer HaL-R32 for 5′ and with HaL-F25 for 3′ analyses. Subsequent nested PCR with primers HaL-R32 and HaL-R31 amplified 700 bp of the upstream region. Three products, of 300, 900, and 1100 bp, of the 3′ end were amplified with primers HaL-F25 and HaL-F03. Sequences of the 700- and 1100-bp products permitted design of a primer pair HaL-F33 and HaL-R33 (Table 1) that amplified a product of 3886 nt extending from position –529 to position +3378 of the gene. A contig of 4229 nt, which was based on independent sequencing of the 5′ region (0.7 kb), a major part of the gene (3.8 kb), and the 3′ region (1.1 kb), was deposited in GenBank as allele 23 of the *H. angustipennis L-fibroin* gene (*Ha23LF*) (accession no. AB354593).

Heavy-Chain Fibroin cDNA in *R. obliterata*

Thirty clones identified in the *R. obliterata* cDNA library were of different lengths but the identity of their 3′ end indicated that they were derived from a single gene. The sequence of 2974 nt based on 17 clones encoded 908 amino acid residues of a truncated protein that was homologous to the previously described H-fibroins of caddisflies (Yonemura et al. 2006). The cDNA sequence apparently corresponded to the 3′ end of the *R. obliterata H-fibroin* gene. It was deposited in GenBank as allele *RoHF3′a* (accession no. AB354588).[2] Most of the sequence consisted of highly conserved blocks of 408 nt. The first one was incomplete and included only 373 nt; the complete blocks began at positions 374, 782, 1190, 1598, and 2006, respectively. They differed by only four transitional point mutations. The regularity of repeats dissipated at nt 2413, which was 315 nt upstream from the stop codon. The polyadenylation signal was located 227 nt after the stop codon and a poly(A) tail began 25 nt farther downstream.

All *H-fibroin* cDNAs detected in our library were truncated at the 5′end. PCR search for the 5′ region was launched with the degenerate primer RoH-F02 (degeneracy index = 384), which corresponded to nt 42–67 (counted from the transcription start) of *H. angustipennis H-fibroin* cDNA (Yonemura et al. 2006), and the specific reverse primer RoH-R01 (Table 1), which matched repetitive sequence close to the 5′ end of the *RoHF3′a* fragment. A 1-kb PCR product contained a continuous open reading frame. Reverse primer RoH-R11, derived from the 5′ end of this product, was used for 5′RACE PCR. The established sequence apparently began with the transcription start that was identified in one clone as the first, and in two clones as the second, nucleotide of the consensus CATCAGTCA. Contig *RoHF5′*, which combined results of the initial PCR with those of 5′ RACE, included 1010 nt and was deposited in GenBank under accession no. AB354587. At its 3′ end, the contig contained one full and one partial repeat of 408 nt.

Deduced H-Fibroin Protein in *R. obliterata*

Putative translation product of the *RoHF5′* fragment contained a predicted signal peptide of 20 amino acid residues followed by 460 residues of the secreted H-fibroin. In the optimized alignment of the nonrepetitive N-termini, the sequences of *R. obliterata* and *H. angustipennis* H-fibroins contained 56% residues in comparable positions (Fig. 4a). The similarity to lepidopteran H-fibroins was rather low; only 20% of amino acid positions were conserved in the nonrepetitive regions of the *R. obliterata* and *G. mellonella* H-fibroins. However, the spacing of certain residues was retained and all H-fibroins contained in their N-terminus a motif rich in Lys and Glu in combination with the nonpolar residues Val and Ile (residues 121–129 in *R. obliterata* H-fibroin). H-fibroin similarities between compared species ended shortly after this motif (Fig. 4a).

Beginning by residue 159, the N-terminal sequence deduced from *RoHF5′* became very similar to the repeats encoded by most of the *RoHF3′a* fragment (Fig. 4b). The N-terminal sequence ended and the C-terminal one began with highly conserved repetitive domains of 136 amino acid residues (corresponding to the blocks of 408 nt). The unidentified central part of the H-fibroin most likely consisted of such domains. In Fig. 4, each domain was divided into repeats RD, RF, and RE to accentuate similarities to the LD, LE, and LF repeats previously identified in the

[2] Several deviations from the *RoHF3'a* sequence were found in 13 cDNA clones. All of them contained G in place of C in position 2654 (counted from the first nucleotide of *RoHF3'a*); this caused replacement of Pro_{-24} (counted from the last residue; Fig. 4b) by Ala. A few point mutations occurred at a low frequency also in the 3' UTR. Point mutations which occurred in the coding region of single cDNAs changed Gly_{-763} to Ser and Glu_{-762} to Gly, Gly_{-219} to Ser, Glu_{-218} to Gly, Gly_{-182} to Glu, and His_{-178} to Pro. All these sequence variations had a negligible influence on the general characteristics of the *H-fibroin* 3' end. However, one cDNA of 2152 nt matched the corresponding part of the *RoHF3'a* sequence but contained 18 deletions and 81 nucleotide replacements in the coding region, and 2 insertions, 7 deletions, and 18 replacements in the 3' URL. This cDNA apparently represented a distinct but rare allelic variation and was registered in GenBank as allele *RoHF3'b* of the *H-fibroin* gene (accession no. AB354589). The translation product (501 residues) of allele *RoHF3'b* differed from the homologous region of the *RoHF3'a* product by 22 transversions and two deletions (data not shown).

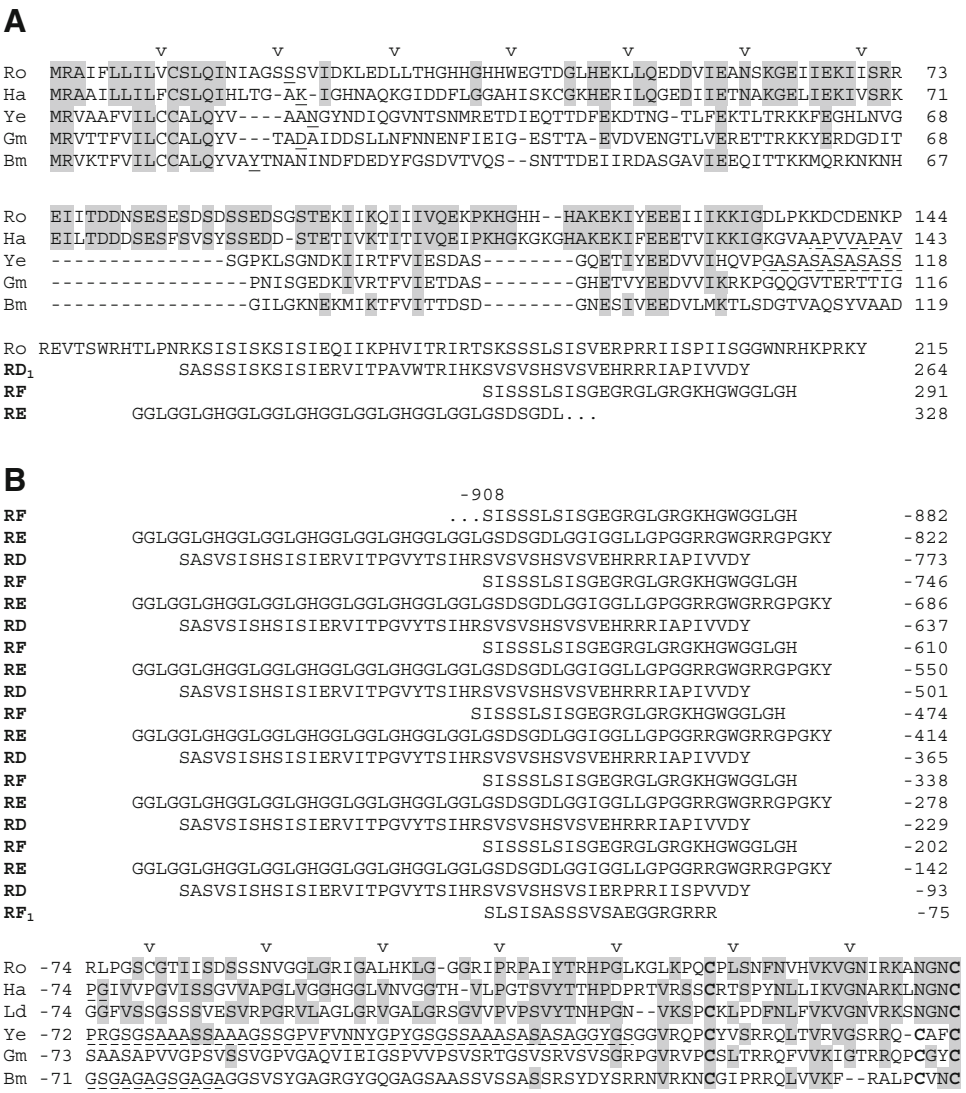

Fig. 4 Amino acid sequences deduced from the analyzed regions of *H-fibroin* cDNA in *R. obliterata* (Ro). Nonrepetitive parts are aligned with the H-fibroins of the caddisflies *H. angustipennis* (Ha) and *L. decipiens* (Ld) and of the moths *Y. evonymella* (Ye), *G. mellonella* (Gm), and *B. mori* (Bm). GenBank entries: Ro 5′ part, AB354587; Ro 3′ part, AB354588; Ld 3′ part, AB214509; Ha 5′ part AB214506; Ha 3′ part, AB214507; Ye 5′ part AB95979; Ye 3′ part, AB195978; Gm 5′ part AF095239; Gm 3′ part, AF095240; Bm, AF226688. Residues conserved in at least two caddisfly species are highlighted gray. The major part of *R. obliterata* H-fibroin is composed of reiterated domains made up of the RD, RF and RE repeats. **a** N-terminal region, with residues numbered from the translation start; the first residue after the predicted signal peptide cleavage is underlined. **b** C-terminal region, with residues numbered backward from the terminal Cys; Cys residues with known function in Lepidoptera are in boldface. Dashed-underlined stretches represent terminal parts of the central repetitive regions (the species differ greatly in the length of the nonrepetitive C-termini)

H-fibroin of *L. decipiens* and the HD repeat found in *H. angustipennis* (Yonemura et al. 2006). The repeats were composed of a few distinct motifs. A stretch of residues of different properties, which was flanked on both sides with a string of five or Ser-X doublets and terminated with Glu, made up a core of the RD repeats. Much longer RE repeats contained reiterated GGL triplets interspaced with a few GH doublets, one GSDSGDL motif, and, at the end, a palindromic LGPGGRRGWGRRGPKY sequence with a central Trp. The RF repeats contained a Ser-rich motif (similar to the SX strings in the RD repeat) flanked on both sides by dodecamers composed of hydrophobic and charged residues. The terminal motif GKHGW of this repeat was similar to the GRRGW of the RE repeats.

The repetitive region of *R. obliterata* H-fibroin ended with a truncated and degenerate repeat RF_1 at –104 to –77 (Fig. 4b). The nonrepetitive stretch comprising 78 C-terminal residues resembled H-fibroins of other caddisfly species and, to a small extent, also those of Lepidoptera. Only two of three Cys residues present at the 3′ end of the typical lepidopteran H-fibroins were found in the H-fibroins of caddisflies.

The *H-Fibroin* Gene of *R. obliterata*

The *H-fibroin* gene of Lepidoptera contains a large intron not far from the start of the coding sequence. The existence of a similar arrangement in the *H-fibroin* of caddisflies was probed by PCR based on primers from the nonrepetitive 5′ cDNA region. A 3-kb fragment was amplified from the genomic DNA of *R. obliterata* larva designated Ro01 with primers RoH-F21 and RoH-R11 (Table 1). The sequence of this fragment revealed overlap with the cDNA sequence and disclosed the presence of an intron 2645 bp long (Fig. 5a). To determine the upstream gene region, the Ro01 DNA was digested with a mix of *Eco*RI and *Mun*I, self-ligated, and used for inverse PCR with primers RoH-R31 and RoH-F31 (Table 1). A 3.5-kb PCR product extended from a *Mun*I restriction site in the intron (+425 from the transcription start) to a *Mun*I site at –3550. A combination of this sequence with that of the 3-kb PCR product yielded a contig that spanned from –3550 to +3017 and. in the regions 1–42 (first exon) and 2713–3655 (part of the second exon), matched precisely the cDNA sequence established with 5′ RACE. Since the cDNA contained an additional 638 nt, the 5′ part of the H-fibroin gene (*Ro03HF5′*) deposited in GenBank under accession no. AB354689 covers the region from –3550 to +3655. The 3′ end of the *R. obliterata H-fibroin* gene (Fig. 5b) was analyzed only at the cDNA level.

The *H-Fibroin* Gene of *H. angustipennis*

Analysis of the *H-fibroin* gene in the *H. angustipennis* larva Ha23 by means of PCR with the cDNA-derived primers HaH-F42 and HaH-R41 (Table 1) yielded a 0.5-kb genomic product. Primers HaH-R51 and HaH-F51 derived from this product were used for inverse PCR with the DNA fragments that had been self-ligated after digestion with the *Hae*II restrictase. A product of 1.5 kb extended the known *H-fibroin* sequence in both directions. Another product of 2.8 kb was obtained from the Ha23 genomic DNA by PCR employing primers HaH-F40 and HaH-R40 (Table 1). A contig combining sequences of the 0.5-, 1.5-, and 2.8-kb products spanned from –1434 to +1476, with a 77-bp intron beginning at +67. The sequence was registered in GenBank as *Ha23HF5′* (accession no. AB354591). The region from about –60 to +70 (end of the first exon) exhibited a high sequence similarity to the *H-fibroin* of *R. obliterata* and, to lesser extent, to that of the moths (Fig. 5a). The distances between the TATA box, transcription start, translation initiation, and the end of the first exon were remarkably similar in all species.

The 3′ region of the *H-fibroin* gene was amplified with the Advantage 2 polymerase mix from the Ha23 genomic DNA self-ligated after *Ban*II digestion. A 3-kb fragment was obtained in a PCR with primers HaH-R46 and HaH-F46. It was cloned and sequenced. The sequence matched

A

```
  -58       v         v         v         v         v         v         v         v         v
Ro tttatttattagaacgcaaatcttgattataaattgagcgagcatttcgataaggcgcATCAGTCATACAAGG-CCCTC----CGCCATGAG   29
Ha ttcagttgaaaacacgaaaaacgggtattataaatgcgggagatttccaagcatcggcATCAGTTCAGCTCCA-CCCTC----CGCCATGAG   29
Gm tacgtcaataatatagcataatttcggtataaaa-gggcctggaaatcagacttcagcATCAGTCCGGTTCCA-CTCTCAATACAATATGAG   33
Bm atacgtcaaaactc-gaaaattttcagtataaaa-aggttcaacttttcaaatcagcATCAGTTCGGTTCCAACTCT-----CAAGATGAG   29

          v         v         v         v         v         v         v         v         v
Ro GGCAATCTTTCTCCTGATCTTAGTCTGCTCACTGCAGgtaaaa<- 2607 bp ->atatatttttttcagATTAACATCGCAGGATCTTCG 2685
Ha GGCGGCAATTCTCCTGATCTTATTCTGCTCATTGCAGgtaagt<--- 56 bp ->ctgctcttattttagATACATTTAACAGGGGCAAAA  134
Gm AGTCACAACCTTCGTGATCTTGTGCTGTGCTCTGCAGgtgagt<- 1289 bp ->ttctcttatttacagTATGTTACTGCAGATGCCATT 1367
Bm AGTCAAAACCTTTGTGATCTTGTGCTGCGCTCTGCAGgtgagt<-- 951 bp ->aacattttgtttcagTATGTCGCTTATACAAATGCA 1029
```

B

```
          v         v         v         v         v         v         v         v         v
Ro  -71 CACAGTGTCCACTTTCAAATTT-CAATGTGCACGTAAAAGTTGGAAATATTCGTAAAGCCAACGGTAATTGTtaagtcattattatggtaca
Ha  -71 GCTCCTGCAGGACTTCTCCATA-CAATCTTTTGATCAAGGTTGGTAATGCTCGCAAGCTTAACGGAAATTGCtaagccttaattgatattca
Ld  -71 CTCCTTGCAAGTTGCC-CGATTTCAACTTGTTCGTGAAAGTGGGAAATGTTCGCAAATCTAACGGCAACTGTtaatctgctttgcccgcaac
Ye  -68 AGCCATGCTACGTCAGCCGGAGGCAACTCACC-GTCAGGGTTGGTAGTCGACGTCAA---TGCGCCTTCTGCtga-agatattatttagacc
Gm  -71 TCCCTTGTAGTCTCACTCGTAGACAATTCGTT-GTTAAGATTGGCACCAGACGCCAACCTTGTGGTTACTGCtaatttgtttggttttttat
Bm  -65 AAAACTGTGGAATTCCTAGAAGACAACTAGTT-GTTAAATTC------AGAGCACTGCCTTGTGTGAATTGCtaatttttaatataaaataa
```

Fig. 5 Similarities in the *H-fibroin* gene among Trichoptera and Lepidoptera (nucleotides conserved in at least two caddisfly species are highlighted in gray); no sequence homologies were detected upstream or downstream from the depicted regions. Exons are in capital letters. **a** The *H-fibroin* 5′ region in *R. obliterata* (Ro; GenBank accession no. AB354689), *H. angustipennis* (Ha; AB354591), *G. mellonella* (Gm; AF095239), and *B. mori* (Bm; AF226688). Sequences are numbered from the transcription start, the TATA box is underlined with a straight line, the initiator consensus sequence (AyCAGyyy) with a wavy line, and the translation initiation codon is double-underlined. Intron donor (gtaagt) and acceptor (tcttattttag) sites are underlined in the Ha sequence. **b** Alignment of the *H-fibroin* 3′ region in *R. obliterata* (Ro; GenBank accession no. AB354588), *H. angustipennis* (Ha; AB214507), *L. decipiens* (Ld; AB214509), *Y. evonymella* (Ye; AB195978), *G. mellonella* (Gm; AF095240), and *B. mori* (Bm; AF226688\0. The termination codon is in boldface

the previously identified cDNA (Yonemura et al. 2006) but continued with 1819 nt beyond its poly(A) attachment site. The genomic sequence representing the 3′ end of *H-fibroin* allele 23 (*Ha23HF3′*) was assigned GenBank accession no. AB354592. The region with similarities to the 3′end of the *H-fibroin* gene of caddisflies and a few moths is shown in Fig. 5b.

Discussion

Conservation of the *L-Fibroin* and *H-Fibroin* Silk Genes in *Trichoptera* and *Lepidoptera*

Analysis of the silk gland-specific cDNAs in *R. obliterata* complements earlier investigations in *H. angustipennis* and *L. decipiens* (Yonemura et al. 2006) and demonstrates that transcripts homologous to the lepidopteran *L-fibroin* and *H-fibroin* genes occur in all three suborders of Trichoptera. This paper shows that the organization of the *L-fibroin* and *H-fibroin* genes in *R. obliterata* and *H. angustipennis* is similar to that of the corresponding lepidopteran genes. Finally, the expression of both genes is restricted to the posterior silk gland section (Yonemura et al. 2006) like in Lepidoptera. These data suggest that ancestral *H-fibroin* and *L-fibroin* genes evolved about 250 mya (million years ago), prior to the separation of Trichoptera and Lepidoptera, and were retained in both orders. Fossil Trichoptera date back to ∼235 mya (Sukacheva 1968, 1973), and fossil Lepidoptera to ∼203 mya (Whalley 1985).

B. mori is the only lepidopteran in which the *L-fibroin* gene was analyzed (GenBank accession no. M76430) and found to contain six introns (Fig. 3). The *L-fibroin* gene of caddisflies *R. obliterata* and *H. angustipennis* (GenBank accession nos. AB354690 and AB354593, respectively) contains five introns that occur at positions comparable to those of introns 1–4 and intron 6 of *B. mori*, indicating that the fifth intron of *B. mori* was inserted into the gene after the moths had split from the caddisflies. The total length of introns is about 12,000 bp in *B. mori,* 9000 in *R. obliterata*, and <1500 in *H. angustipennis*. The first intron is much longer than any other and the fourth is the shortest in all species. Donor sequences at the exon/intron boundaries are close to the AGGTRAGT motif (exon sequence underlined), while the acceptor motif YYYYYYYNCAG is in some cases modified to YYYYYYYNTAG. Sequences TTAATC, TTAACT, TTGCTT, and TTGATG within the introns can be regarded as intron branch points.

The TATA box was identified in the *L-fibroin* gene of *H. angustipennis* at position –29 relative to the transcription start, similarly to *B. mori* (–31). *R. obliterata* contains a TATA box at –286 followed by a putative transcription initiation sequence at –255, but no corresponding transcript was detected. The first Met codon ATG follows 29 nt after the transcription start in *H. angustipennis*, 27 nt in *R. obliterata*, and 42 nt in *B. mori.* The lengths of exons are alike in all species except for an insertion of 71 bp into the second exon and a split of the original fifth exon into two in *B. mori* and a terminal extension of exon 6 in *H. angustipennis.* The sequence of the first exon is about 57% identical in the caddisfly species and different from that in *B. mori.* The last exon in all species contains a putative polyadenylation signal within 20 bp after the termination codon (not shown in Fig. 3) but signals in more distant positions are obviously used for the poly(A) attachment (Fig. 3). The *L-fibroin* gene of *H. angustipennis* harbors three putative signals in the region 2450–2530 (only one of them is indicated in Fig. 3) and the poly(A) chain is attached at 2617. Additional polyadenylation signals occur at 2946 and 2998 and their use is indicated by the finding of cDNAs with a poly(A) tail attached at 2974 and 3032, respectively.

In the 5′ region the *H-fibroin* gene of *R. obliterata* and *H. angustipennis* includes (GenBank accession nos. AB354689 and AB354591, respectively) a short first and a long second exon, similar to the *H-fibroin* gene of the moths, for example, *G. mellonella* and *B. mori* (GenBank accession nos. AF095239 and AF226688). A low but distinct sequence similarity between the *H-fibroins* of caddisflies and moths begins in the 5′ region about 60 nt upstream from the transcription start and increases around the TATA box at –31 in *R. obliterata* and at –30 in the other species (Fig. 5a). Maximal similarity is found in the first exon, which is nearly identical in the moths analyzed (Yonemura and Sehnal 2006). The nucleotide sequence (but not the encoded peptide) of this exon is more diversified among the caddisflies; some regions of the first exon in *H. angustipennis* resemble the *H-fibroin* genes of moths more than that of *R. obliterata.* The nucleotide sequence similarity among and between Trichoptera and Lepidoptera is retained in the exon/intron boundary but rapidly dissipates in the intron. The donor and acceptor sites are conserved and close to the consensus sequences.

Most of the *H-fibroin* second exon is occupied by species-specific repetitive blocks in both Lepidoptera (e.g., Mita et al. 1994; Sezutsu and Yukuhiro 2000; Zhou et al. 2001; Žurovec and Sehnal 2002; Fedič et al. 2003; Yonemura and Sehnal 2006) and Trichoptera (Yonemura et al. 2006; present data). The blocks of 408 nt found in *R. obliterata* are exceptionally uniform. At the end of the coding regions the blocks are replaced by nonrepetitive sequences that exhibit some similarities among the caddisflies and, to a lesser extent, also in the moths (Fig. 5b). The untranslated 3′ terminus contains a single polyadenylation signal 143 nt after the termination codon in *L. decipiens* and 226 nt after in *R. obliterata*; the *H-fibroin*

gene of *H. angustipennis* contains one polyadenylation signal 97 nt after and another one 157 nt after the termination codon.

Conservation of the H-Fibroin and L-Fibroin Proteins

The linkage of L-fibroin and H-fibroin proteins by a disulfide bridge between Cys_{-22} of the H-fibroin (Tanaka et al. 1999b) and Cys_{170} of the L-fibroin (Yamaguchi et al. 1989) proved indispensable for the secretion of both components in *B. mori* (Takei et al. 1987). Silk filament formation in most Lepidoptera further involves a P25 glycoprotein. For *B. mori* it was shown that noncovalent interaction of P25 with the H-fibroin N-terminus is important for gel/filament processing of the hydrophobic H-fibroin/L-fibroin dimers (Tanaka et al. 1999a; Inoue et al. 2000), while L-fibroin protects attachment of the sugar moieties that render P25 hydrophilic (Inoue et al. 2004). Our failure to detect any sequence homologous to P25 among 2304 silk gland-specific cDNAs in *R. obliterata* and the negative results of our previous search in *H. angustipennis* and *L. decipiens* (Yonemura et al. 2006) suggest strongly that this type of protein is lacking in caddisfly silk.

The conservation of certain amino acid motifs in the H-fibroin N-terminus in both moths and caddisflies (Fig. 4a) indicates that the role of this region in fibroin processing is not limited to the interaction with P25. The similarity of signal peptide sequences is obviously related to H-fibroin transport into the endoplasmatic reticulum. We cannot appreciate functions of the motifs like IYEEEIIIK in *R. obliterata* (residues 121–129; Fig. 5a) or somewhat similar sequences of the alternating charged and hydrophobic residues (e.g., KGEIIEKII at positions 62–70 in *R. obliterata*), which occur with slight modifications in all H-fibroins identified so far (Fig. 4a).

The nonrepetitive H-fibroin C-terminus also contains similar features across Trichoptera and Lepidoptera (Fig. 4b). In the caddisfly H-fibroins, a region of about 75 amino acid residues contains relatively high representations of Gly and Pro and a number of single residues or their small groupings, such as GRR, SVYT, HPG, VKVGN, and NGNC, at conserved positions. The spacings of Cys_{-1}, Cys_{-22}, and a few other residues in the C-terminus are similar to those in the lepidopteran H-fibroins. Cys_{-22} may participate in H-fibroin linkage with the L-fibroin but the absence of Cys_{-4} precludes the formation of the intramolecular disulfide bridge Cys_{-1}/Cys_{-4} that is known in *B. mori*.

The L-fibroins of Trichoptera and Lepidoptera exhibit a similar distribution of many residues with characteristic properties such as size, hydrophobicity, and charge (Fig. 1), including three Cys residues whose significance for L-fibroin function was demonstrated in *B. mori* (Tanaka et al. 1999b). Specific features of caddisfly L-fibroins include the conservation of regions with alternating groups of hydrophobic and hydrophilic residues, the presence of three to six additional Cys residues (one of them, Cys_{129} in *R. obliterata*, is conserved in all three caddisfly species), and the occurrence of short and rather unusual motifs such as ProTrpPro (residues 134–136 in *R. obliterata).* The features unique to caddisflies might be important for L-fibroin/H-fibroin interaction in the absence of P25.

Filament Polymerization

The total amino acid composition of the repetitive H-fibroin region is similar in all examined caddisflies and distinct from the known lepidopteran silks. In *R. obliterata* the region contains 28.2% Gly, 15.9% Ser, 9.7% Leu, 8.3 % Ile, 6.4% Val, 8.3% Arg, 6.3% His, and 3.2% Pro (Fig. 4). High representation of Gly and Ser, the bulky hydrophobic residues (Leu, Ile, Val), charged residues (Arg, His, Lys, Asp, and Glu), and Pro was found also in *H. angustipennis* and *L. decipiens* (Yonemura et al. 2006). Another characteristic feature of caddisflies is the low amount of Ala (1.6% in *R. obliterata*), in contrast to its high content in the H-fibroin repeats of Lepidoptera (Sehnal and Sutherland 2008).

Filament formation depends on weak molecular interactions between motifs that are arranged in higher-order repeats (Lucas and Rudall 1968). Comparison of reiterated sequences is the first step to understanding the evolution and mechanisms of filament polymerization. H-Fibroins of all three examined caddisfly species contain similar motifs, but their relative representation, their assembly into repeats, and the regularity of repeats are different. To facilitate their comparison, we designate each repeat with two letters: the first is derived from the family name (Hydropsychidae, Limnephilidae, and Rhyacophilidae, respectively), and the second specifies the type of repeat based on the presence of distinct amino acid motifs and the repeat length. The A, B, and C repeat types were recognized in *H. angustipennis*, the E and F types in *L. decipiens* and *R. obliterata*, and the D type in all three species (Fig. 6). The HA, HC, HD, LD, LE, LF, RE, and RF repeats contain a similar stretch of about 11 amino acid residues with a central Trp. Another conserved region of 27–31 residues occurs in the D-type repeats and includes a central amphiphilic region sandwiched between two Ser-rich motifs. It is likely that both conserved regions, which are absent in the H-fibroins of Lepidoptera, evolved prior to the caddisfly separation into three suborders.

Molecular conformation of the repeats determines physical properties of the silk filament. Conformation of some motifs has been demonstrated or proposed on

```
HA       GPSGLYGDSGIVDGVYGPGLVGPGGWGRRPYGGYSASRSVSAE---GPRGWYGPRGL
HB                                                GPRGLGPLGLDSDIIGDGYGPYGL
HC                                  GPRGLGPLGGLGRRPYGGYSASGSVSAE---GPRGWYGPRGL
HD APVVYHAPIIRRRPKISRSSSYSVERIVAPTVIT----RISGSHSVSAE--GRRGVWGPHGV

LD                    VSISRSVSIERIVTPGIYT----KISRSSSVSVEGGRRRGPWGYGRG
LE                                  LSGSGDLDGLGGVGGLGGLGGLGGRRGPWGRGYG
LF                                        SSGTVSVSVSVEEGRRRGPWGRRGK

RD              SASVSISHSISIERVITPGVYTSIHRSVSVSHSVSVEHRRRIAPIVVDYSI
RF                                        SSSLSISGEGRGLGRGKHGWGGLGH
RE       GGLGGLGHGGLGGLGHGGLGGLGHGGLGGLGSDSGDLGGIGGLLGPGGRRGWGRRGPGKY
```

Fig. 6 Types of H-fibroin repeats identified in the caddisflies *H. angustipennis* (HA-HD), *L. decipiens* (LD-LE), and *R. obliterata* (RD-RE). The alignment of repeats accentuates their similarities. A region conserved in the D repeats of all species is underlined and a motif occurring in most repeats is shaded in gray

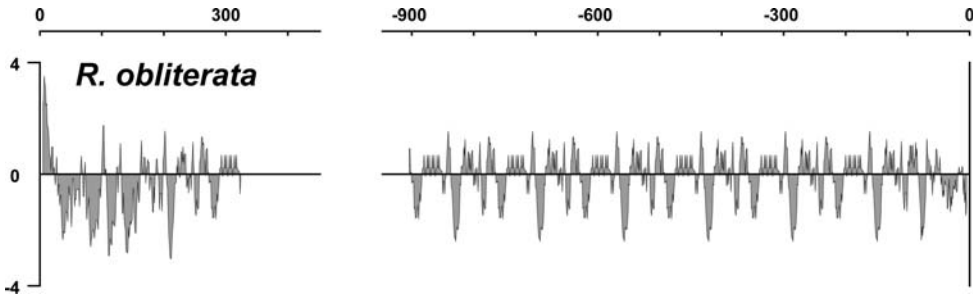

Fig. 7 Kyte-Doolittle hydropathy plot of the N- and C-terminal regions of *R. obliterata* H-fibroin. The first 150 and last 75 residues make up the nonrepetitive ends, while the major central part is composed of repeats

theoretical grounds. For example, the Ser-rich motifs and the strings of $(SX)_n$ in repeats of types A, C, D, E, and F are likely to form rather rigid β-sheets due to hydrogen bonding via polar zipper interactions (Bini et al. 2004). The GGL motif reiterated in the E repeats provides a 3(1) helix conformation, as shown for LGG triplets (Ashida et al. 2003). The GPGXX motif, which occurs in the HA and RE repeats, is believed to form a β-spiral conferring elasticity to the protein polymer (Hayashi and Lewis 1998). Physical properties of the filament are further affected by repeat regularity that ensures precise registration of the interacting motifs (Sehnal and Žurovec 2004). The length of repeats varies in the H-fibroins of *H. angustipennis* and *L. decipiens* but the repeats of *R. obliterata* form very precise modules of 135 residues.

Bini et al. (2004) suggested that proteins forming water-resistant silk filament must include hydrophilic terminal domains flanking a very long central portion constructed from the hydrophobic blocks alternating with short hydrophilic regions. The strong predominance of hydrophobicity in the major polymerizing part was regarded as essential for water exclusion during β-crystallite formation, while the hydrophilic terminal regions allowed fibroin hydration during storage in the silk gland lumen. These rules may apply to aerial silks spun by caterpillars or spiders but are contradicted by the content and distribution of hydrophilic amino acid residues in the repetitive domain of caddisfly H-fibroins. The amphiphilic plot of *R. obliterata* H-fibroin reveals a predominance of hydrophilic regions (Fig. 7), very similar to the situation described for *H. angustipennis* and *L. decipiens* (Yonemura et al. 2006) and different from that in most Lepidoptera (Sehnal and Žurovec 2004).

Acknowledgments We appreciate the help of Dr. K. Novák, who instructed us on when and how to collect *R. obliterata* larvae; several other colleagues helped us with the collections. Dr. C. Hayashi, of University of California, Riverside, kindly read the manuscript and provided very useful comments. The work was supported by a research fellowship awarded to N. Yonemura by the Japan Society for the Promotion of Science and by grant IAA5007402 received by F. Sehnal from the Grant Agency of the Academy of Sciences of the Czech Republic. Part of the work was done in the framework of Project Z50070508 of the Biology Centre ASCR.

References

Akai H, Hakim RS, Kristensen NP (2003) Exocrine glands: saliva and silk. In: Kristensen NP (ed) Lepidoptera: moths and butterflies 2. Handbuch der Zoologie [*Handbook of Zoology*] IV. Walther de Gruyter, Berlin, Vol 36, pp 377–388

Ashida J, Ohgo K, Komatsu K, Kubota A, Asakura T (2003) Determination of the torsion angles of alanine and glycine residues of model compounds of spider silk $(AGG)_{10}$ using solid-state NMR methods. J Biomol NMR 25:91–103

Bini E, Knight DP, Kaplan DL (2004) Mapping domain structures in silks from insects and spiders related to protein assembly. J Mol Biol 335:27–40

Cline J, Braman JC, Hogrefe HH (1996) PCR fidelity of Pfu DNA polymerase and other thermostable DNA polymerases. Nucleic Acids Res 24:3546–3551

Craig CL (1997) Evolution of arthropod silks. Annu Rev Entomol 42:231–267

Engster M (1976) Studies on silk secretion in the Trichoptera (F. Limnephilidae). I. Histology, histochemistry, and ultrastructure of the silk glands. J Morphol 150:183–211

Fedič R, Žurovec M, Sehnal F (2002) The silk of Lepidoptera. J Insect Biotech Sericol 71:1–15

Fedič R, Žurovec M, Sehnal F (2003) Correlation between fibroin amino acid sequence and physical silk properties. J Biol Chem 278:35255–35264

Hayashi CY, Lewis RV (1998) Evidence from flagelliform silk cDNA for the structural basis of elasticity and modular nature of spider silks. J Mol Biol 275:773–784

Inoue S, Tanaka K, Arisaka F, Kimura S, Ohtomo K, Mizuno S (2000) Silk fibroin of *Bombyx mori* is secreted, assembling a high molecular mass elementary unit consisting of H- chain, L-chain, and P25, with a 6:6:1 molar ratio. J Biol Chem 275:40517–40528

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 (2004) Assembly of the silk fibroin elementary unit in endoplasmic reticulum and a role of L-chain for protection of α1, 2-mannose residues in N-linked oligosaccharide chains of fibrohexamerin/P25. Eur J Biochem 271:356–366

Lucas F, Rudall KM (1968) Extracellular fibrous proteins: The silks. In: Florkin M, Stota EH (eds) Comprehensive biochemistry, vol 26B. Elsevier, Amsterdam, pp 475–558

Mita K, Ichimura S, James TC (1994) Highly repetitive structure and its organization of the silk fibroin gene. J Mol Evol 38:583–592

Mita K, Morimyo M, Okano K, Koike Y, Nohara J, Kwasaki H, Kadono-Okuda K, Yamamoto K, Suzuki MG, Shimada T, Goldsmith MR (2003) The construction of an EST database for *Bombyx mori* and its application. Proc Natl Acad Sci USA 100:14121–14126

Sehnal F, Akai H (1990) Insects silk glands: their types, development and function, and effects of environmental factors and morphogenetic hormones on them. Int J Insect Morphol Embryol 19:79–132

Sehnal F, Sutherland T (2008) Silks produced by insect labial glands. Prion 2:1–9

Sehnal F, Žurovec M (2004) Construction of silk fiber core in Lepidoptera. Biomacromolecules 5:666–674

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

Sukacheva ID (1968) Mesozoic caddis flies (Trichoptera) of Transbaikalia. Paleontol J 1968(2):202–216

Sukacheva ID (1973) New caddis-flies (Trichoptera) from the Mesozoinc of Soviet Central Asia. Paleontol J 1973(3):377–384

Takei F, Kikuchi Y, Kikuchi A, Mizuno S, Shimura K (1987) Further evidence for importance of the subunit combination of silk fibroin in its efficient secretion from the posterior silk gland cells. J Cell Biol 105:175–180

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

Tanaka K, Mizuno S (2001) Homologues of fibroin L-chain and P25 of *Bombyx mori* are present in *Dendrolimus spectabilis* and *Papilio xuthus* but not detectable in *Antheraea yamamai*. Insect Biochem Mol Biol 31:665–677

Tanaka K, Kajiyama N, Ishikura K, Waga S, Kikuchi A, Ohtomo K, Takagi T, Mizuno S (1999a) Determination of the site of disulfide linkage between heavy and light chains of silk fibroin produced by *Bombyx mori*. Biochim Biophys Acta 1432:92–103

Tanaka K, Inoue S, Mizuno S (1999b) Hydrophobic interaction of P25, containing Asn-linked oligosaccharide chains, with the H-L complex of silk fibroin produced by *Bombyx mori*. Insect Biochem Mol Biol 29:269–276

Whalley PES (1986) A review of the current fossil evidence of Lepidoptera in the Mesozoic. Biol J Linn Soc 28:253–271

Yamaguchi K, Kikuchi Y, Takagi T, Kikuchi A, Oyama F, Shimura K, Mizuno S (1989) Primary structure of the silk fibroin light chain determined by cDNA sequencing and peptide analysis. J Mol Biol 210:127–139

Yonemura N, Sehnal F (2006) The design of silk fiber composition in moths has been conserved for more than 150 million years. J Mol Evol 63:42–53

Yonemura N, Sehnal F, Mita K, Tamura T (2006) Protein composition of silk filaments spun under water by caddisfly larvae. Biomacromolecules 7:3370–3378

Zaretschnaya SN (1965) Glands of caddisworms. III. Spinning glands. Plankton and Benthos of Inland Water Reservoirs. Proc Acad Sci USSR 12:293–303

Zhou CZ, Confalonieri F, Jacquet M, Perasso R, Li ZG, Janin J (2001) Silk fibroin: structural implications of a remarkable amino acid sequence. Proteins Struct Funct Genet 448:119–122

Žurovec M, Sehnal F (2002) Unique molecular architecture of silk fibroin in the waxmoth, *Galleria mellonella*. J Biol Chem 277:22639–22647