Molecular cloning, gene expression analysis, and recombinant protein expression of novel silk proteins from larvae of a retreat-maker caddisfly, *Stenopsyche marmorata*

Xue Bai [a], Mayo Sakaguchi [a], Yuko Yamaguchi [a], Shiori Ishihara [a], Masuhiro Tsukada [a], Kimio Hirabayashi [a,b], Kousaku Ohkawa [c], Takaomi Nomura [a], Ryoichi Arai [a,d]*

[a] Division of Applied Biology, Faculty of Textile Science and Technology, Shinshu University, Ueda, Nagano 386-8567, Japan

[b] Institute of Mountain Science, Interdisciplinary Cluster for Cutting Edge Research, Shinshu University, Minamiminowa, Nagano 399-4598, Japan

[c] Institute for Fiber Engineering, Interdisciplinary Cluster for Cutting Edge Research, Shinshu University, Ueda, Nagano 386-8567, Japan

[d] Institute for Biomedical Sciences, Interdisciplinary Cluster for Cutting Edge Research, Shinshu University, Matsumoto, Nagano 390-8621, Japan

* Corresponding author. Division of Applied Biology, Faculty of Textile Science and Technology, Shinshu University, 3-15-1 Tokida, Ueda, Nagano 386-8567, Japan.

E-mail address: rarai@shinshu-u.ac.jp (R. Arai).

Highlights

- The cDNAs of novel silk proteins were cloned from the caddisfly *Stenopsyche marmorata*.
- The Smsp-2 has an array of GYD-rich repeat motifs and two $(SX)_4E$ motifs.
- The Smsp-4 has a number of GW-rich repeat motifs and three $(SX)_4E$ motifs.
- Gene expression of Smsp-2 and Smsp-4 varied seasonally.
- Recombinant protein expression of Smsps was successfully performed in *E. coli*.

Graphical abstract

Novel silk proteins from a caddisfly, *Stenopsyche marmorata*

Smsp-2

MKFFVFLFAICLVLAVTDAC
GGGGYYGGYGDYGGYGDYGG
YGCGHAVAAKT*SCSTSTSVE*
TRHVGCGGYYGDYDDGYYGG
YDGGYYGGYGDGYYGDYDGG
YYGGYGDGYYGDYDDGYYGG
YYGCGRPCCRPRPPPVVKT*S*
*ISTSTSVE*TKFMRRRPCFSP
CASPCGY

Smsp-4

MKFFAFLFLACLAFVATDAC
AHVGGYWPVGRG*SASHSVSW*
*E*HGGWGGWHGGWGGWYPGCG
WGYPGYHG*SASHSVSWE*HNG
WGGWHGGRGGWYPGCGWGRW
GHWGWPQHKW*SASHSASWE*N
GGWVRPACGCHW

ABSTRACT

Retreat-maker larvae of *Stenopsyche marmorata,* one of the major caddisfly species in Japan, produce silk threads and adhesives to build food capture nets and protective nests in water. Research on these underwater adhesive silk proteins potentially leads to the development of new functional biofiber materials. Recently, we identified four major *S. marmorata* silk proteins (Smsps), Smsp-1, Smsp-2, Smsp-3, and Smsp-4 from silk glands of *S. marmorata* larvae. In this study, we cloned full-length cDNAs of Smsp-2, Smsp-3, and Smsp-4 from the cDNA library of the *S. marmorata* silk glands to reveal the primary sequences of Smsps. Homology search results of the deduced amino acid sequences indicate that Smsp-2 and Smsp-4 are novel proteins. The Smsp-2 sequence [167 amino acids (aa)] has an array of GYD-rich repeat motifs and two $(SX)_4E$ motifs. The Smsp-4 sequence (132 aa) contains a number of GW-rich repeat motifs and three $(SX)_4E$ motifs. The Smsp-3 sequence (248 aa) exhibits high homology with fibroin light chain of other caddisflies. Gene expression analysis of Smsps by real-time PCR suggested that the gene expression of Smsp-1 and Smsp-3 was relatively stable throughout the year, whereas that of Smsp-2 and Smsp-4 varied seasonally. Furthermore, Smsps recombinant protein expression was successfully performed in *Escherichia coli*. The study provides new molecular insights into caddisfly aquatic silk and its potential for future applications.

Keywords: caddisfly; Trichoptera; *Stenopsyche marmorata*; silk protein; aquatic silk; Smsp.

1. Introduction

Caddisflies (order Trichoptera) are a large group of aquatic insects. *Stenopsyche marmorata* (suborder Annulipalpia, called "retreat-maker") is one of the most common large caddisflies in rivers in Japan, such as the Chikuma (Shinano) River [1]. The larvae feed, mature, and pupate underwater and spin aquatic adhesive silk to build essential structures including food capture nets and protective nests [2,3]. Research on silk proteins from caddisfly larvae could lead to novel biopolymer materials for underwater adhesive and biomedical purposes [4]. Recently, we identified four major *S. marmorata* silk proteins (Smsps) extracted from the silk glands of *S. marmorata* larvae [5]. The high-molecular-mass protein (>~300 kDa) was designated as Smsp-1; the three low-molecular-mass proteins were designated as Smsp-2, Smsp-3, and Smsp-4 (~26–32 kDa, ~21–26 kDa, and ~16–17 kDa, respectively, in SDS–PAGE analysis) [3,5,6]. We also reported the biochemical characterization and amino acid sequences of Smsp-1 that are the major component of the larval net silk/adhesive precursor [3,6]. However, detailed molecular information on Smsp-2, Smsp-3, and Smsp-4 is still unknown. To reveal the molecular characteristics of the aquatic silk proteins, we first report the primary sequences of Smsp-2, Smsp-3, and Smsp-4 deduced from their full-length cDNA clones, which were isolated from the cDNA library of the *S. marmorata* silk glands. In addition, we performed gene expression analysis and recombinant protein expression of Smsps.

2. Materials and methods

2.1 N-terminal amino acid sequencing of Smsps by Edman degradation

The fifth instar larvae of *S. marmorata* were collected in the middle reaches of the Chikuma River in Nagano Prefecture, Japan. The silk glands were dissected from the larvae; the major silk proteins (P3′ fraction) were extracted from the silk glands as previously

reported [5,6]. The proteins were separated by SDS–PAGE and blotted onto a PVDF membrane. The individual bands of Smsp-2, Smsp-3, and Smsp-4 were cut out from the membrane; the N-terminal amino acid sequencing was performed on an automated protein sequencer PPSQ-21 (Shimadzu).

2.2 cDNA cloning of Smsps

The cDNA libraries were constructed from silk glands of *S. marmorata* larvae as previously described [6]. In brief, the silk glands were dissected from the larvae; the total RNA from the silk glands was isolated using the ISOGEN reagent (Nippon Gene). The polyA$^+$ RNAs were purified using the Oligotex-dT30 <Super> mRNA purification kit (Takara Bio). The cDNA libraries were constructed using the cDNA library construction kit with a cloning vector pAP3neo (Takara Bio) or the CloneMiner cDNA library construction kit with a cloning vector pDONR222 (Life Technologies).

The 3′ end cDNA fragments of Smsp-2 and Smsp-4 were amplified from the cDNA library (pAP3neo) by polymerase chain reaction (PCR) using KOD -Plus- Neo DNA polymerase (Toyobo) with the degenerate primer, Smsp-2N-FW or Smsp-4N-FW, and the vector-specific primer, T3promoter(pAP3neo). (Table S1.) Each of the amplified DNA fragments of Smsp-2 and Smsp-4 was cloned into the plasmid vector pBluescript II KS(+) (Agilent Technologies) and digested with *Eco*RV by blunt-end ligation. DNA sequence analysis was performed using Applied Biosystems 3130xl Genetic Analyzer (Life Technologies). The 5′ end cDNA fragments were amplified from the cDNA library by PCR with the specifically designed primer, Smsp-2-RV1 or Smsp-4-RV1, and the vector-specific primer, T7FWlongpAP3neo. The amplified 5′ end cDNA fragments of Smsp-2 and Smsp-4 were cloned into pBluescript II KS(+); DNA sequences were analyzed. The 5′ and 3′ ends of cDNA sequences were assembled into the full-length cDNA sequences.

The cDNA clone of Smsp-3 was isolated from the cDNA library (pDONR222) using PCR with the vector-specific primer M13FW and the designed degenerate primer Smsp-3-RV1.

The determined cDNA sequences were registered in DNA data bank of Japan (DDBJ) with accession numbers, LC057251 for Smsp-2, LC057252 for Smsp-3, and LC057253 for Smsp-4.

2.3 Western blotting to detect phosphoserine

Western blot analysis was performed using rabbit anti-phosphoserine (Life Technologies) primary and $F(ab')_2$-goat anti-rabbit IgG (H+L) HRP-conjugated (Life Technologies) secondary antibodies. Immunoreactions were visualized by Immobilon Western Chemiluminescent HRP Substrate (Merck Millipore).

2.4 Mass spectrometry

After SDS–PAGE of the P3′ fraction from the silk glands (Fig. S1), the gel bands of Smsp-3 and Smsp-4 stained by the reverse staining method [7] were excised, destained, and crushed. The whole proteins were eluted from the gel fragments by formic acid/water/2-propanol (1:3:2 v/v/v) [8] and analyzed using a MALDI-TOF mass spectrometer, TOF/TOF 5800 System (AB SCIEX) using a sinapic acid matrix.

2.5 Gene expression analysis of Smsps by real-time PCR

Fifth instar larvae of *S. marmorata* were collected around Kamuriki Bridge in the middle part of the Chikuma River in Nagano Prefecture, Japan, every month from May to December 2012 (Table S2). The silk glands were dissected from the larvae; total RNA from the silk glands was extracted using a High pure RNA tissue kit (Roche Diagnostics). Quantitative real-time PCR was performed on a MiniOpticon Real-Time PCR System (Bio-Rad). The

cDNA template, the forward and reverse primers for each Smsp (Table S3), and SYBR Premix Ex Taq II (Tli RNaseH Plus) (Takara Bio) in a total of 25 µl were applied to the following PCR programs: 95°C for 30 s (initial denaturation); 95°C for 5 sec and 60°C for 30 s, repeated for 40 cycles (amplification). The ribosomal protein genes *rpL11 and rpL31* from *S. marmorata* (DDBJ accession numbers LC057254 and LC057255, respectively) were used for normalization.

2.6 Recombinant protein expression in Escherichia coli

Each of the cDNA encoding Smsps without the secretory signal sequence was cloned into pENTR-TEVL, a modified pENTR vector with a Tobacco Etch Virus (TEV) protease cleavage site, derived from pENTR1A (Life Technologies). The cDNA of Smsp-1 (clone 3-54) encoded the C-terminal fragment region, Smsp-1c (479 residues) [6]. The cDNA of Smsp-4 was modified with codon optimization for expression in *E. coli*. The protein expression vectors pDEST17-Smsps and pCold-TF-Smsps were constructed using GATEWAY technology (Life Technologies) with pENTR-TEVL-Smsps and pDEST17 (Life Technologies) or pCold-TF-GW, which was the modified pCold TF (Takara Bio) with a GATEWAY reading frame cassette (Life Technologies). The Smsps with a His_6 tag (H-Smsps) were expressed in *E. coli* BL21 Star (DE3) (Life Technologies) harboring pDEST17-Smsps (with a T7 promoter) using LB broth (50 µg/mL ampicillin) at 37°C for ~7 h. For the Smsps fusion protein with a His_6 and trigger factor (TF) tag (TF-Smsps), *E. coli* BL21 Star (DE3) harboring pCold-TF-Smsps (cold shock promoter) was cultured using LB broth (50 µg/mL ampicillin) at 37°C for ~2 h. At OD_{660} = ~0.5, expression was induced with 50 µM isopropyl β-D-1-thiogalactopyranoside (IPTG) and cold shock at 15°C; the cells were further cultured for 16 h at 15°C. The harvested cells were disrupted by sonication in a lysis/wash buffer (50 mM sodium phosphate buffer (pH 7.0) containing 300 mM NaCl, 10%

glycerol). The soluble and insoluble fractions were separated by centrifugation. The proteins including H-Smsps in the insoluble fractions were solubilized with a solubilization/wash buffer (20 mM Tris–HCl buffer (pH 8.0) containing 8 M urea, 1 mM dithiothreitol, and 1 mM EDTA). The proteins were purified by immobilized metal ion affinity chromatography (IMAC) with cOmplete His-tag purification resin (Roche Diagnostics) for H-Smsps or TALON metal affinity resin (Takara Bio) for TF-Smsps; the proteins were eluted using 250 mM imidazole in the lysis/wash buffer or the solubilization/wash buffer.

3. Results and discussion

3.1 cDNA cloning and amino acid sequences of Smsp-2 and Smsp-4

For cDNA cloning of Smsp-2 and Smsp-4, the degenerate primers, Smsp-2N-FW and Smsp-4N-FW (Table S1), were designed based on the N-terminal amino acid sequences of Smsp-2 and Smsp-4 by Edman degradation sequencing (Table S4). The full-length cDNA clones of Smsp-2 and Smsp-4 (Figs. S2 and S3) were successfully isolated from the cDNA library of the *S. marmorata* silk glands by PCR cloning. Homology search results of basic local alignment search tool (BLAST) [9] showed no homologous proteins with significant similarity, indicating that Smsp-2 and Smsp-4 are novel proteins.

Figure 1A shows the deduced amino acid sequence of Smsp-2. The full-length Smsp-2 protein is composed of 167 amino acid residues including the N-terminal secretory signal sequence (19 residues) predicted by SignalP [10] (Fig. S4A). The subsequent sequence is essentially consistent with the N-terminal amino acid sequence (Fig. 1A and Table S4). The experimental analysis of the amino acid composition indicated that Smsp-2 highly contained Gly (25.4 mol%), Tyr (16.9 mol%), and Asx (9.9 mol%) [3,6], which is consistent with that deduced from the cDNA of Smsp-2 (Table S5). The molecular mass of Smsp-2 without the signal sequence is calculated to be 15.7 kDa, which is lower than ~26–32 kDa estimated by

SDS–PAGE, possibly implying potential post-translational modifications and/or unusual mobility shift on electrophoresis. Smsp-2 has an array of unusual GYD-rich repeat motifs comprising Gly, Tyr, and Asp, and two $(SX)_4E$ motifs (Fig. 1A and S5A). Our recent study on Smsp-1 showed that many Ser residues of $(XS)_n$ motif in Smsp-1 were *O*-phosphorylated, and the *O*-phosphoserine residue occurred in a clustered manner, probably serving a cement function for Smsp-1 [6]. In addition, a repeating $(SX)_n$ motif conserved in the fibroin heavy chain (H-fibroin) of a case-maker larva of caddisfly (suborder Integripalpia), *Hesperophylax consimilis* (corrected in the subsequent paper [11]), was also densely phosphorylated [12]. These data suggest that the Ser residues of the $(SX)_4E$ motifs of Smsp-2 are potential *O*-phosphorylation sites.

Figure 1B shows the deduced amino acid sequence of Smsp-4. The full-length Smsp-4 protein is composed of 132 amino acid residues including the N-terminal secretory signal sequence (19 residues), which is suggested by the N-terminal amino acid sequence (Table S4) and predicted by SignalP [10] as the second candidate (Fig. S4C). The subsequent sequence is mostly consistent with the N-terminal amino acid sequence (Fig. 1B and Table S4). The experimental analysis of the amino acid composition indicated that Smsp-4 highly contained Gly (20.3 mol%), Trp (8.9 mol%), His (7.5 mol%), and Ser (7.3 mol%) [3,6], which is roughly consistent with that deduced from the cDNA of Smsp-4 (Table S5). The molecular mass of Smsp-4 without the signal sequence is calculated to be 12.4 kDa, which is a roughly reasonable value compared to ~16–17 kDa estimated by SDS–PAGE considering experimental errors and/or potential post-translational modifications. As shown in Figures 1B and S5B, Smsp-4 has a significant number of characteristic GW-rich repeat motifs comprising Gly and Trp. Also Smsp-4 has three $(SX)_4E$ motifs, which are suggested as potential *O*-phosphorylation sites. Western blot analysis with an anti-phosphoserine antibody suggested that Smsp-4 was phosphorylated at Ser residues (Fig. 1C). Furthermore, the

MALDI–TOF mass spectrum of the Smsp-4 protein extracted from the silk glands (Fig. 1D) shows that the first main peak of *m/z* 13440.5 is probably assignable to a fully phosphorylated Smsp-4 monomer (~13.4 kDa) considering mass increase due to phosphorylation of twelve Ser residues (+80 Da × 12 = +960 Da: twelve Ser residues of potential phosphorylation sites at three of the $(SX)_4E$ motif in Smsp-4). The asymmetric shape of the relatively-broad mass peak with a lower-mass tail probably suggests its heterogeneity of partial phosphorylation. Also, the second and third peaks of *m/z* 26934.3 and 40509.9 are assignable to the phosphorylated Smsp-4 dimer and trimer, respectively, suggesting that Smsp-4 potentially forms homo-oligomers possibly because the GW-rich repeats interact with each other by stacking and hydrophobic interactions of Trp residues.

3.2 cDNA cloning and amino acid sequence of Smsp-3

As in the case of Smsp-2 and Smsp-4, we first tried cDNA cloning of Smsp-3 from the cDNA library by PCR with the degenerate primers, Smsp-3N-FW or Smsp-3N-FW2 (Table S1), designed based on the N-terminal amino acid sequence of Smsp-3 (Table S4). However, we attempted unsuccessfully to obtain a cDNA fragment of Smsp-3. Then, we changed to another strategy for the cDNA cloning of Smsp-3. Because the N-terminal amino acid sequence of Smsp-3 shares significant homology with fibroin light chain (L-fibroin) from other caddisflies (Fig. S6A), we tried cDNA cloning of L-fibroin from *S. marmorata* using the sequence alignment information of L-fibroin from other caddisflies. We designed another degenerate primer, Smsp-3-RV1 (Table S1), based on the highly homologous sequence region "NN(V/I)GAAATSAAT" found in the sequence alignment of L-fibroin from three caddisfly species, *Limnephilus decipiens*, *Rhyacophila obliterate*, and *Hydropsyche angustipennis* [13,14] (Fig. S6B). The full-length cDNA clone of *S. marmorata* L-fibroin (Fig. S7) was successfully isolated from the cDNA library of the silk glands by PCR with the

new degenerate primer. Finally, we confirmed that Smsp-3 was identical to *S. marmorata* L-fibroin because the tryptic-digested Smsp-3 was significantly identified as *S. marmorata* L-fibroin (the amino acid sequence deduced from the cloned cDNA) by using tandem mass spectrometry (Fig. S8).

Figure 2 shows the deduced amino acid sequence of Smsp-3 (*S. marmorata* L-fibroin) on the sequence alignment of L-fibroin from other caddisfly species. The full-length Smsp-3 protein is composed of 248 amino acid residues including the predicted N-terminal secretory signal sequence (18 residues) (Fig. S4B). The molecular mass of Smsp-3 without the signal sequence is calculated to be 24.1 kDa, which corresponds to ~21–26 kDa estimated by SDS–PAGE. In addition, the molecular mass of the whole protein of Smsp-3 extracted from the silk gland was measured by MALDI–TOF mass spectrometry (Fig. S9). The doublet peak of *m/z* 23291.8 and 23482.6 is probably assignable to the Smsp-3 monomer considering experimental errors. As shown in Figure 2, Smsp-3 shares significantly high homology with L-fibroin from other caddisflies, *H. angustipennis* (71% identities), *Hesperophylax occidentalis* (54%), *L. decipiens* (54%), and *R. obliterata* (51%), suggesting that Smsp-3 has the same function as L-fibroin of other caddisflies.

3.3 Gene expression analysis by real-time PCR

Using real-time PCR, we performed gene expression analysis of Smsps in the silk glands from natural larvae samples taken every month from May to December (Table S2) (Fig. 3). The results show that the Smsp-1 and Smsp-3 genes were expressed at relatively stable levels, irrespective of seasons. In contrast, the expression levels of the Smsp-2 and Smsp-4 genes were higher in summer and lower in winter (significant difference between July–August and November–December in Smsp-2 ($p < 0.05$) and in Smsp-4 ($p < 0.01$)). However, the expression pattern of Smsp-2 was somewhat elusive, and it was possibly controlled by

on/off-like regulation. These results suggest that Smsp-1 and Smsp-3 play fundamental roles as backbones of the silk protein complex through all seasons. However, Smsp-2 and Smsp-4 may play additional roles because the Smsp-4 expression varied seasonally, and the Smsp-2 expression was strictly controlled as may be necessary.

3.4 Recombinant protein expression of Smsps in E. coli

We constructed recombinant protein expression systems of Smsps in *E. coli* to facilitate further research and future applications. First, we tried T7 expression system for Smsps with a His_6 tag (H-Smsps). All H-Smsps were expressed in the insoluble fractions with and without IPTG induction (Fig. S10A). H-Smsp-3 and H-Smsp-4 were highly expressed. The H-Smsps proteins were successfully solubilized and purified by IMAC under denaturing conditions with 8 M urea (Fig. 4A). Second, we tried cold shock expression system for Smsps with a His_6 tag and TF tag (TF-Smsps). TF-Smsp-2, TF-Smsp-3, and TF-Smsp-4 were expressed in soluble fractions with IPTG and cold shock induction (Fig. S10B). TF-Smsp-3 and TF-Smsp-4 were highly expressed in the soluble fractions and successfully purified by IMAC under native conditions (Fig. 4B). However, TF-Smsp-2 was hardly purified and TF-Smsp-1c could not be purified in the soluble fractions, probably due to protein degradation.

3.5 A putative complex model of Smsps with various interactions

The gene expression analysis suggests that Smsp-1 and Smsp-3 play fundamental roles as backbones of the silk protein complex through all seasons. In the domestic silkworm *Bombyx mori*, H-fibroin and L-fibroin form a complex by a disulfide linkage between Cys-c20 of H-fibroin and Cys-172 of L-fibroin [15]. Amino acid sequences of H-fibroin and L-fibroin from *B. mori* around these regions were homologous to H-fibroin and L-fibroin from

caddisflies, *H. angustipennis* and *L. decipiens* [13], and also to Smsp-1 and Smsp-3 from *S. marmorata*, respectively (Fig. S11). The two Cys residues of Cys-c20 (twentieth amino acids from C-terminal) of H-fibroin and Cys-172 of L-fibroin from *B. mori* are strictly conserved among these caddisflies including *S. marmorata*, suggesting that Smsp-1 and Smsp-3 also potentially form a heterodimeric complex by a disulfide linkage between the two Cys residues (Fig. S11).

The western blot analysis with anti-phosphoserine antibody (Fig. 1C) and the mass spectrum of Smsp-4 (Fig. 1D) suggest that the Ser residues of Smsp-4 are phosphorylated probably at the $(SX)_4E$ motif, similar to Smsp-1. In our previous study, the addition of EDTA induced the separation of Smsp-1 from the other proteins containing Smsp-2, Smsp-3, and Smsp-4 [6]. These results suggest that the complex formation of Smsp-1 and Smsp-4 is driven by cross-bridging of the anionic phosphoserine clusters, the $(pSX)_n$ motifs of Smsp-1 and Smsp-4 together with the cationic metal ions such as Ca^{2+}. In the silk from a case-maker caddisfly, *H. consimilis*, the structural model, in which the phosphorylated serine repeats $(pSX)_4$ complex with divalent cations Ca^{2+} and Mg^{2+} to form rigid nanocrystalline β-sheet structures, was also reported [16].

The Smsp-2 expression was strictly controlled, suggesting that Smsp-2 is not a major component of the silk but an additional and optional factor. The newly-discovered GYD-rich repeat motif of Smsp-2 implies its unique function, which is probably related to its surprisingly-high composition of Tyr (20.3%) (Table S5). Recently, dityrosine crosslinking, catalyzed by peroxinectin in the adhesive underwater silk of a case-maker caddisfly, *H. occidentalis*, was reported [17]. The Smsp-2 with many Tyr residues at the GYD-rich repeat motif may be a potential substrate protein dityrosine-crosslinked by peroxinectin to enhance the molecular network and mechanical properties of caddisfly silk fibers.

The Smsp-4 expression varied seasonally and tended to be higher in summer and lower in winter, suggesting that Smsp-4 plays more important roles in summer than in winter. Since the caddisfly larvae live an active life in summer, they produce more silk fibers in summer than in winter, implying that Smsp-4 may be associated with efficient silk production. In the silkworm *B. mori*, the silk fibroin is efficiently secreted from the posterior silk gland as an elementary unit, 2.3 MDa protein complex, comprising six sets of a disulfide-linked H-fibroin–L-fibroin heterodimer and one molecule of fibrohexamerin/P25 [18]. In the caddisfly *S. marmorata*, Smsp-1 forms a large complex with Smsp-2, Smsp-3, and Smsp-4 as previously reported [5]. In addition, the present study suggests that the Smsp-4 oligomer potentially assembles a fundamental complex of Smsps with several sets of a disulfide-linked Smsp-1–Smsp-3 heterodimer.

In the present study, we successfully performed cDNA cloning, gene expression analysis, and recombinant protein expression of Smsps including the unusual novel proteins, Smsp-2 and Smsp-4. These results provide new molecular information and insights into a relatively unexplored field of aquatic silk proteins.

Acknowledgments

The authors thank Ms. Hikari Zukeran, Mr. Aki Nishimura, Mr. Tetsuya Musha, Mr. Naohiko Watanabe, Ms. Misaki Nakayama, Ms. Yumi Miura, Ms. Mai Kanamori, and Mr. Masaaki Takeda at Shinshu University for collecting the caddisfly samples; Ms. Eri Ishikawa at Shinshu University for assistance in mass spectrometry; Dr. Mikako Shirouzu and Dr. Shigeyuki Yokoyama at RIKEN for providing the modified expression vectors. We are indebted to Divisions of Gene Research and Instrumental Analysis, Research Center for Human and Environmental Sciences, Shinshu University, for providing facilities. This work was supported by JSPS KAKENHI Grant Numbers, 22113508, 24113707, and 24780097 to

RA; 22510028 to KH; 22350103 and 23651083 to KO; and 22580060 and 26288101 to MT. This study was also supported by Program for Dissemination of Tenure-Track System funded by MEXT and Exploratory Research Grant for Young Scientists funded by Shinshu University to RA. Finally, this paper is dedicated to the memory of our great colleague, Prof. Koji Abe, who was one of founders of this research project.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.bbrc.2015.07.041.

References

[1] G. Kimura, E. Inoue, K. Hirabayashi, Seasonal abundance of adult caddisfly (Trichoptera) in the middle reaches of the Shinano river in central Japan, in: W.H. Robinson, D. Bajomi (Eds.) Proc. 6th Int. Conf. Urban Pests, OOK-Press Kft, Budapest, 2008, pp. 259-266.

[2] M. Tsukada, M.M. Khan, E. Inoue, G. Kimura, J.Y. Hun, M. Mishima, K. Hirabayashi, Physical properties and structure of aquatic silk fiber from *Stenopsyche marmorata*, Int. J. Biol. Macromol., 46 (2010) 54-58.

[3] K. Ohkawa, T. Nomura, R. Arai, K. Abe, M. Tsukada, K. Hirabayashi, Characterization of Underwater Silk Proteins from Caddisfly Larva, *Stenopsyche marmorata*, in: T. Asakura, T. Miller (Eds.) Biotechnology of Silk, Biologically-Inspired Systems 5, Springer, Dordrecht, 2014, pp. 107-122.

[4] M. Tszydel, A. Zablotni, D. Wojciechowska, M. Michalak, I. Krucinska, K. Szustakiewicz, M. Maj, A. Jaruszewska, J. Strzelecki, Research on possible medical use of silk produced by caddisfly larvae of *Hydropsyche angustipennis* (Trichoptera, Insecta), J. Mech. Behav. Biomed. Mater., 45 (2015) 142-153.

[5] K. Ohkawa, Y. Miura, T. Nomura, R. Arai, K. Abe, M. Tsukada, K. Hirabayashi, Isolation of silk proteins from a caddisfly larva, *Stenopsyche marmorata*, J. Fiber Bioeng. Inform., 5 (2012) 125-137.

[6] K. Ohkawa, Y. Miura, T. Nomura, R. Arai, K. Abe, M. Tsukada, K. Hirabayashi, Long-range periodic sequence of the cement/silk protein of *Stenopsyche marmorata*: purification and biochemical characterisation, Biofouling, 29 (2013) 357-367.

[7] C. Fernandez-Patron, M. Calero, P.R. Collazo, J.R. Garcia, J. Madrazo, A. Musacchio, F. Soriano, R. Estrada, R. Frank, L.R. Castellanos-Serra, et al., Protein reverse staining: high-efficiency microanalysis of unmodified proteins detected on electrophoresis gels, Anal. Biochem., 224 (1995) 203-211.

[8] S.L. Cohen, B.T. Chait, Mass spectrometry of whole proteins eluted from sodium dodecyl sulfate-polyacrylamide gel electrophoresis gels, Anal. Biochem., 247 (1997) 257-267.

[9] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res., 25 (1997) 3389-3402.

[10] T.N. Petersen, S. Brunak, G. von Heijne, H. Nielsen, SignalP 4.0: discriminating signal peptides from transmembrane regions, Nat. Methods, 8 (2011) 785-786.

[11] N.N. Ashton, D.S. Taggart, R.J. Stewart, Silk tape nanostructure and silk gland anatomy of trichoptera, Biopolymers, 97 (2012) 432-445.

[12] R.J. Stewart, C.S. Wang, Adaptation of caddisfly larval silks to aquatic habitats by phosphorylation of H-fibroin serines, Biomacromolecules, 11 (2010) 969-974.

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

[14] N. Yonemura, K. Mita, T. Tamura, F. Sehnal, Conservation of silk genes in Trichoptera and Lepidoptera, J. Mol. Evol., 68 (2009) 641-653.

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

[16] J.B. Addison, N.N. Ashton, W.S. Weber, R.J. Stewart, G.P. Holland, J.L. Yarger, beta-Sheet nanocrystalline domains formed from phosphorylated serine-rich motifs in caddisfly larval silk: a solid state NMR and XRD study, Biomacromolecules, 14 (2013) 1140-1148.

[17] C.S. Wang, N.N. Ashton, R.B. Weiss, R.J. Stewart, Peroxinectin catalyzed dityrosine crosslinking in the adhesive underwater silk of a casemaker caddisfly larvae, *Hysperophylax occidentalis*, Insect Biochem. Mol. Biol., 54 (2014) 69-79.

[18] S. Inoue, K. Tanaka, F. Arisaka, S. Kimura, K. Ohtomo, S. Mizuno, 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 (2000) 40517-40528.

Figure legends

Fig. 1. The novel silk proteins Smsp-2 and Smsp-4 from *S. marmorata*. (A) The deduced amino acid sequence of Smsp-2. The secretory signal sequence predicted by SignalP is underlined. The GYD-rich repeat motif and the $(SX)_4E$ motif are shown in blue bold and red italic fonts, respectively. (B) The deduced amino acid sequence of Smsp-4. The predicted secretory signal sequence is underlined. The GW-rich repeat motif and the $(SX)_4E$ motif are shown in green bold and red italic fonts, respectively. (C) Western blot analysis of Smsp-4. Phosphorylation of Ser was detected with an anti-phosphoserine antibody (right). All proteins on the PVDF membrane were stained with Ponceau S (left). M: molecular mass marker; P3′: the P3′ fraction from the *S. marmorata* silk glands [5,6]. (D) The MALDI–TOF mass spectrum of the Smsp-4 protein extracted from the *S. marmorata* silk glands.

Fig. 2. The deduced amino acid sequence of Smsp-3 (*S. marmorata* L-fibroin) on the sequence alignment of L-fibroin from other caddisfly species. The secretory signal sequence of Smsp-3 predicted by SignalP is underlined in red. The amino acid residues completely conserved are highlighted in black. The amino acid residues conserved with 60–80% similarities are highlighted in gray. Accession numbers for L-fibroin: *Hydropsyche angustipennis,* BAF62094 [13]; *Hesperophylax occidentalis,* AIO11229 [17]; *Limnephilus decipiens,* BAF62096 [13]; *Rhyacophila obliterate,* BAH80180 [14].

Fig. 3. The gene expression analysis of Smsps from natural samples taken every month from May to December (Table S2). The expression data were normalized to the value of Smsp-1 in May. Error bars represent ±1 standard deviation (n = 2–4).

Fig. 4. Recombinant protein expression of Smsps in *E. coli*. (A) SDS–PAGE (15% gel) of His_6-tagged Smsps (H-Smsps). (B) SDS–PAGE (10% gel) of trigger factor-tagged Smsps (TF-Smsps). M: molecular mass marker; S: solubilized samples from the insoluble fractions; E: eluted samples after IMAC purification; L: cell lysate samples. Proteins were stained with Coomassie brilliant blue.

Footnotes

Abbreviations: DDBJ, DNA data bank of Japan, H-fibroin, fibroin heavy chain; HRP, horseradish peroxidase; IMAC, immobilized metal ion affinity chromatography; IPTG, isopropyl β-D-1-thiogalactopyranoside; L-fibroin, fibroin light chain; MALDI–TOF, matrix-assisted laser desorption/ionization time-of-flight; SDS–PAGE, sodium dodecyl sulfate–polyacrylamide gel electrophoresis; Smsp, *Stenopsyche marmorata* silk protein; TEV, Tobacco Etch Virus; TF, trigger factor.

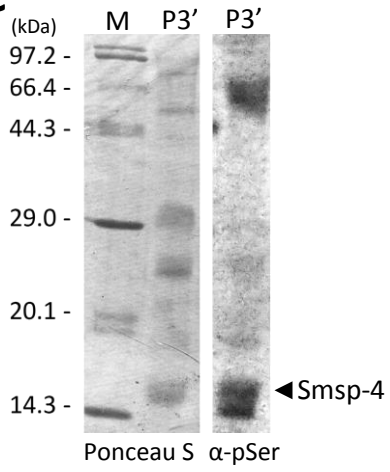
A **Smsp-2**

MKFFVFLFAICLVLAVTDAC
GGGGYYGGYGDYGGYGDYGG
YGCGHAVAAKT*SCSTSTSVE*
TRHVGCGGYYGDYDDGYYGG
YDGGYYGGYGDGYYGDYDGG
YYGGYGDGYYGDYDDGYYGG
YYGCGRPCCRPRPPPVVKT*S*
*ISTSTSVE*TKFMRRRPCFSP
CASPCGY

B **Smsp-4**

MKFFAFLFLACLAFVATDAC
AHVGGYWPVGRG*SASHSVSW*
*E*HGGWGGWHGGWGGWYPGCG
WGYPGYHG*SASHSVSWE*HNG
WGGWHGGRGGWYPGCGWGRW
GHWGWPQHKW*SASHSASWE*N
GGWVRPACGCHW

C



D

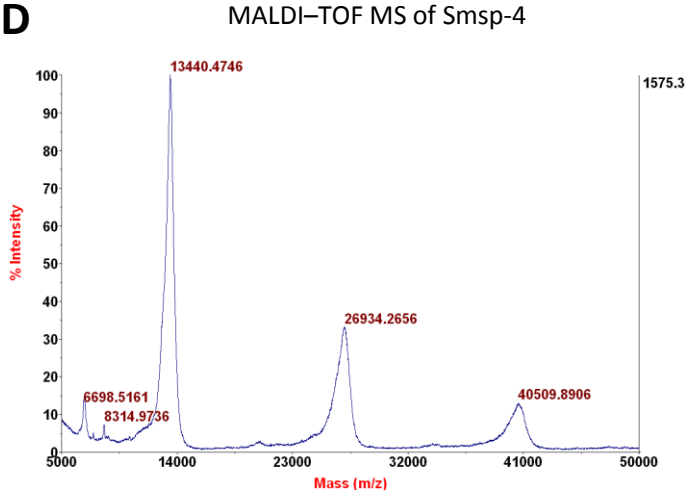


Fig. 1

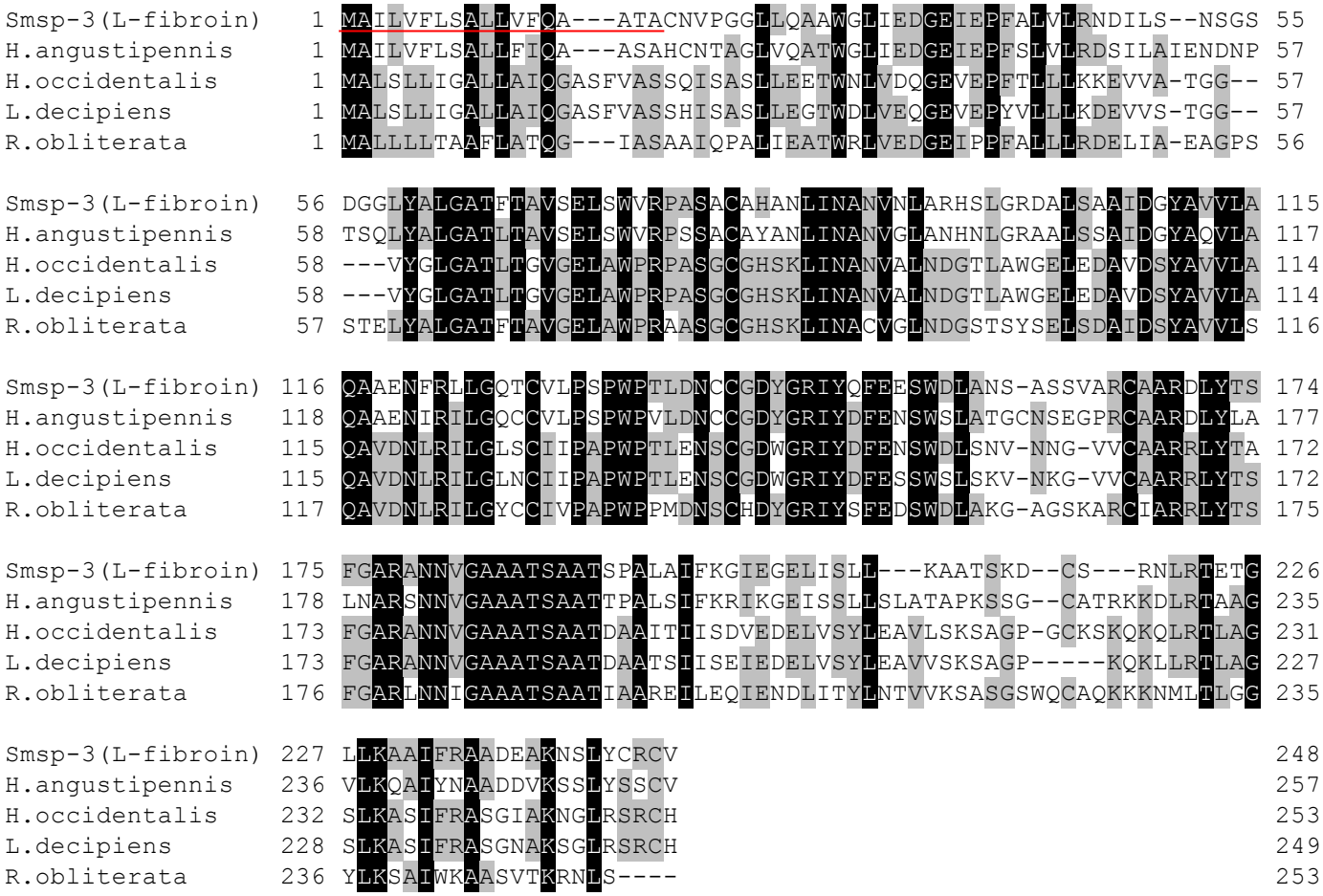

Smsp-3(L-fibroin) 1 MAILVFLSALLVFQA---ATACNVPGGLLQAAWGLIEDGEIEPFALVLRNDILS--NSGS 55
H.angustipennis 1 MAILVFLSALLFIQA---ASAHCNTAGLVQATWGLIEDGEIEPFSLVLRDSILAIENDNP 57
H.occidentalis 1 MALSLLIGALLAIQGASFVASSQISASLLEETWNLVDQGEVEPFTLLLLKKEVVA-TGG-- 57
L.decipiens 1 MALSLLIGALLAIQGASFVASSHISASLLEGTWDLVEQGEVEPYVLLLKDEVVS-TGG-- 57
R.obliterata 1 MALLLLTAAFLATQG---IASAAIQPALIEATWRLVEDGEIPPFALLLRDELIA-EAGPS 56
Smsp-3(L-fibroin) 56 DGGLYALGATFTAVSELSWVRPASACAHANLINANVNLARHSLGRDALSAAIDGYAVVLA 115
H.angustipennis 58 TSQLYALGATLTAVSELSWVRPSSACAYANLINANVGLANHNLGRAALSSAIDGYAQVLA 117
H.occidentalis 58 ---VYGLGATLTGVGELAWPRPASGCGHSKLINANVALNDGTLAWGELEDAVDSYAVVLA 114
L.decipiens 58 ---VYGLGATLTGVGELAWPRPASGCGHSKLINANVALNDGTLAWGELEDAVDSYAVVLA 114
R.obliterata 57 STELYALGATFTAVGELAWPRAASGCGHSKLINACVGLNDGSTSYSELSDAIDSYAVVLS 116
Smsp-3(L-fibroin) 116 QAAENFRLLGQTCVLPSPWPTLDNCCGDYGRIYQFEESWDLANS-ASSVARCAARDLYTS 174
H.angustipennis 118 QAAENIRILGQCCVLPSPWPVLDNCCGDYGRIYDFENSWSLATGCNSEGPRCAARDLYLA 177
H.occidentalis 115 QAVDNLRILGLSCIIPAPWPTLENSCGDWGRIYDFENSWDLSNV-NNG-VVCAARRLYTA 172
L.decipiens 115 QAVDNLRILGLNCIIPAPWPTLENSCGDWGRIYDFESSWSLSKV-NKG-VVCAARRLYTS 172
R.obliterata 117 QAVDNLRILGYCCIVPAPWPPMDNSCHDYGRIYSFEDSWDLAKG-AGSKARCIARRLYTS 175
Smsp-3(L-fibroin) 175 FGARANNVGAAATSAATSPALAIFKGIEGELISLL---KAATSKD--CS---RNLRTETG 226
H.angustipennis 178 LNARSNNVGAAATSAATTPALSIFKRIKGEISSLLSLATAPKSSG--CATRKKDLRTAAG 235
H.occidentalis 173 FGARANNVGAAATSAATDAAITIISDVEDELVSYLEAVLSKSAGP-GCKSKQKQLRTLAG 231
L.decipiens 173 FGARANNVGAAATSAATDAATSIISEIEDELVSYLEAVVSKSAGP-----KQKLLRTLAG 227
R.obliterata 176 FGARLNNIGAAATSAATIAAREILEQIENDLITYLNTVVKSASGSWQCAQKKKNMLTLGG 235
Smsp-3(L-fibroin) 227 LLKAAIFRAADEAKNSLYCRCV 248
H.angustipennis 236 VLKQAIYNAADDVKSSLYSSCV 257
H.occidentalis 232 SLKASIFRASGIAKNGLRSRCH 253
L.decipiens 228 SLKASIFRASGNAKSGLRSRCH 249
R.obliterata 236 YLKSAIWKAASVTKRNLS---- 253

Fig. 2

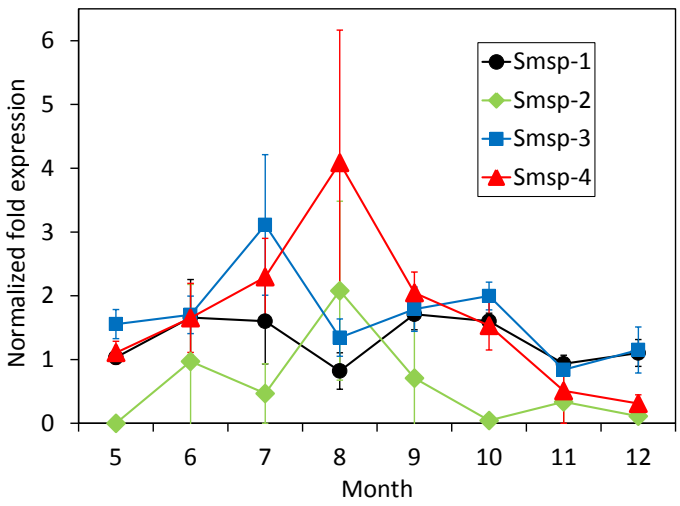

Normalized fold expression
0
1
2
3
4
5
6
5
6
7
8
9
10
11
12
Month
Smsp-1
Smsp-2
Smsp-3
Smsp-4

Fig. 3

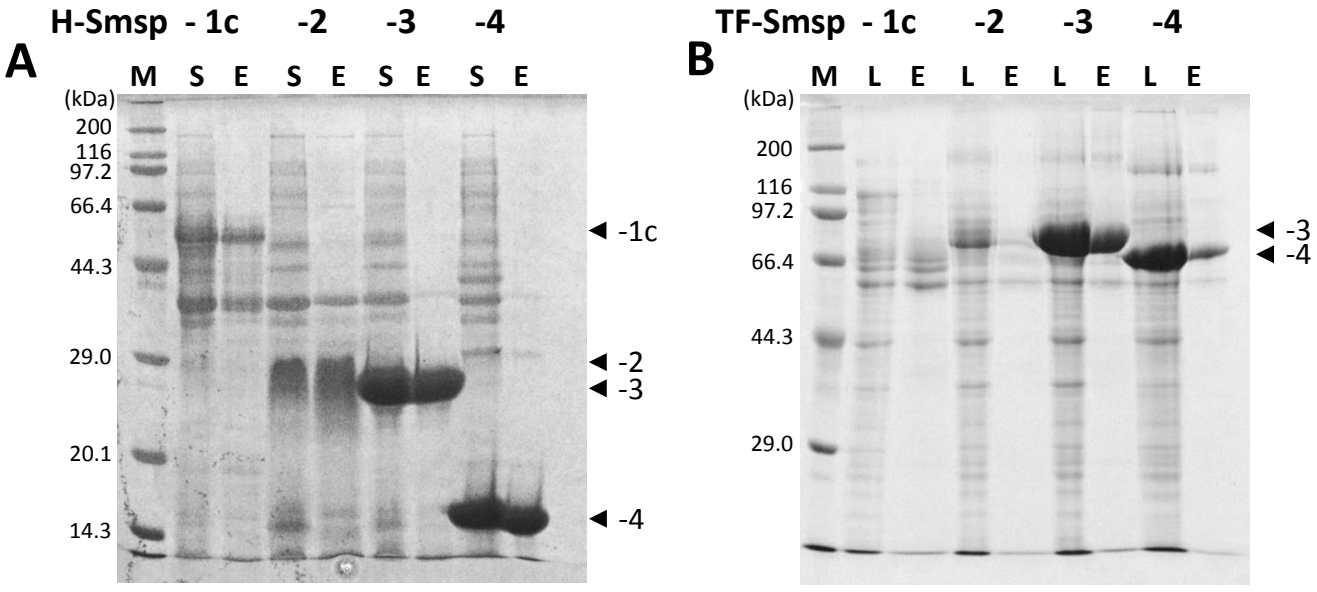

A
H-Smsp - 1c -2 -3 -4
M S E S E S E S E
(kDa)
200
116
97.2
66.4
44.3
29.0
20.1
14.3
-1c
-2
-3
-4
B
TF-Smsp - 1c -2 -3 -4
M L E L E L E L E
(kDa)
200
116
97.2
66.4
44.3
29.0
-3
-4

Fig. 4

Supplementary data

Molecular cloning, gene expression analysis, and recombinant protein expression of novel silk proteins from larvae of a retreat-maker caddisfly, *Stenopsyche marmorata*

Xue Bai [a], Mayo Sakaguchi [a], Yuko Yamaguchi [a], Shiori Ishihara [a], Masuhiro Tsukada [a], Kimio Hirabayashi [a,b], Kousaku Ohkawa [c], Takaomi Nomura [a], Ryoichi Arai [a,d,*]

[a] Division of Applied Biology, Faculty of Textile Science and Technology, Shinshu University, Ueda 386-8567, Japan
[b] Institute of Mountain Science, Interdisciplinary Cluster for Cutting Edge Research, Shinshu University, Minamiminowa, Nagano 399-4598, Japan
[c] Institute for Fiber Engineering, Interdisciplinary Cluster for Cutting Edge Research, Shinshu University, Ueda 386-8567, Japan
[d] Institute for Biomedical Sciences, Interdisciplinary Cluster for Cutting Edge Research, Shinshu University, Matsumoto 390-8621, Japan

*Corresponding author e-mail address: rarai@shinshu-u.ac.jp (R. Arai).

Table S1
Sequences of the primers used for the cDNA cloning experiments in this study.

Primer Name	DNA sequence (5'→3')
M13FW	GTA AAACGACGGCCAGT
M13RV	GGAAACAGCTATGACCATG
Smsp-2N-FW	ATHGGNGGNGGNGGITAYTAYGG
Smsp-3N-FW	ATHGGNTAYATHAARCCIWSIACIGGIGC
Smsp-3N-FW2	GCNGTNTGGGGIYTIATHGARGAYGGIGARATIG
Smsp-3-RV1	GTNGCNGCRCTNGTIGCIGCIGCICCIACRTTRTT
Smsp-4N-FW	GTNGGNGGNTAYTGGCCNGTIGG
T3promoter(pAP3neo)	ATTAACCCTCACTAAAGGGCG
Smsp2RV1	AGGTGGTGGTCGAGGACGG
Smsp1SeqRV1	TCCATATGGTCCAAAATCATCTA
Takara_T7pro_primer	TAATACGACTCACTATAGGG
Smsp-4RV1	GGCAGATCCATGGTATCCTGGATAACCC
T7FWlongpAP3neo	CGGGCCCTTAGGACGCGTAATACGACTCAC
T3RVlongpAP3neo	GATTTAAATTAACCCTCACTAAAGGGCGGC
Smsp3-RV-2	CAATTCCCTTGAAGATAGCCAAAGC

Table S2

Data on the field-work sample collections and environments in the Chikuma (Shinano) River, in the samples for gene expression analysis of Smsps.

Collection date	No. of larvae	Air temp. (℃)	Water temp. (℃)	Flow rate (cm/s)	pH	EC (mS/m)	DO (mg/L)	Water depth (cm)
5/9/2012	3	15.0	14.0	45.6	8.08	13.13	10.97	14.9
6/8/2012	3	25.2	21.2	50.0	9.81	13.72	14.85	24.6
7/11/2012	2	26.3	22.8	42.6	7.96	13.63	9.34	20.0
8/22/2012	4	34.7	26.0	54.0	8.17	17.84	9.98	23.5
9/21/2012	4	19.7	21.1	46.9	7.46	17.8	8.82	20.6
10/26/2012	4	14.4	13.1	60.0	8.22	16.69	11.79	15.5
11/13/2012	4	6.8	7.1	42.1	8.15	17.4	12.53	20.0
12/22/2012	4	−0.7	3.0	47.5	8.13	18.4	13.91	16.4

EC: electric conductivity; DO: dissolved oxygen.

Table S3

Sequences of the primers used for gene expression analysis of Smsps in this study.

Target gene		DNA sequence (5'→3')
Smsp-1	Forward	GTCGCACCTTTGGTTTATGG
	Reverse	AGTCCTGGGATCAGGATGTCT
Smsp-2	Forward	TCGACCACCACCTGTAGTCA
	Reverse	AGGCAGGATGTGTCAGGAAG
Smsp-3	Forward	CCTCACCTGCTTTGGCTATC
	Reverse	TCGGTTCTCAAGTTCCTGCT
Smsp-4	Forward	GGTAAGGAGCGCACTTGCTA
	Reverse	CCAAGTAGTCAATTATGCCATCAG
rpL11	Forward	TCATTGCACAGTTCGAGGAG
	Reverse	CTTGGATACCGAAACCGAAA
rpL31	Forward	GCCTCAACAAATTCCTCTGG
	Reverse	GGCTGAATCTTCATCGTCGT

Table S4

N-terminal amino acid sequences of Smsps by Edman degradation sequencing.

Protein name	N-terminal amino-acid sequences
Smsp-2	IGGGGYYGGY
Smsp-3	IGYIKPSTGAVWGLIEDGEIGPFAL
Smsp-4	XAXVGGYWPVGGG

The underlined sequence regions were used to design the degenerate primers for cDNA cloning.

Table S5
Amino acid compositions (mol%) of Smsp-2 and Smsp-4.

	Smsp-2		Smsp-4	
	Experimental	Deduced	Experimental	Deduced
Ala (A)	4.6	2.7	8.3	5.3
Arg (R)	5.7	4.7	6.4	3.5
Asx (D+N)	**9.9**	**8.8**	4.4	1.8
Cys (C)	0.4	6.1	0.7	4.4
Glx (E+Q)	4.3	1.4	4.4	3.5
Gly (G)	**25.4**	**27.7**	**20.3**	**28.3**
His (H)	2.1	1.4	**7.5**	**10.6**
Ile (I)	1.5	1.4	1.2	0.0
Leu (L)	1.5	0.0	1.5	0.0
Lys (K)	4.1	2.0	6.3	0.9
Met (M)	0.6	0.7	0.2	0.0
Phe (F)	2.0	1.4	1.2	0.0
Pro (P)	6.4	5.4	8.1	5.3
Ser (S)	4.2	6.8	**7.3**	**10.6**
Thr (T)	4.8	5.4	2.3	0.0
Trp (W)	0.9	0.0	**8.9**	**16.8**
Tyr (Y)	**16.9**	**20.3**	6.3	4.4
Val (V)	4.7	4.1	4.6	4.4

Experimental: the amino acid composition obtained by the experimental analysis [3,6].
Deduced: the amino acid composition deduced from the cloned cDNA sequences in this study.

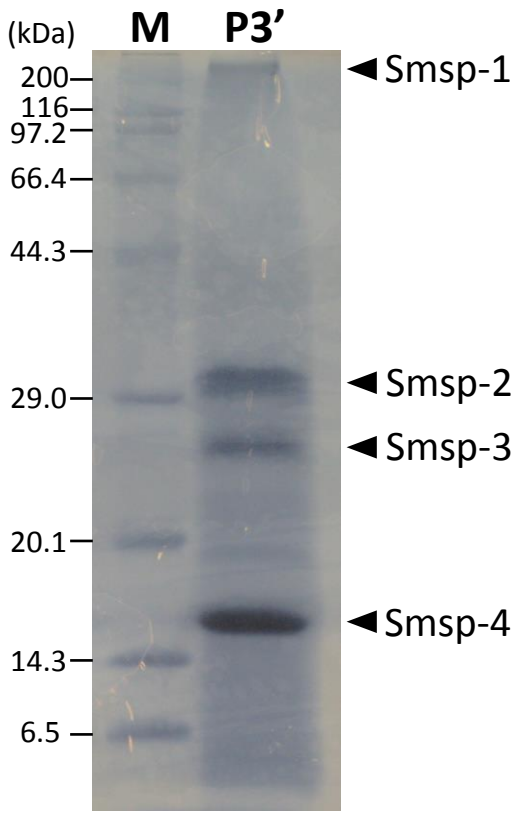


Fig. S1. SDS-PAGE analysis (20% gel) of the P3′ fraction of the major silk proteins extracted from the silk glands [5,6]. The proteins were stained by the reverse staining method [7] for the subsequent mass spectrometry analysis.

```
  1  TTTGGTTGGATTCATATCTGCTGAAATTACCAAGATGAAGTTCTTCGTATTCTTGTTTGC     60
  1                                  M  K  F  F  V  F  L  F  A      9

 61  CATCTGCTTGGTCCTTGCGGTGACCGATGCTTGTGGCGGTGGTGGATATTATGGTGGCTA    120
 10    I  C  L  V  L  A  V  T  D  A  C  G  G  G  G  Y  Y  G  G  Y     29

121  TGGAGACTATGGCGGATATGGAGACTATGGCGGATATGGATGTGGACATGCTGTCGCAGC    180
 30    G  D  Y  G  G  Y  G  D  Y  G  G  Y  G  C  G  H  A  V  A  A     49

181  CAAAACCTCATGCAGTACCAGCACAAGCGTTGAAACTCGCCATGTAGGTTGTGGCGGATA    240
 50    K  T  S  C  S  T  S  T  S  V  E  T  R  H  V  G  C  G  G  Y     69

241  TTATGGAGACTATGATGATGGTTATTATGGCGGCTATGATGGCGGTTATTATGGCGGCTA    300
 70    Y  G  D  Y  D  D  G  Y  Y  G  G  Y  D  G  G  Y  Y  G  G  Y     89

301  TGGTGATGGTTATTATGGCGACTATGATGGCGGTTATTATGGCGGCTATGGTGATGGATA    360
 90    G  D  G  Y  Y  G  D  Y  D  G  G  Y  Y  G  G  Y  G  D  G  Y    109

361  TTATGGTGACTATGATGATGGTTATTATGGTGGCTATTATGGTTGTGGAAGACCTTGCTG    420
110    Y  G  D  Y  D  D  G  Y  Y  G  G  Y  Y  G  C  G  R  P  C  C    129

421  CCGTCCTCGACCACCACCTGTAGTCAAGACATCAATCAGTACCAGCACAAGCGTTGAAAC    480
130    R  P  R  P  P  P  V  V  K  T  S  I  S  T  S  T  S  V  E  T    149

481  AAAGTTCATGAGAAGAAGACCTTGTTTTTCACCCTGTGCATCCCCCTGTGGTTATTAGTT    540
150    K  F  M  R  R  R  P  C  F  S  P  C  A  S  P  C  G  Y  *      168

541  CTGCACAATCTTCCTGACACATCCTGCCTTTGCATAAGACGATCGAATTCTATTATTTAA    600

601  GCTTTTTTTTTTTTTCATCTCCTTAGTATACATTTTAAATAAAAGCATATTTTCTGAAGC    660

661  ATAAAAAAAAAAAAAAAA    678
```

Fig. S2. The cloned cDNA sequence of Smsp-2 consists of 34 bp of 5'-UTR (1–34), 504 bp of coding region (35–538), and 140 bp of 3'-UTR with a poly A tail (539–678) (DDBJ accession no. LC057251**).** The Kozak translation start consensus sequence, which is consistent with the Kozak sequence [(C/A)AA(A/C)ATG] from an insect, *Drosophila* [19], is shown in red. The predicted polyadenylation signal sequence (AATAAA) is underlined. The deduced amino acid sequence is represented with single-letter codes below the corresponding triplet codons.

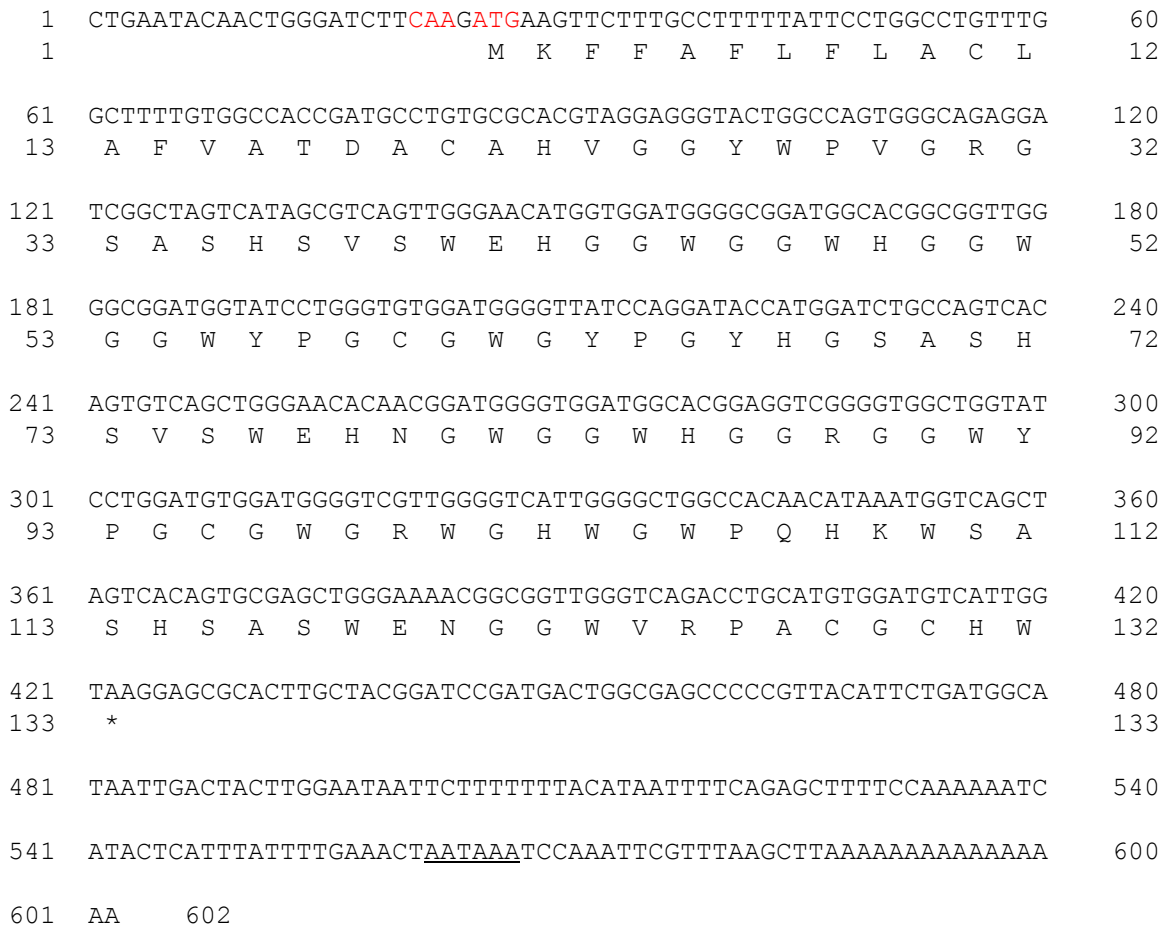
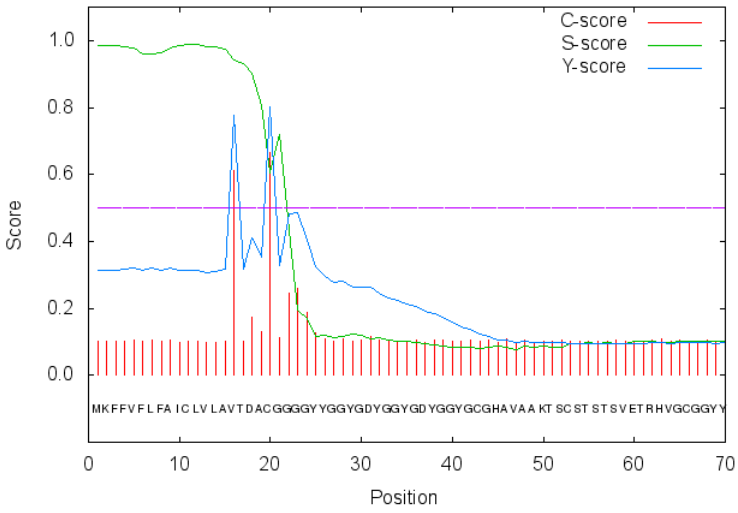


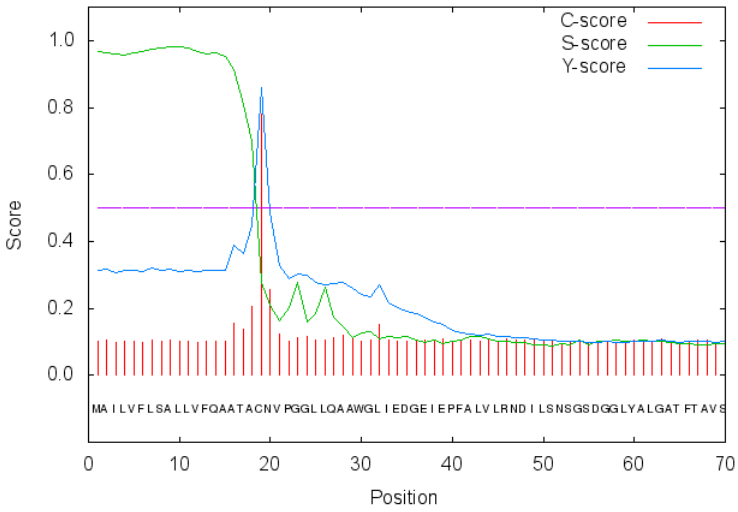
Fig. S3. The cloned cDNA sequence of Smsp-4 consists of 24 bp of 5′ UTR (1–24), 399 bp of coding region (25–423), and 179 bp of 3′ UTR with a poly A tail (424–602) (DDBJ accession no. LC057253). The Kozak translation start consensus sequence, which is consistent with the Kozak sequence [(C/A)AA(A/C)ATG] from an insect, *Drosophila* [19], is shown in red. The predicted polyadenylation signal sequence (AATAAA) is underlined. The deduced amino acid sequence is represented with single-letter codes below the corresponding triplet codons.

A Smsp-2 SignalP-4.1 prediction (euk networks): Smsp2



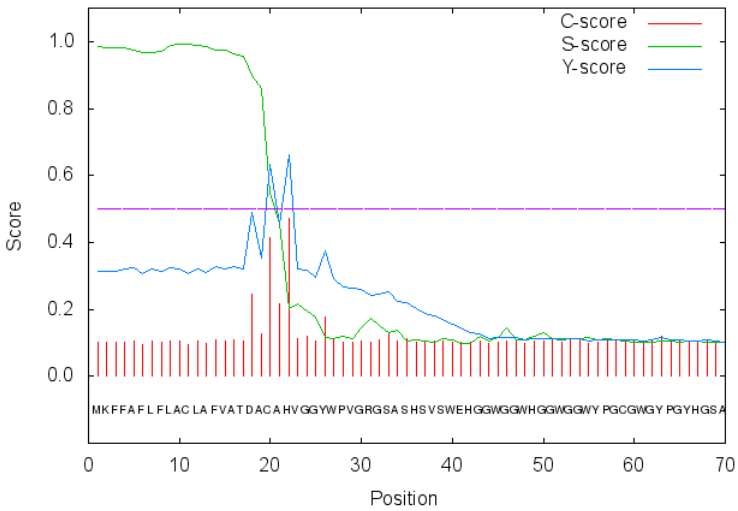
Signal peptide? 'YES'
Cleavage site between pos. 19 and 20: TDA-CG
D=0.887 (D-cutoff=0.450)

B Smsp-3 SignalP-4.1 prediction (euk networks): Smsp3



Signal peptide? 'YES'
Cleavage site between pos. 18 and 19: ATA-CN
D=0.904 (D-cutoff=0.450)

C Smsp-4 SignalP-4.1 prediction (euk networks): Smsp4



Signal peptide? 'YES'
Cleavage site between pos. 21 and 22: ACA-HV or between pos. 19 and 20: TDA-CA
D=0.803 (D-cutoff=0.450)

Fig. S4. Secretory signal sequence prediction of Smsp-2 (A), Smsp-3 (B), and Smsp-4 (C) using SignalP 4.1 server [10].

A **Smsp-2**

MKFFVFLFAICLVLAVTDA
CGGGGYYGGYGD
YGGYGD
YGGYGCGHAVAAKT
SCSTSTSVETRHVGCGGYYGDYDDG
YYGGYDGG
YYGGYGDG
YYGDYDGG
YYGGYGDG
YYGDYDDG
YYGGYYGCGRPCCRPRPPPVVKT
SISTSTSVETKFMRRRPCFSPCASPCGY

B **Smsp-4**

MKFFAFLFLACLAFVATDA
CAHVGGYWPVGRG
SASHSVSWEHGGWGGW
HGGWGGWYPGCGWG
YPGYHG
SASHSVSWEHNGWGGW
HGGRGGWYPGCGW
GRWGHWGWPQHKW
SASHSASWENGGWVRPACGCHW

Fig. S5. The significant patterns of the novel repeat motifs in the amino acid sequences of the novel silk proteins Smsp-2 and Smsp-4 from *S. marmorata*. The repeat motif patterns are emphasized by line feed and alignment. (A) The deduced amino acid sequence of Smsp-2. The secretory signal sequence predicted by SignalP [10] is underlined. The GYD-rich repeat motif and the potentially-phosphorylated $(SX)_4E$ motif are shown in blue bold and red italic fonts, respectively. (B) The deduced amino acid sequence of Smsp-4. The secretory signal sequence predicted by SignalP is underlined. The GW-rich repeat motif and the potentially-phosphorylated $(SX)_4E$ motif are shown in green bold and red italic fonts, respectively.

A

Download GenPept Graphics

fibroin light chain [Hydropsyche angustipennis]

Sequence ID: dbj|BAF62094.1| Length: 257 Number of Matches: 1

See 1 more title(s)

Range 1: 28 to 43 GenPept Graphics Next Match Previous Match

Score	Expect	Identities	Positives	Gaps
43.5 bits(95)	0.003	13/16(81%)	13/16(81%)	0/16(0%)

```
Query  10  AVWGLIEDGEIGPFAL  25
          A WGLIEDGEI PF L
Sbjct  28  ATWGLIEDGEIEPFSL  43
```

Download GenPept Graphics

fibroin light chain [Rhyacophila obliterata]

Sequence ID: dbj|BAH80185.1| Length: 253 Number of Matches: 1

Range 1: 28 to 43 GenPept Graphics Next Match Previous Match

Score	Expect	Identities	Positives	Gaps
36.3 bits(78)	0.69	12/16(75%)	13/16(81%)	0/16(0%)

```
Query  10  AVWGLIEDGEIGPFAL  25
          A W L+EDGEI PFAL
Sbjct  28  ATWRLVEDGEIPPFAL  43
```

B

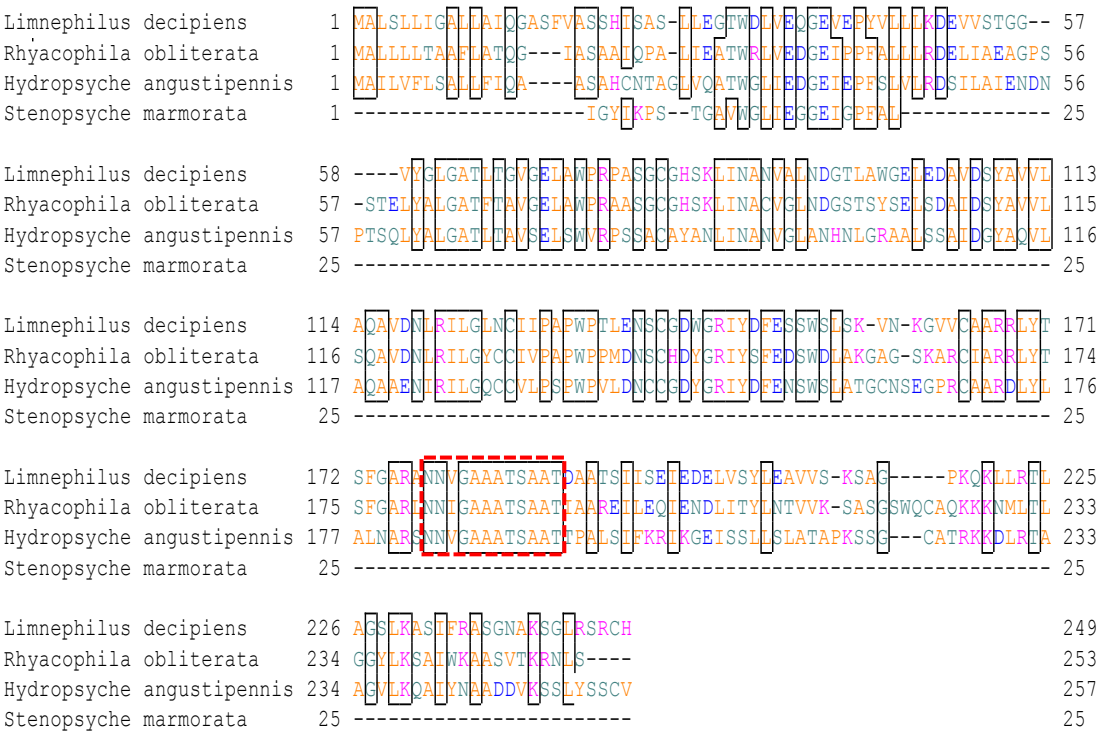


Fig. S6. Homology information on fibroin light chain (L-fibroin) from caddisflies for cloning of Smsp-3. (A) The BLAST homology search results of the N-terminal amino acid sequence of Smsp-3 by Edman sequencing. It shares significant homology with L-fibroin from other caddisflies. (B) The sequence alignment of L-fibroin from three caddisfly species, *Limnephilus decipiens*, *Rhyacophila obliterate*, and *Hydropsyche angustipennis* [13,14] and the N-terminal amino-acid sequence of Smsp-3 from *Stenopsyche marmorata*. The highly homologous sequence region, used for design of the degenerate primer, Smsp-3-RV1, is surround with red broken lines.

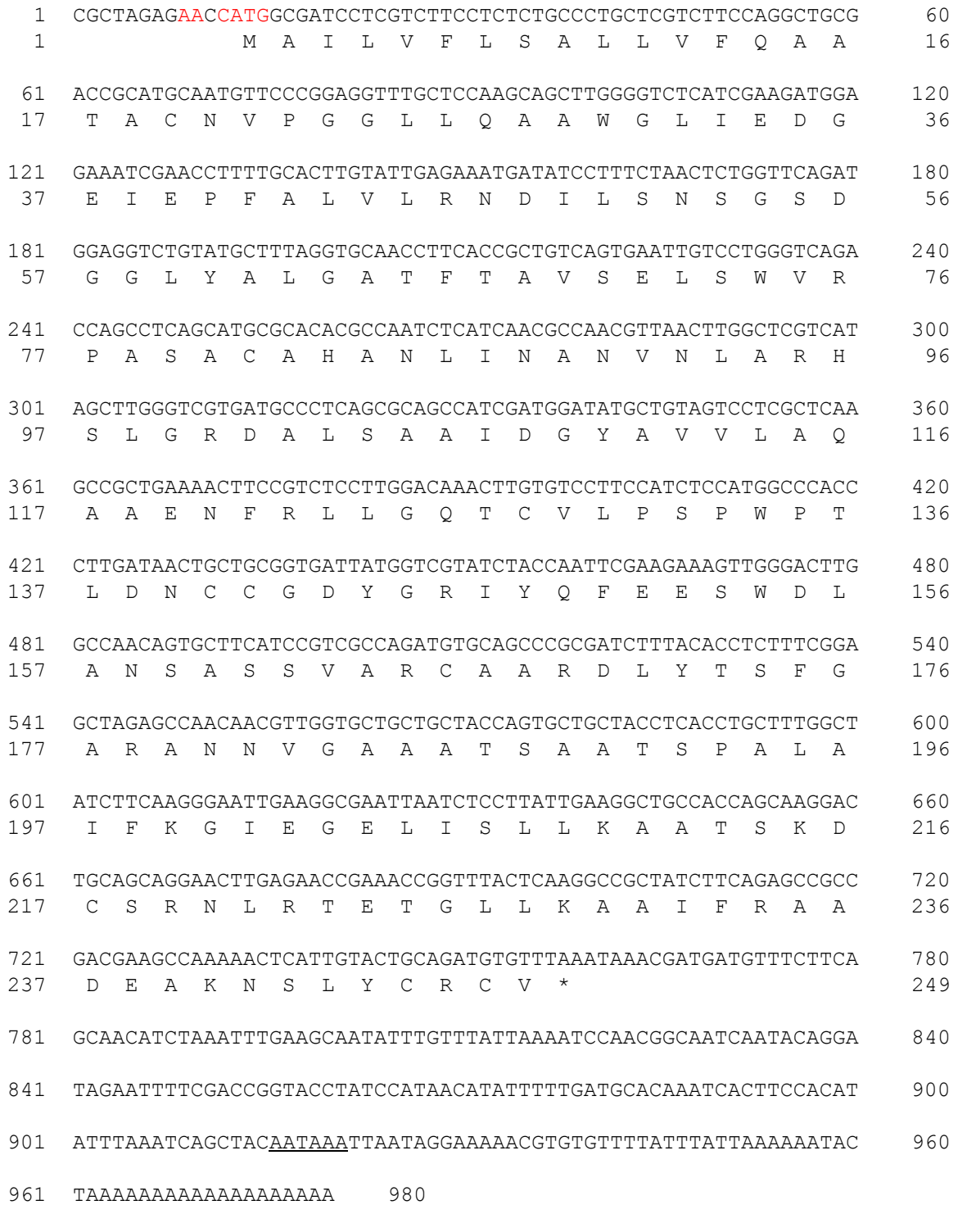


Fig. S7. The cloned cDNA sequence of Smsp-3 (*S. marmorata* L-fibroin) consists of 12 bp of 5′ UTR (1–12), 747 bp of coding region (13–759), and 221 bp of 3′ UTR with a poly A tail (760–980) (DDBJ accession no. LC057252). The Kozak consensus sequence flanking translation start sites, which is consistent with the Kozak sequence [(C/A)AA(A/C)ATG] from an insect, *Drosophila* [19], is shown in red. The predicted polyadenylation signal sequence (AATAAA) is underlined. The deduced amino acid sequence is represented with single-letter codes below the corresponding triplet codons.

A

Protein name	PLGS score	Matched peptides	Coverage (%)
S. marmorata L-fibroin	33437	67	49.2

B

C

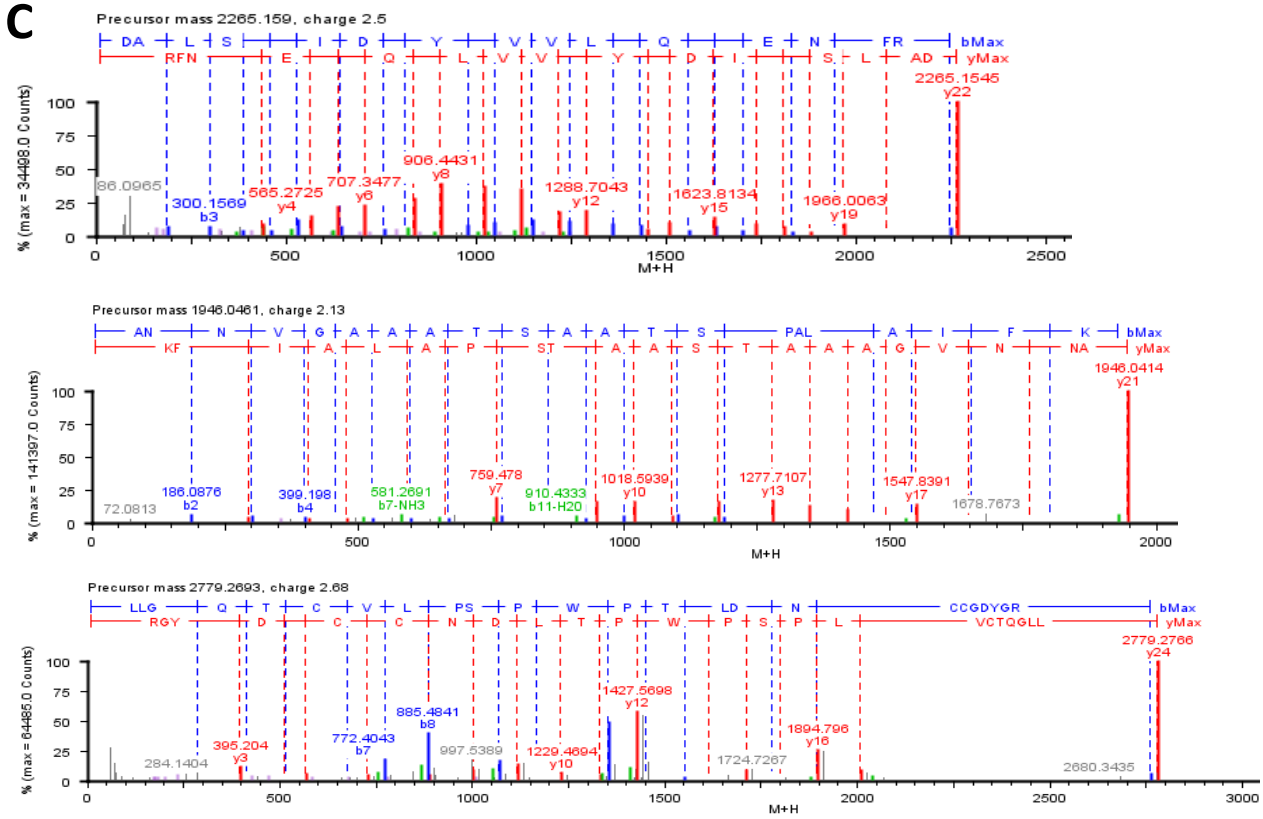


Fig. S8. Protein identification of Smsp-3 using tandem mass spectrometry. (Refer to additional materials and methods in this supplementary data.) (A) Summary of results of the database search with the tandem mass data of the tryptic-digested peptides of Smsp-3 from the *S. marmorata* silk glands. The top hit protein is *S. marmorata* L-fibroin and its ProteinLynx Global Server (PLGS) score (= 33437) is highly significant, indicating that Smsp-3 is identified as *S. marmorata* L-fibroin deduced from the cloned cDNA. (B) The coverage map of *S. marmorata* L-fibroin (Smsp-3). The regions of the protein sequence that match peptides are highlighted in color: blue, matched to a peptide; green, matched to a modified peptide (carbamidomethyl Cys by treatment with iodoacetamide). (C) Some examples of the MS/MS data matched to the trypsin-digested peptides from Smsp-3. Fragment ions of b-series and y-series are colored in blue and red, respectively.

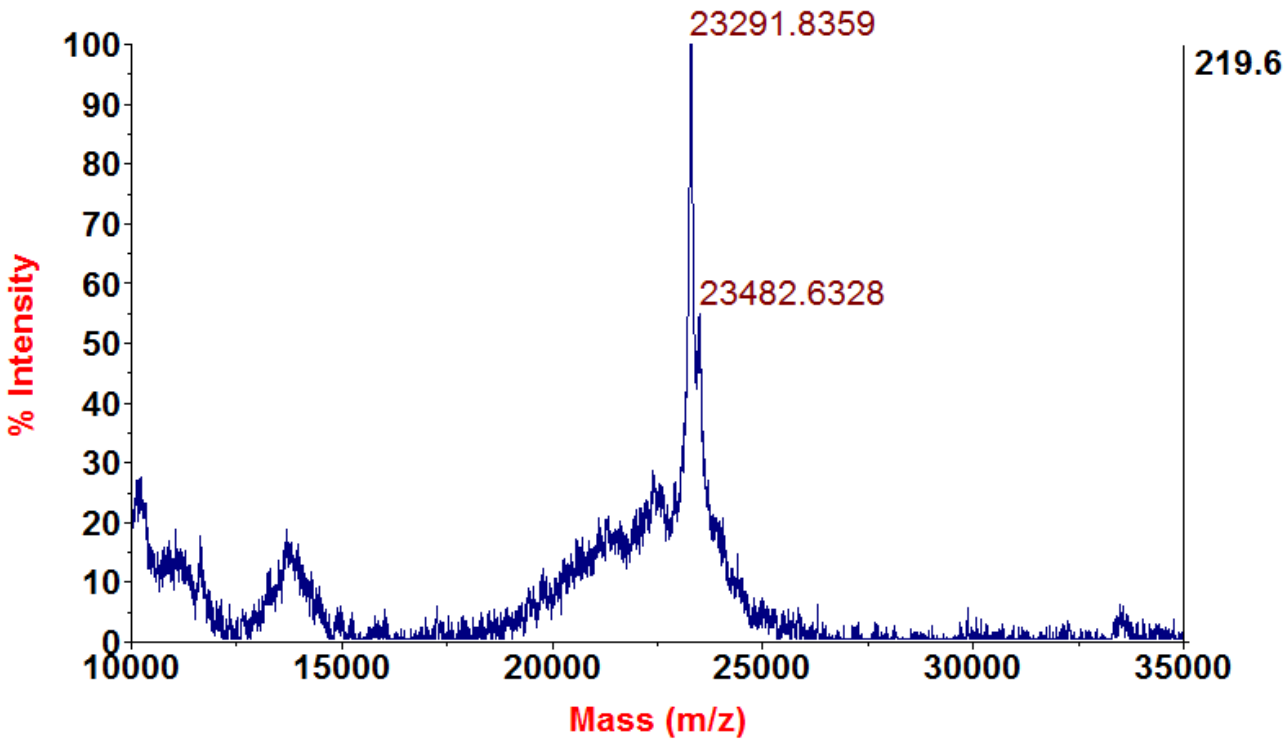


Fig. S9. The MALDI–TOF mass spectrum of the Smsp-3 protein extracted from the *S. marmorata* silk glands.

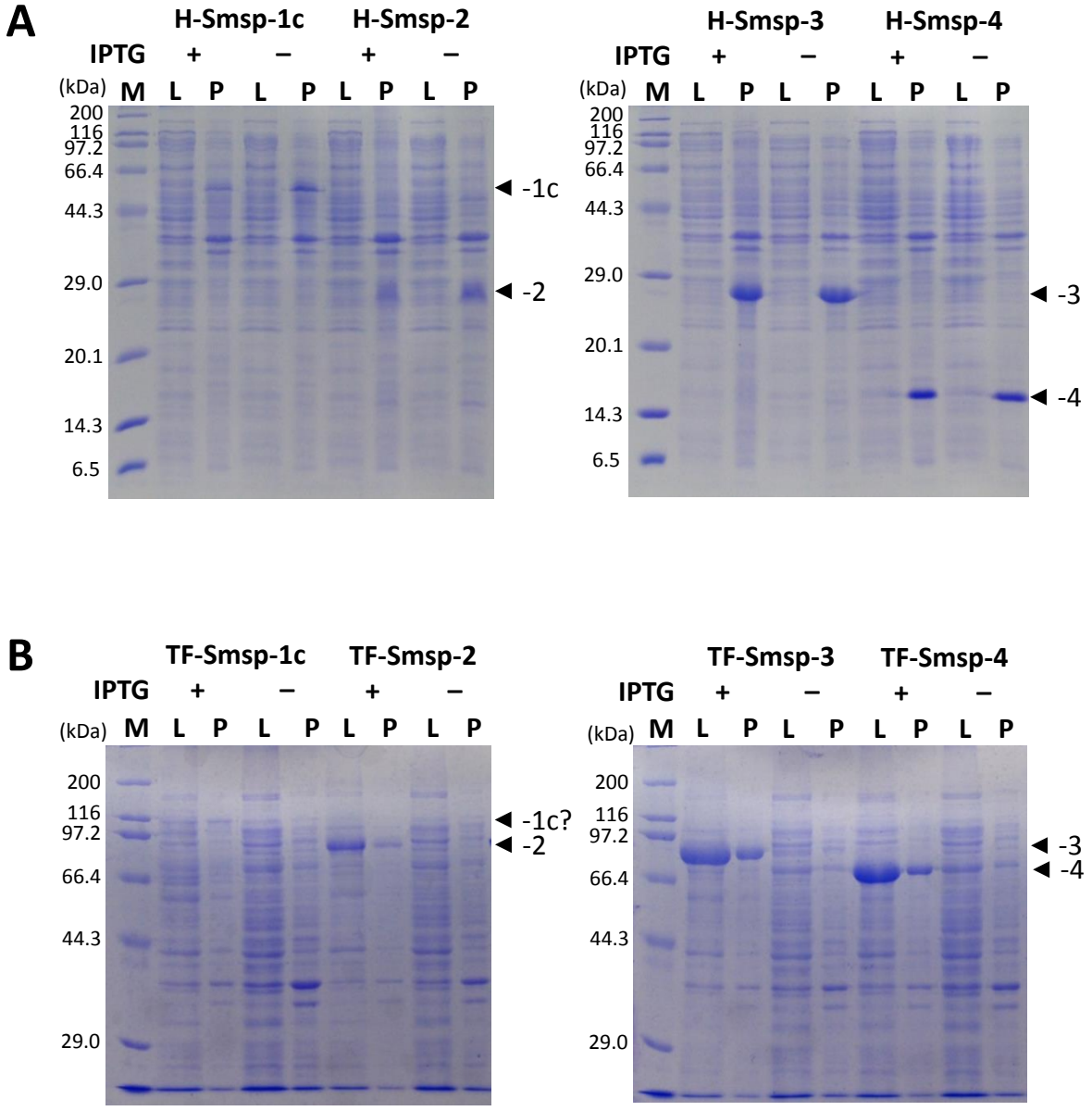


Fig. S10. Recombinant protein expression of Smsps in *E. coli*. (A) SDS–PAGE (15% gel) of His_6-tagged Smsps (H-Smsps). (B) SDS–PAGE (10% gel) of trigger factor-tagged Smsps (TF-Smsps). (+): Protein expression was induced with 50 μM IPTG; (–): IPTG was not added.
M: molecular mass marker (Protein Molecular Weight Marker (Broad), Takara Bio); L: cell lysate samples; P: pellet samples separated from cell extract by centrifugation. Proteins were stained with Coomassie brilliant blue. The calculated molecular masses of the expressed proteins including tags are as follows: H-Smsp-1c, 53.6 kDa; H-Smsp-2, 19.8 kDa; H-Smsp-3, 28.3 kDa; H-Smsp-4, 16.5 kDa; TF-Smsp-1c, 105.4 kDa; TF-Smsp-2, 71.5 kDa; TF-Smsp-3, 80.0 kDa; TF-Smsp-4, 68.3 kDa.

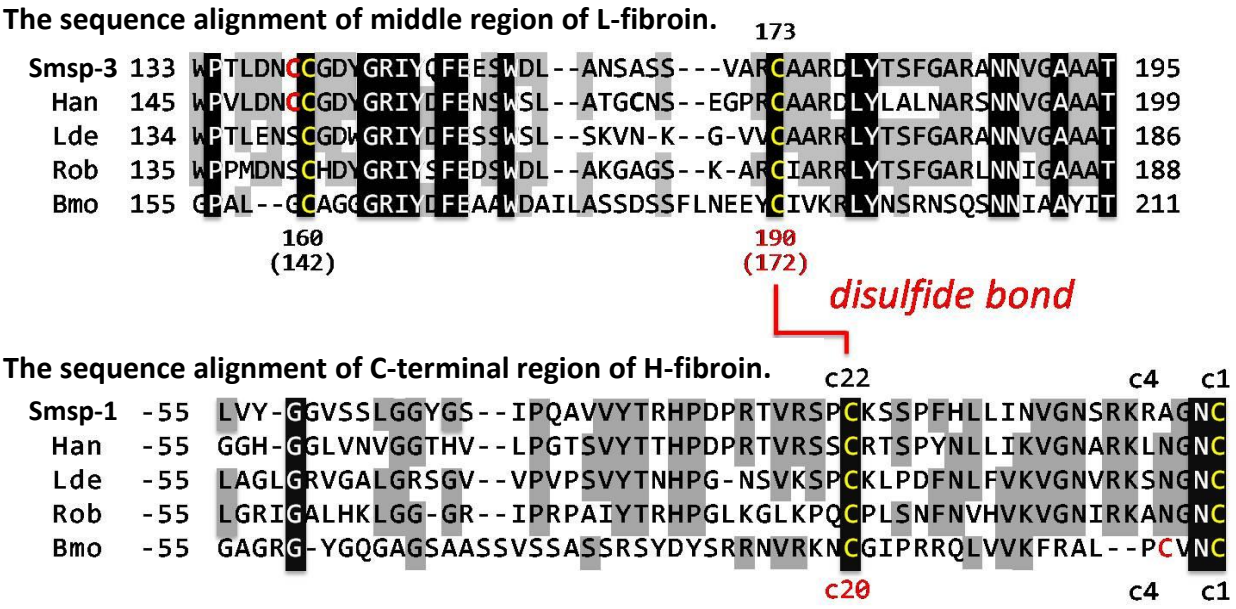


Fig. S11. The putative disulfide linkage between Smsp-1 and Smsp-3. The amino acid sequence alignments of L-fibroin (upper panel) and H-fibroin (lower panel) from silkworm and caddisflies. The disulfide-linked Cys residues of Cys-c20 (H-fibroin) and Cys-172 (L-fibroin) from *B. mori* [15] are strictly conserved among these caddisflies including *S. marmorata*, suggesting that Smsp-1 and Smsp-3 also potentially form a heterodimeric complex by a disulfide linkage between Cys-c22 (Smsp-1) and Cys-173 (Smsp-3). The amino acid residues completely conserved are highlighted in black. The amino acid residues conserved with 60–80% similarities are highlighted in gray. Han: *Hydropsyche angustipennis*; Lde: *Limnephilus decipiens*; Rob: *Rhyacophila obliterata*; Bmo: *Bombyx mori*. GenBank accession numbers of each fibroin light chain (L-fibroin): Han, BAF62094; Lde, BAF62096; Rob, BAH80180; Bmo, Q8IT50. GenBank accession numbers of each fibroin heavy chain (H-fibroin): Han, BAH80182; Lde, BAF62095; Rob, BAH80178; Bmo, P05790.

Additional Materials and Methods

Protein identification of Smsp-3 using tandem mass spectrometry

After SDS-PAGE of the P3′ fraction from the silk glands (Fig. S1), the gel band of Smsp-3 stained by the reverse staining method [7] was excised, crushed, and digested by trypsin in the gel, according to the modified protocol [20] based on the standard method [21], as follows:

- The gel band was excised and crushed, and the gel pieces were washed and destained with destaining solution (25 mM Tris–HCl, 192 mM glycine, pH 8.3) several times.
- Dehydration by soaking in 100 μL of acetonitrile and evaporation of the solvent were performed.
- The gel pieces were treated with 50 μL of 10 mM DTT in 20 mM NH_4HCO_3 buffer at 60 ℃ for 60 min.
- They were washed with 100 μL of 20 mM NH_4HCO_3 buffer.
- They were treated with 50 μL of 55 mM iodoacetamide in 100 mM NH_4HCO_3 buffer at room temperature for 30 min in the dark.
- They were washed with 100 μL of 20 mM NH_4HCO_3 buffer, and dehydrated in 100 μL of acetonitrile.
- They were washed again with 100 μL of 20 mM NH_4HCO_3 buffer, dehydrated again in 100 μL of acetonitrile, and dried up by evaporation.
- Then the gel pieces were treated with 25 μL of 10 ng/μL trypsin (sequencing grade modified trypsin, Promega) in 20 mM NH_4HCO_3 at 37 ℃ for 16 h.
- The peptides were extracted twice with 50 μL of 50% acetonitrile, 5% formic acid and 45% 20 mM NH_4HCO_3 buffer and once with 50 μL of 80% acetonitrile, 5% formic acid and 15% 20 mM NH_4HCO_3 buffer.
- The extracts were concentrated to about 10–20 μL by evaporation.

The extracted peptides were analyzed by a nano-LC Q-TOF mass spectrometer, ACQUITY UPLC Xevo QTOF (Waters). To identify the protein, the MS^E data were analyzed using ProteinLynx Global Server (PLGS) (Waters) with the *S. marmorata* protein sequence database deduced from the cDNA library sequences of the *S. marmorata* silk glands.

Additional References

[19] D.R. Cavener, Comparison of the consensus sequence flanking translational start sites in *Drosophila* and vertebrates, Nucleic Acids Res., 15 (1987) 1353-1361.

[20] R. Arai, M. Nishimoto, M. Toyama, T. Terada, S. Kuramitsu, M. Shirouzu, S. Yokoyama, Conserved protein TTHA1554 from *Thermus thermophilus* HB8 binds to glutamine synthetase and cystathionine beta-lyase, Biochim. Biophys. Acta, 1750 (2005) 40-47.

[21] A. Shevchenko, M. Wilm, O. Vorm, M. Mann, Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels, Anal. Chem., 68 (1996) 850-858.