



## Structure and expression of the silk adhesive protein Ser2 in *Bombyx mori*

Barbara Kludkiewicz<sup>a,b</sup>, Yoko Takasu<sup>c</sup>, Robert Fedic<sup>a</sup>, Toshiki Tamura<sup>c</sup>,  
Frantisek Sehnal<sup>a</sup>, Michal Zurovec<sup>a,\*</sup>

<sup>a</sup>Biology Centre, Academy of Sciences, and the Faculty of Natural Sciences, University of South Bohemia, Branisovska 31, 370 05 Ceske Budejovice, Czech Republic

<sup>b</sup>Department of Protein Biosynthesis, Institute of Biochemistry and Biophysics 5A, Pawlowskiego Str, Warsaw, Poland

<sup>c</sup>National Institute of Agrobiological Sciences, Tsukuba, Ibaraki, Japan

### ARTICLE INFO

#### Article history:

Received 22 October 2009

Received in revised form

27 November 2009

Accepted 30 November 2009

#### Keywords:

Silkworm

Sericin

Cocoon

Silk gland

Repetitive sequence

### ABSTRACT

Sericins are soluble silk components encoded in *Bombyx mori* by three genes, of which *Ser1* and *Ser3* have been characterized. The *Ser1* and *Ser3* proteins were shown to appear later in the last larval instar as the major sericins of cocoon silk. These proteins are, however, virtually absent in the highly adhesive silk spun prior to cocoon spinning, when the larvae construct a loose scaffold for cocoon attachment. We show here that the silk-gland lumen of the feeding last instar larvae contains two abundant adhesive proteins of 230 kDa and 120 kDa that were identified as products of the *Ser2* gene. We also describe the sequence, exon–intron structure, alternative splicing and deduced translation products of this gene in the Daizo p50 strain of *B. mori*. Two mRNAs of 5.7 and 3.1 kb are generated by alternative splicing of the largest exon. The predicted mature proteins contain 1740 and 882 amino acid residues. The repetitive amino acid sequence encoded by exons 9a and 9b is apparently responsible for the adhesiveness of *Ser2* products. It has a similar periodic arrangement of motifs containing lysine and proline as a highly adhesive protein of the mussel *Mytilus edulis*.

© 2009 Elsevier Ltd. All rights reserved.

### 1. Introduction

The silk produced by caterpillars such as the silkworm, *Bombyx mori*, is a composite mix of several proteins that are secreted by a pair of tubular glands (Fedic et al., 2003). Heavy chain fibroin (H-fibroin) associated with the light chain fibroin and the P25 glycoprotein are produced and assembled into micelles in the posterior silk gland (PSG) section, forming a gel flowing into the middle silk gland (MSG) section, where several layers of sericins are subsequently added. Whereas a significant amount of the partitioned jelly dope accumulates in the MSG lumen during feeding, even more collects during the post-feeding wandering period when the caterpillar seeks a suitable place for cocoon attachment. The silk dope is converted to solid fiber during spinning, when it flows through the narrow anterior silk gland section, and is discharged from the spinneret. Proteins from the PSG solidify into a backbone silk filament. A pair of filaments from each gland is then sealed into a single fiber by sericins, which also glue the fibers to one another. The silk spun at the beginning and in later phases of cocoon construction differs by sericins derived from MSG (Couble et al., 1987).

Two bends delineate the distal (adjacent to PSG), central and anterior sections of MSG, relative to the head. Each section produces different sericins that are layered consecutively around the fibroin core. Abundant sericins include sericin P (150 kDa), sericin M (400 kDa) and sericin A (250 kDa) identified in the distal, central and anterior MSG sections, respectively (Takasu et al., 2002). At least three genes are responsible for sericin production. *Ser1* was discovered earlier by Okamoto et al. (1982) and almost fully sequenced by Garel et al. (1997). It was determined that the gene consists of 9 exons and generates 4 mRNAs (2.8, 4.0, 9.0 and 10.5 kb) by alternative splicing (Garel et al., 1997). The *Ser2* gene was also discovered earlier and mapped with restriction enzymes (Michaille et al., 1990a). Its transcription start, part of the final exon, and the major repetitive motif were identified. Two *Ser2* mRNA variants, 3.1 kb and 5.0–6.4 kb were proposed to arise via alternative splicing mechanism (Michaille et al., 1990b). The structure and sequence of the third gene, *Ser3*, were described and shown to generate a single transcript of 4.5 kb (Takasu et al., 2007).

Relationships between the identified sericin genes and the sericin proteins are not fully understood. Sericin M and sericin P were identified as products of the *Ser1* gene and sericin A, which occurs mainly in the floss and outer layer of the cocoon, is most likely a product of *Ser3* (Takasu et al., 2007). Insufficient characterization of the *Ser2* gene hampered unequivocal assignment of a sericin protein(s) to this gene. The developmental profile of *Ser2*

\* Corresponding author. Tel.: +420 38775283; fax: +420 385310354.  
E-mail address: [zurovec@entu.cas.cz](mailto:zurovec@entu.cas.cz) (M. Zurovec).

gene expression suggested that it peaked before cocoon spinning (Michaille et al., 1990b). It is possible that the products of *Ser2* provide the sticky coating of fibers that function as a scaffold firmly attaching the cocoon to a substrate, such as a tree branch. To verify this hypothesis, we analyzed sericin proteins accumulated in the lumen of proximal MSG at the end of the feeding period and matched this information with a parallel analysis of the *Ser2* gene. Since the *Ser2* gene was reported to be remarkably polymorphic (Michaille et al., 1990a), we decided to use the inbred silkworm strain Daizo p50 in an effort to have an isogenetic background.

## 2. Materials and methods

### 2.1. Insect material

The silkworm strain Daizo p50, which was also used in the *B. mori* genome-sequencing project (Suetsugu et al., 2007), was reared on mulberry leaves in the standard way (Takasu et al., 2002). The MSG and PSG were dissected from the last instar, larvae anaesthetized by submergence in water and used as described in the following text. Our work also employed genomic clones 2001, 2002 and 2004, which were derived from a hybrid between the European *B. mori* strains 200 and 300 (Michaille et al., 1990a). The clones were kindly provided by Dr. A. Garel.

### 2.2. Analysis of the genomic DNA

The sequencing of PCR products obtained with specific primers was the basis of *Ser2* gene analysis in the provided genomic clones as well as the genomic DNA extracted from the Daizo p50 strain. To obtain genomic DNA, about 1 g PSG was crushed in a mortar under liquid nitrogen and homogenized in 10 ml lysis buffer (0.1 M NaCl, 0.05 M EDTA, 0.5% SDS and 0.01 M Tris, pH 7.5). The sample was treated with RNase A and proteinase K for 30 min at 37 °C. The genomic DNA was extracted with phenol/chloroform and precipitated with ethanol. Desired fragments were typically amplified from 40 ng DNA in 25 µl reaction volumes. Denaturation at 94 °C for 1 min was followed by 35 cycles, each consisting of 30 s at 94 °C, 25 s at 51 °C, and 90 s at 72 °C, and a final extension at 72 °C for 10 min. The primers were first derived from the short sequence tags published earlier by Michaille et al. (1990a) and then from other gene regions identified in our study. Genomic DNA of Daizo p50 was also used for Southern analysis. Individual 5 µg DNA samples were digested with the restriction enzymes *EcoRI*, *KpnI*, *HincII* or *XbaI*, and the resulting fragments were separated on a 1% agarose gel, which was blotted onto the Hybond-N+ membrane (Amersham Pharmacia Biotech). Hybridization probes covering specific gene regions were prepared by random priming of gel-purified cDNA inserts with [ $\alpha$ -<sup>32</sup>P]dATP in the reaction (Multiprime DNA Labeling System, Amersham Pharmacia Biotech). Southern blot hybridizations were conducted at 65 °C.

### 2.3. RNA isolation and analysis

Total RNA was extracted from 1 g of homogenized MSG of the Daizo p50 strain with Trizol reagent (Invitrogen) followed by ethanol precipitation. Aliquots of 0.5–1 µg RNA in 20 µl reaction mixtures containing 200 U SuperScript II, 1× First-strand buffer, 10 mM DTT, 500 µM dNTP, 40 U RNaseOUT ribonuclease inhibitor (all from Invitrogen), and 20 pmol oligo d(T) were used for reverse transcription (RT-PCR). The reaction was carried out at 42 °C for 50 min followed by heat inactivation at 70 °C for 15 min. The resulting cDNA was diluted 10-fold and 50 ng aliquots were taken as templates for PCR amplification with selected specific primers (Supplementary Table 1). A typical PCR profile consisted of 1 min

initial denaturation at 94 °C, 30 cycles of 30 s at 94 °C, 25 s at 53 °C, and 90 s at 72 °C, followed by 10 min at 72 °C for a final extension.

For the Northern analysis, 5 µg aliquots of total RNA were resolved by formaldehyde-agarose gel electrophoresis and transferred to Hybond-N+ nylon membranes (Amersham Pharmacia Biotech). The membranes were stained with 1% methylene blue to verify the integrity of rRNA bands. The probes and Northern blot hybridizations were performed as with the Southern blots previously described.

### 2.4. Nucleotide sequencing and computational analysis

PCR products were separated on an agarose gel, extracted with QIA Quick Gel Extraction Kit (Qiagen) and either used for direct sequencing or cloned into the pGEM T-easy vector (Promega). The inserts were amplified for sequencing with either gene specific primers or those complementary to the vector polylinker. Sequences were obtained with the ABI prism sequencer (Perkin Elmer model 310) and analyzed with the MEGALIGN program (DNASTAR). Homologous sequences were searched for in cDNA libraries derived from silk glands and in the *B. mori* genome Silkbase (<http://kaikoblast.dna.affrc.go.jp>). Deduced proteins were analyzed with several web-based programs, including SignalP 3.0 (Bendtsen et al., 2004) for signal sequence detection, Dotlet 1.5 (Junier and Pagni, 2000) for the detection of repeats and NetNGlyc (<http://www.cbs.dtu.dk/services/NetNGlyc/>) for prediction of glycosylation sites (R. Gupta, E. Jung, and S. Brunak, unpublished data).

### 2.5. Extraction, separation and N-terminal sequencing of the silk gland proteins

Silk glands were taken from individual silkworm larvae of the Daizo p50 strain on the fourth day of the final instar. The anterior MSG sections (18 mg) were excised, submerged in 300 µl water and the contents of their lumen were allowed to flow out for 30 min. The tissue was removed, the solution centrifuged (5 min, 12 000g), and the supernatant (280 µl) was mixed with 560 µl sample buffer (10% glycerol, 2.5% SDS, 62.5 mM Tris–HCl, 5% 2-mercaptoethanol, pH 6.8) and boiled for 5 min. Aliquots of 2 µl were loaded in Tris–glycine Laemmli buffer on a 6% polyacrylamide gel without a stacking gel and electrophoresed at 20 mA for 90 min. Proteins in the gel were transferred to a PVDF membrane (Bio-Rad, Hercules, CA, USA) in a transfer buffer (pH 8.8; contained 48 mM Tris, 39 mM glycine, 0.0375% SDS, and 20% methanol) using the SemiPhor™ Semi-dry transfer unit (Amersham Pharmacia Biotech, Piscataway, NJ, USA) with a current density of 0.8 mA/cm<sup>2</sup> for 1 h. The membrane was stained with Coomassie brilliant blue R-250, and destained in 50% methanol. The protein bands were excised from the membrane for N-terminal sequencing (ABI procise 491HT, Applied Biosystems, Foster City, CA, USA) at the Nippi Research Institute of Biomatrix (Tokyo, Japan).

### 2.6. Pull-off test of adhesion

The adhesion properties of the collected sericin proteins were compared to the properties of bone glue (Kittfort) and starch paste (Malbenka) by quantification with a weight holder attached by wire to a cylindrical wooden stub. The smooth surface of the stub cross-section (1 cm<sup>2</sup>) was coated with the test material and then pressed against a wooden plate. The assembly was allowed to dry at room temperature for 24 h. The plate was then held in a horizontal position with the stub facing down and exposed to an increasing load (loading rate of 1 kg/min) until it detached from the plate. The vertical force required to pull off the stub was taken as a measure of the tensile strength and expressed in Newtons per square centimeter (N/cm<sup>2</sup>).

### 3. Results

#### 3.1. Molecular cloning of two *Ser2* cDNAs

Michaille et al. (1990a,b) discovered the *Ser2* gene in the European silkworm strains 200 and 300. They detected two transcript variants of 3.1 and 5.0–6.4 kb, and sequenced three short areas of the gene. The areas included a 140 bp region surrounding the transcription start, 200 bp from the 3' end of the gene (including a polyadenylation signal), and a 315 bp internal sequence containing a repetitive motif of 45 bp and *MboI* restriction site. We used these sequences for the design of primers to start a comprehensive gene analysis.

Since there was remarkable polymorphism of the *Ser2* locus in the previously analyzed European strains, we chose the highly inbred silkworm strain Daizo p50 for our investigations. MSG-specific RNA from the fully-grown larvae of this strain was taken for RT-PCR. Most of the 5' region of the *Ser2* cDNA was amplified with the primer F1F (derived from the area surrounding the transcription start) and F1R (designed from the 45 bp repeats) (Supplementary Table 1). A single 1.8 kb fragment was obtained, cloned and sequenced. The sequence contained a long ORF encoding a protein rich in lysine and serine. The 3' region of the cDNA was generated with the forward primer F2F derived from the repetitive region and the reverse primer F2R matching the 3' end sequence (Supplementary Table 1). The amplified fragment of 800 bp also contained a long ORF. To connect both cDNA fragments, primers F3F and F3R were designed from the sequence adjacent to the region with the 45 bp repeats. A single 300 bp product was amplified and sequenced. The assembly of all three cDNA pieces yielded a composite sequence of 2924 bp that represented a full-length cDNA corresponding to the shorter mRNA splice variant (Fig. 1a). The cDNA contained a duplicated fragment of 389 bp that flanked both sides of a 309 bp region containing the 45 bp repeats.

To determine whether the short *Ser2* transcript was produced by skipping an exon included in the large transcript, several internal PCR primers based on the short cDNA were used in search of the missing sequence. PCR performed with the primer pair F4F-F1R (Supplementary Table 1) generated a product that contained a previously unidentified sequence (Fig. 1b). Subsequently, a combination of primer F6F, which was derived from this sequence, with the F3R primer

produced a large 2.5 kb fragment containing a long array of the 45 bp repeats. The result suggested that at least one more exon composed of the 45 bp repeats was localized upstream of the duplication. The 3' end of the newly discovered region was specified by amplification of a 600 bp fragment with the primers F2F and F5R. Sequencing of this fragment revealed an insertion of 389 bp between two regions containing the 45 bp repeats. Insertion of the newly amplified overlapping fragments into the sequence of the short *Ser2* mRNA yielded a 5.5 kb sequence of the long mRNA variant.

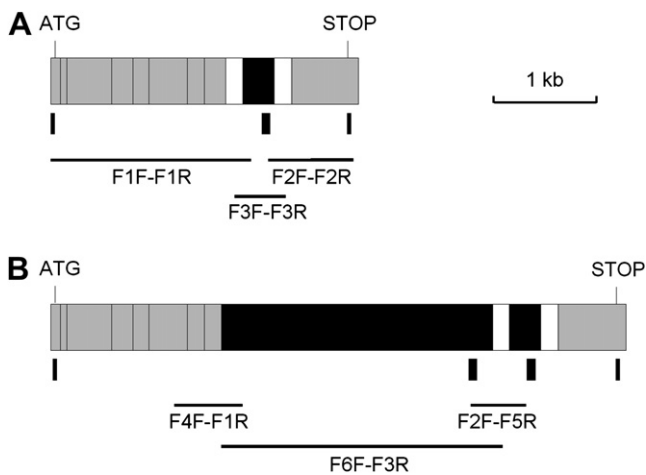
#### 3.2. Isolation of the *Ser2* gene

The identified cDNA sequences were used to design primers for the amplification of selected *Ser2* gene regions from genomic DNA. Combinations of several primers amplified an array of partially overlapping genomic sequences that ranged in size from 600 to 3000 bp. Step-wise sequencing of the long F6F-F3F product revealed a continuous reading frame composed largely of the 45 bp repeats. Comparison of the genomic and the cDNA sequences yielded the exon–intron gene structure (Fig. 2). When assembled, the established genomic sequences covered the entire *Ser2* gene with the exception of the promoter. We chose to identify this region first in the genomic clone 2004 derived from the European silkworm strains 200 and 300 (Michaille et al., 1990a). The sequence extending 400 bp upstream of the transcription start allowed us to design the specific primer G0F (Supplementary Table 1). This primer in combination with the reverse primer G0R, derived from the first exon, led to PCR amplification of the *Ser2* promoter in the Daizo p50 genomic DNA. The entire sequence of the *Ser2* gene of the Daizo p50 strain spanned over 12 kb and has been deposited in the DDBJ/EMBL/GenBank databases as *B. mori Ser2* allele D (Accession number GQ381286).

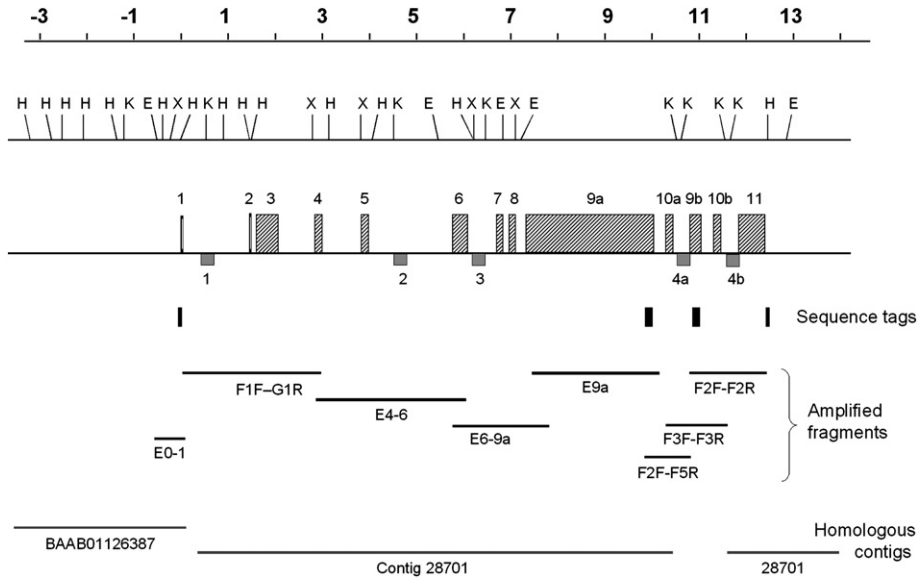
Regions of sequence similarity with the established *Ser2* sequence were detected in several BAC clones used in the *Bombyx* genome-sequencing project. The most complete *Ser2* sequence was identified in BAAB01186731 and BAAB01064864 clones and was localized to chromosome 11 (<http://kaikoblast.dna.affrc.go.jp/>). The contig # 28701, which was assembled from the sequenced parts of the BAC clones, covered the entire central part of the *Ser2* gene and confirmed most of the gene structure found in our study. However, the contig lacked the duplicated region (about 1.2 kb) present at the 3' end of our gene sequence. This discrepancy was apparently due to an error in the contig assembly from the sequenced BAC regions. Such errors occur relatively frequently because the software fails to recognize similar adjacent sequences as different and misaligns them as a single sequence (Eichler, 2001). By integrating our information with the database sequences, we were able to extend our map at the 5' and 3' ends by using the BAAB01126387 clone and contig # 28701, respectively (Fig. 2). The position of the *XbaI* site at the 3' extended sequence of contig # 28701 correlated with the Southern blot probed by exon 11 (data not shown).

#### 3.3. Structural analysis of the *Ser2* gene

The identified *Ser2* gene (Fig. 2) contained 13 exons that ranged in size from 28 to 2574 bp. The duplication of 1.2 kb at the 3' end included an approximately 300 nt fragment of exon 9 (truncated at the 5' end), copies of intron 9, exon 10 and intron 10, as well as a short duplication of the beginning of exon 11. The overall exon arrangement in the *Ser2* gene D allele could be described as 1, 2, 3, 4, 5, 6, 7, 8, 9a, 10a, 9b, 10b, and 11. The promoter included a consensus TATA box sequence in position –23 to –28; the region further upstream contained multiple copies of the homeodomain protein-binding consensus motif TAAT (Kissinger et al., 1990). The transcription start ACACGAG determined earlier by Michaille et al. (1990a) corresponded to the beginning of exon 1, which encoded the putative translation start site and the signal peptide



**Fig. 1.** Schematic drawing of two cDNAs derived from the *Ser2* gene by alternative splicing. The vertical lines indicate exon boundaries. Exons 9a and 9b are highlighted in black, exons 10a and 10b are shown in white. Areas sequenced earlier (Michaille et al., 1990a) are indicated by black rectangles under the map. PCR products that were crucial for cDNA analysis are shown as bars designated with primer names. A. Short transcript with single exon highlighted in black and called 9b. B. Long transcript with two exons highlighted in black (9a and 9b).



**Fig. 2.** Composite restriction map and the amplification strategy of the *Ser2* gene in the silkworm strain Daizo p50. Areas sequenced earlier<sup>6</sup> are indicated by black rectangles (“sequenced tags”). Exons are represented by diagonally striped boxes; grey boxes below the line indicate putative transposons (numbered Ins1–Ins4B). Fragments amplified by PCR and sequenced are shown as black lines below the map. The grey lines at the bottom demarcate sequences detected in the BAC clone BAAB01126387 or in the contig # 28701 of the Kaikoblast database (<http://kaikoblast.dna.affrc.go.jp>). Abbreviations for restriction sites are: H (*HincII*), K (*KpnI*), E (*EcoRI*) and X (*XbaI*).

region ending in exon 2. The predicted initiation codon occurred in the context of CATCATG, in accordance with the Kozak consensus initiation sequence CACCATG (Cavener, 1987). The last exon (#11) consisted of 578 bp and contained a stop codon (TGA) and 89 nt 3' UTR, including the polyadenylation signal AATAAA. Taken as a whole, the coding region contained a high number of adenosine residues (42.4%) and a relatively low representation of thymidine (14.3%).

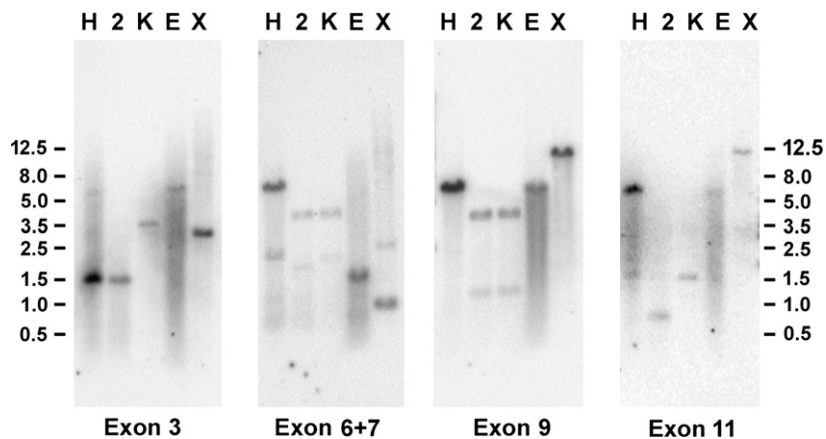
The sizes of introns ranged from 85 to 1702 bp. The sequences at the exon/intron boundaries were highly conserved with respect to the canonical acceptor/donor splicing site (ag/gt). Introns 1, 5, 6, 10a and 10b contained Gypsy-Ty3-like elements (data not shown). The Ty3-like elements present in introns 10a and 10b were almost identical. The sequences of introns 9a vs. 9b were completely identical and introns 10a vs. 10b displayed 97% identity, suggesting that the segmental gene duplication event was relatively recent.

To confirm the deduced gene organization, Southern blot analysis was performed with probes derived from four different parts of the gene (Fig. 3). The genomic DNA was digested with *EcoRI*, *HincII*, *KpnI*,

*HincII* + *KpnI*, and *XbaI*, respectively, and hybridized to probes corresponding to exons 3, 6 + 7, 9a and 11, respectively. The length of the hybridizing restriction fragments was compared with the expected sizes calculated from the sequencing data. No significant size differences from our predictions were detected (data not shown). Southern blotting provided further evidence for the partial duplication of exon 9a, intron 9a, exon 10a and intron 10a. As shown in Fig. 3, hybridization of the Southern blot membrane with a probe corresponding to exon 9a revealed two hybridizing fragments of the predicted sizes in the DNA preparations restricted with *KpnI* or with *HincII* + *KpnI*. One fragment consisted of about 4200 bp and the other of 1200 bp. If there were no duplication, hybridization would occur with a single DNA fragment corresponding to a single exon 9.

3.4. Alternative splicing of the *Ser2* gene

It was shown earlier that the *Ser2* gene is expressed exclusively in MSG and generates two mRNA variants (Michaille et al., 1990a,b).



**Fig. 3.** Southern analysis of the *Ser2* gene in the silkworm strain Daizo p50. Aliquots of 5 µg genomic DNA were digested with *HincII* (H), *HincII* + *KpnI* (2), *KpnI* (K), *EcoRI* (E) and *XbaI* (X), respectively, and hybridized with radiolabeled PCR fragments derived from exon 3, exons 6 + 7, exon 9 (repetitive region present in both 9a and 9b) and exon 11. Size markers (kb) are indicated on both sides.

Our analyses of the cDNAs (Fig. 1) and genomic DNA (Fig. 2) revealed that the mRNAs are products of alternative splicing of exon 9a. This conclusion was verified by Northern blot analysis of the RNA isolated from the MSG of the fully-grown *B. mori* larvae. The blots of total RNA were probed with radiolabeled fragments corresponding to individual *Ser2* exons. Probe 9A specific for exon 9a was derived from its unique 5' region, while probe 9B matched the 45 bp repeats present in exons 9a and 9b. As expected, probe 9A gave a signal only with the longer transcript (Fig. 4, lane F), confirming that the shorter mRNA lacked exon 9a. All other probes detected two mRNAs of 3.1 and 5.7 kb (including polyA chains). Probe 9B gave a stronger signal with the larger mRNA (Fig. 4, lane G), because this transcript included more 45 bp repeats than the smaller mRNA. The probe matching exon 10 (lane H) revealed an additional band above the two mRNAs, perhaps a result of cross-hybridization with an unknown RNA.

Alternative splicing of exon 9a could also be deduced from the alignment of the *Ser2* gene sequence with EST sequences retrieved from <http://kaikoblast.dna.affrc.go.jp/>. The BLAST search identified several independent ESTs that covered the duplicated region, spanning either the exons 9a-10a-9b or 10a-9b-10b, consistently with the proposed duplication arrangement 9a-10a-9b-10b-11. Exon 9a was detected in the EST's *MSV3 23F01*, *MSV3 13D10*, *Rswab0 00812* and *ET MSV3 03D09*, while it was missing in the ESTs *msgV0234*, *MSV3 18A12F*, *Rswab0 006125.yl*, and *Rswab0 006125.yl*. This observation is consistent with the existence of two cDNA classes detected in the present study (Fig. 1). We conclude that there are two major *Ser2* mRNAs that differ by the presence or absence of the largest (2.6 kb) exon 9a.

### 3.5. Deduced amino acid sequences

The open reading frame of the longer splice variant of *Ser2* encoded a protein of 198 kDa (1758 amino acids), whereas the predicted MW of the shorter variant was 102 kDa (900 amino acids). The proteins were characterized by a high representation of hydrophilic amino acids, especially those containing a charge at physiological pH. The larger protein contained 17.1% Lys, 15.1% Ser, 11.7% Asp, 11.1% Glu, 5.8% Thr and 5.6% Pro residues. The smaller product lacked the repetitive Lys- and Pro-rich region encoded by exon 9a, but the remaining sequence was identical with that of the larger protein. The presence of the repetitive region affected the isoelectric point of the proteins, which was calculated 8.5 and 5.4 for the large and small sericin2, respectively. Both deduced proteins contained several N-glycosylation sites predicted by the NetNGlyc 1.0 software.

The deduced amino acid sequence of *Ser2* began with a putative hydrophobic signal peptide MKIPYVLLFLVGVAVVNA encoded by the first and second exons (Fig. 5). The peptide encoded by the 3rd exon started with a short unique sequence and in the central part contained 7 reiterations of the KFENLDKDNVGE repeat. About the last

third of exon 3, the entire exons 4, 5, 6, 7, and 8, and the first 140 residues of exon 9a encoded a long unique sequence characterized by a high content of charged amino acids that occurred often as doublets (KK, EE). Most of exon 9a encoded 45 reiterations of the well-conserved repeat RSPSHKDTEKAKPND; the repetitive region was followed by a unique sequence of 43 residues. There was some sequence similarity between the short motifs encoded by exons 6 and 9a (Fig. 5). Translation products of both exons began with an EKK motif and their internal sequence (only the non-repetitive part of the exon 9a product) differed from the rest of the deduced protein by the presence of DD doublets and a relatively high representation of Q. At the C-terminus, both products contained a similar distribution of the oligopeptides ERE, KSES, EFEN, and KxAxS.

The translation product of exon 9b began with a sequence similar to that encoded by exon 11. When aligned, astounding 13 out of 22 amino acid residues were in identical positions. The remainder of the product of exon 9b appeared as a truncation of the peptide encoded by exon 9a, including an octapeptide degenerated from the 15-residue repeat, one perfect and one degenerated full-length repeat, and an exact copy of the 43-mer found at the end of the exon 9a product. The peptides encoded by exons 10a and 10b were identical (Fig. 5).

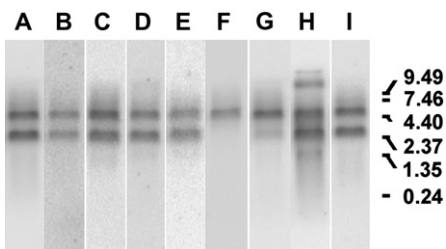
The proteins contained 6 potential N-glycosylation sites (Fig. 5). Structural predictions revealed that the shorter protein could form a coiled coil and alpha helix, while the region encoded by exon 9a could not. The regular distribution of proline and the prevalence of charged residues in the repeats also restrict the formation of strong  $\beta$ -sheets. The crosslinking of the *Ser2* protein probably relied on the electrostatic interactions between these residues.

### 3.6. Identification of the *Ser2* proteins

Since the *Ser2* mRNA is localized mostly in the anterior part of MSG and its expression occurs earlier than that of *Ser1* (Michaille et al., 1989), we analyzed proteins from the lumen of the anterior MSG section in the last instar larvae, about 2 days before the wandering stage (i.e. 3 days before the start of cocoon spinning). Proteins collected in distilled water were separated by SDS-PAGE in a 6% gel and detected by staining with Coomassie brilliant blue R-250 (Fig. 6). The sizes of two dominant silk proteins were estimated by comparison with standard bands at 230 and 120 kDa. The N-terminal sequencing of both proteins yielded the identical sequence LFGGLVKLSLS, which perfectly matched the peptide LFGGLVKLSLS identified in the N-terminus of both proteins deduced from the *Ser2* gene. We surmised that these bands correspond to the larger (deduced size ca. 197 kDa) and the smaller (ca. 101 kDa) products of the *Ser2* gene.

### 3.7. Adhesion properties of crude extract of the *Ser2* proteins

Analysis of proteins stored in the silk gland lumen showed that the *Ser2* sericins occurred in the anterior MSG section before the wandering period (Fig. 6). They were discharged early during spinning because of the predominance of the other sericins in the MSG content analyzed at the time of cocoon spinning (data not shown). The timing of expression of the *Ser2* proteins suggests that they provide the highly adhesive coating of the silk filaments spun prior to cocoon construction. Since we were unable to obtain sufficient amounts of pure 230 and 120 kDa sericins to individually test their adhesive properties, we used all material present in the lumen of anterior MSG dissected from the last instar larvae, 2 days before the wandering period. Commercial starch and bone glues were used for comparison in the measurements of tensile strength needed to detach wooden surfaces glued together over a 1 cm<sup>2</sup> area. The adherence of the crude *Ser2* proteins extract was 120 ± 30 N/cm<sup>2</sup>. For comparison, this was less than the adherence of commercial bone glue Kittfort (502 ± 132 N/cm<sup>2</sup>), but more than the adherence of the commercial starch glue Malbenka (42 ± 20 N/cm<sup>2</sup>).



**Fig. 4.** Northern analysis of MSG-specific RNA with probes derived from *Ser2* exons. Lane A, hybridization with probe matching exon 3; B, exons 4 + 5; C, exon 6; D, exon 7; E, exon 8; F, exon 9a (unique 5' end); G, exon 9b (repetitive region, common for both 9a and 9b); H, exon 10; I, exon 11. Size markers (kb) are indicated on the right.



exon10b and intron 10b. Exon 9b included a 40 bp sequence with 77% identity to the start of exon 11, followed by 243 bp matching the 3' region of exon 9a (Fig. 5). The duplication may be consequence of unequal crossing-over. The presence of several mobile elements of the Ty3/gypsy/mariner in *Ser2* gene type suggests that they may play a role in the duplication. Since there is no such

element in the intron 9a or 9b, the distribution of transposons does not support this hypothesis. Previous mapping of three *Ser2* alleles from European 200 and 300 *B. mori* strains revealed high level of polymorphism near the 3' end of the gene (Michaille et al., 1990a). The sequence of the European C allele suggests that the duplicated segment may be larger in some silkworm strains and may include

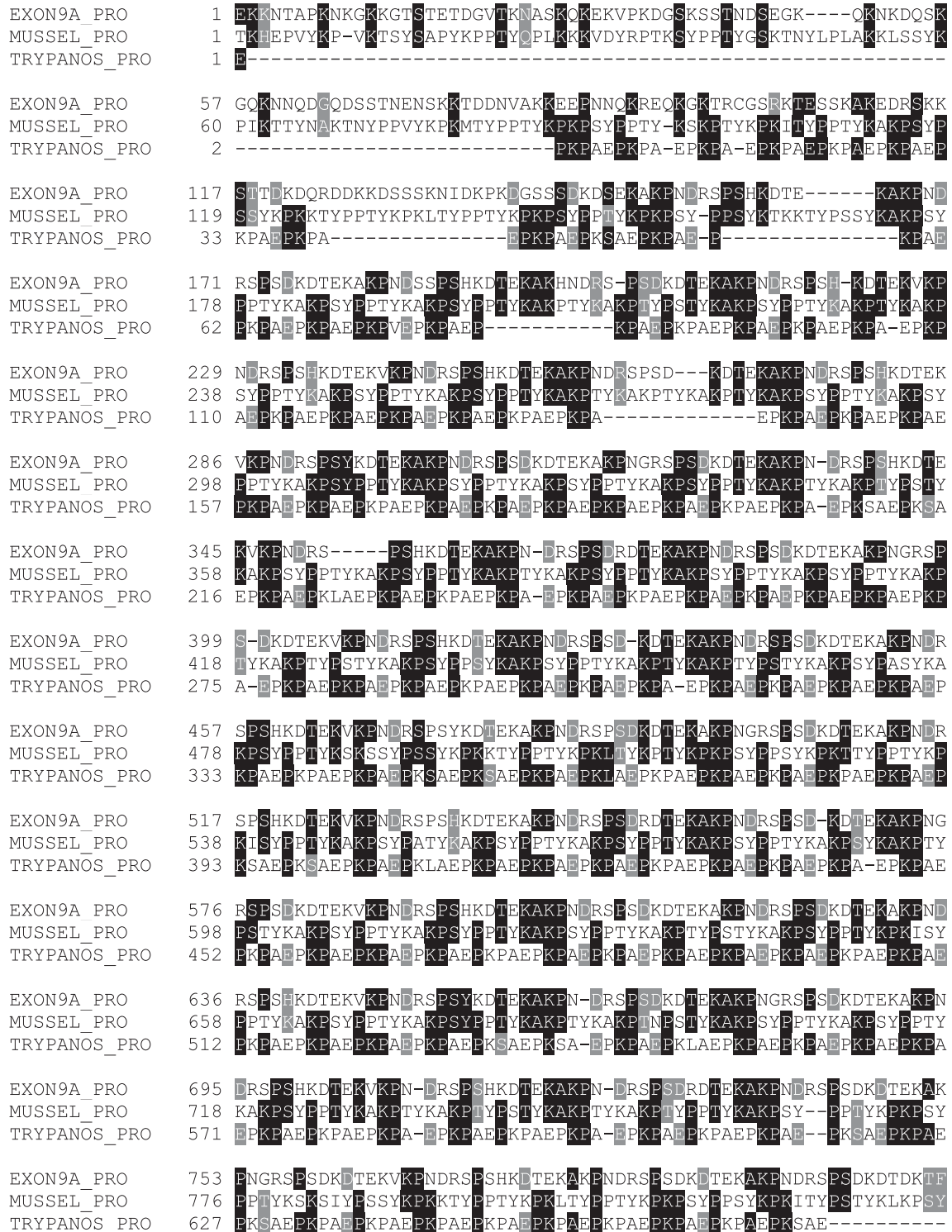


Fig. 7. Alignment of the *Ser2* deduced protein sequence encoded by exon 9a (starting with amino acid 534) with the Mussel Adhesive Plaque Protein (Q25460) and the *Trypanosoma cruzi* protein trans-sialidase (XM\_808493). In the alignment, first amino acid of the mussel protein corresponds to the first amino acid in the sequence Q25460 and in the trypanosomal protein XM\_808493 first amino acid corresponds to amino acid 821. Identical amino acids are highlighted in black and similar amino acids (based on the hydrophobicity score of each residue) are shaded in grey.

exons 8 and 11 (Kludkiewicz et al., unpublished). More information is needed to study the evolution of *Ser2* gene structure.

#### 4.2. Alternative splicing and deduced proteins

The two splicing variants of the *Ser2* transcripts we describe here (Fig. 1) were also detected earlier (Michaille et al., 1990a). We showed that they differ by the presence of large exon 9a. Two abundant proteins of 230 kDa and 120 kDa, which were present in the anterior MSG lumen of the feeding last instar larvae (Fig. 6), had the identical N-terminal peptide LFGGLVKLSL, matching a short stretch of the amino acid sequence deduced from the *Ser2* mRNAs (Fig. 5). The deduced protein began with MKIPYVLLFLVGVAVVNALPNLFGGLVKLSL, where signal peptide cleavage was predicted to occur after Ala<sub>18</sub>, i.e. four residues upstream from the detected N-terminus of the proteins. Their production required hydrolysis between Pro<sub>22</sub> and Leu<sub>23</sub>, but canonical signal peptidases usually do not cut after proline (Von Heijne, 1986). We therefore propose that a unique peptidase is involved in the proteolytic processing of *Ser2* proteins. This processing might be similar to that of the light chain fibroin in *Galleria mellonella*, where we implicated the N-terminal dipeptidyl dipeptidase IV (Zurovec et al., 1995) that often cuts after the proline residues (Kreil et al., 1980). The predicted sizes of the *Ser2* secreted proteins were 197 kDa and 101 kDa, respectively. The electrophoretic mobility of the proteins indicated somewhat larger sizes, probably due to sugar attachments to some of the 6 potential N-glycosylation sites. Its amino acid composition, such as the relatively high serine content, could also slow down its mobility (Huang et al., 2003).

The unique amino acid composition of the deduced sericin proteins invites comparison with previous reports about the primary structure of native cocoon silk proteins. A protein named Src2 by Gamo (1987) was regarded as a *Ser2* gene product because of its demonstrated size variations and tissue specificity in the anterior part of MSG (Michaille et al., 1990b). The total amino acid composition of the genuine *Ser2* proteins predicted from our sequence, however, does not support this idea. The most striking differences between the deduced amino acid composition of the larger *Ser2* protein and any sericin that has been isolated from cocoons included abundance of Lys (17.1%, i.e. at least 3 times more than in the analyzed cocoon sericins) and Pro (5.8%, in contrast to trace amounts detected in all sericins extracted from the cocoons) and paucity of Gly and Ser (4.1% and 15.1%, i.e. 3–4 times and 2 time less than in the analyzed sericins). These data are consistent with our conclusion that the adhesive *Ser2* proteins coat silk filaments spun prior to cocoon construction. Because the supply of *Ser2* is quickly exhausted, very little is left to become incorporated in the cocoon silk, explaining why *Ser2* cannot be isolated from this source.

#### 4.3. Specific function of the *Ser2* proteins

The transcription of *Ser2* mRNAs was high in the feeding last instar larvae and declined to zero in post-feeding larvae Michaille et al., 1989, 1990a,b. According to Couble et al. (1987) the gene was expressed in the whole MSG in young larvae, but later became restricted to the anterior MSG. In accord with this expression pattern, we were able to isolate both *Ser2* proteins from the lumen of the anterior MSG of the feeding larvae. The expression of *Ser1* and *Ser3* is low at this time. Since the production of both *Ser2* proteins occurs before cocoon spinning, these proteins seem to be needed for silk functions in larvae and are required to a lesser degree, if at all, for cocoon spinning. The larvae spin a small silky pad before each molt and a loose scaffold just before cocoon spinning. The pad provides a firm holding for the molting larva and the scaffold fixes the cocoon to a suitable substrate, for example a tree branch. The high adhesiveness of silk fibers used for these structures is due to the presence of the *Ser2* proteins.

The adhesiveness of *Ser2* protein may be connected to high content of charged amino acid residues, which could provide for electrostatic interactions with molecules in substrate surfaces. Interactions with residues of the opposite charge within the repetitive *Ser2* sequence are probably hindered by the high incidence of Pro, which stabilizes bends in the peptide chain. About half of the large *Ser2* protein and the whole small *Ser2* protein contain virtually no Pro, but the proteins have high occurrence of both positively and negatively charged residues.

We also propose that the stickiness of these proteins is due to the presence of a repetitive sequence encoded by exons 9a and 9b, exhibiting a remarkable similarity (35% identity over 600 amino acids; Fig. 7) with that of the adhesive protein that fixes the byssal threads of blue mussels to rocks (Filipula et al., 1990). The byssal threads have a fibrous collagenous core that is coated with the highly adhesive glue protein, largely made from reiterations of the decapeptide PPTY-KAKPSY. The high Tyr content was interpreted to facilitate a requirement for protein crosslinking under water (Burzio and Waite, 2000). The *Ser2* repeats are also similar to the repetitive motif of a trans-sialidase from the protozoan parasite *Trypanosoma cruzi*, which is believed to mediate adhesion to host cells (Chuenkova and Pereira, 1995). This trans-sialidase consists of a long array of the pentapeptide PKPAE. Except for the high content of Lys and Pro, this repeat shows no obvious similarity with the pentapeptidic repeat RSPSHKDKTEKAKPND present in the large *Ser2* protein (Fig. 7). However, the repeats in mussel glue, trans-sialidase and *Ser2* all comprise domains with a similar frequency of short KAK and PK motifs, single Lys and Pro residues, and large hydrophilic residues (Tyr, Asp, Glu). We believe that the distribution of these specific amino acid residues mediate the stickiness of these proteins. A low level of similarity to the *Ser2* repetitive region occurs also in the cellulosome anchoring protein of *Clostridium*. These various adhesive proteins are of very different origin and hence the similar amino acid patterns responsible for stickiness are the result of convergent evolution.

#### Acknowledgments

Genomic clones of the C allele of the *Ser2* genes were kindly provided by Dr. Annie Garel of the Claude Bernard University Lyon, France. We also acknowledge comments on the manuscript by Dr. Cheryl Hayashi of the University of California, Riverside. Conducted research was supported by grant IAA5007402 from the Grant Agency of the Academy of Sciences, MSM 60076605801 and by Research Center Program MSMT – LC06077. The nucleotide sequence for the *B. mori* sericin 2 gene has been deposited in the GenBank database under GenBank Accession Number (GQ381286).

#### Appendix. Supplementary material

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ibmb.2009.11.005.

#### References

- Bendtsen, J.D., Nielsen, H., von Heijne, G., Brunak, S., 2004. Improved prediction of signal peptides: signalP 3.0. *J. Mol. Biol.* 340, 783–795.
- Burzio, L.A., Waite, J.H., 2000. Cross-linking in adhesive quinoproteins: studies with model decapeptides. *Biochemistry* 39, 11147–11153.
- Cavener, D.R., 1987. Comparison of the consensus sequence flanking translational start sites in *Drosophila* and vertebrates. *Nucleic Acids Res.* 15, 1353–1361.
- Couple, P., Michaille, J.J., Couble, M.L., Prudhomme, J.C., 1987. Developmental switches of sericin mRNA splicing in individual cells of *Bombyx mori* silkgland. *Dev. Biol.* 124, 431–440.
- Chuenkova, M., Pereira, M.E., 1995. *Trypanosoma cruzi* trans-sialidase: enhancement of virulence in a murine model of Chagas' disease. *J. Exp. Med.* 181, 1693–1703.
- Eichler, E., 2001. Segmental duplications: what's missing, misassigned, and misassembled – and should we care? *Genome Res.* 11, 653–656.

- Fedic, R., Zurovec, M., Sehnal, F., 2003. Correlation between fibroin amino acid sequence and physical silk properties. *J. Biol. Chem.* 278, 35255–35264.
- Filipula, D.R., Lee, S.M., Link, R.P., Strausberg, S.L., Strausberg, R.L., 1990. Structural and functional repetition in a marine mussel adhesive protein. *Biotechnol. Prog.* 6, 171–177.
- Gamo, T., 1987. Component of silk proteins and their gene loci in the silkworm. *JARQ* 21, 53–58.
- Garel, A., Deleage, G., Prudhomme, J.C., 1997. Structure and organization of the *Bombyx mori* Sericin 1 gene and of the sericin 1 deduced from the sequence of the Ser 1B cDNA. *Insect Biochem. Mol. Biol.* 27, 469–477.
- Huang, J., Valluzzi, R., Bini, E., Vernaglia, B., Kaplan, D.L., 2003. Cloning, expression and assembly of sericin-like protein. *J. Biol. Chem.* 278, 46117–46123.
- Junier, T., Pagni, M., 2000. Dotlet: diagonal plots in a web browser. *Bioinformatics* 16, 178–179.
- Kissinger, C.R., Liu, B.S., Martin-Blanco, E., Kornberg, T.B., Pabo, C.O., 1990. Crystal structure of an engrailed homeodomain–DNA complex at 2.8 Å resolution: a framework for understanding homeodomain–DNA interactions. *Cell* 63, 579–590.
- Kreil, G., Haiml, L., Suchanek, G., 1980. Stepwise cleavage of the pro part of rome-littin by dipeptidase IV. *Eur. J. Biochem.* 111, 49–58.
- Michaille, J.J., Garel, A., Prudhomme, J.C., 1989. The expression of five middle silk gland specific genes is territorially regulated during the larval development of *Bombyx mori*. *Insect Biochem.* 19, 19–27.
- Michaille, J.J., Garel, A., Prudhomme, J.C., 1990a. Cloning and characterization of the highly polymorphic ser2 gene of *Bombyx mori*. *Gene* 86, 177–184.
- Michaille, J.J., Garel, A., Prudhomme, J.C., 1990b. Expression of Ser1 and Ser2 genes in the middle silk gland of *Bombyx mori* during the fifth instar. *Sericologia* 30, 49–60.
- Okamoto, H., Ishikawa, E., Suzuki, Y., 1982. Structural analysis of sericin genes. Homologies with fibroin genes in the 5' flanking nucleotide sequences. *J. Biol. Chem.* 257, 15192–15199.
- Suetsugu, Y., Minami, H., Shimomura, M., Sasanuma, S.I., Narukawa, J., Mita, K., Yamamoto, K., 2007. End-sequencing and characterization of silkworm (*Bombyx mori*) bacterial artificial chromosome libraries. *BMC Genomics* 8, 314.
- Takasu, Y., Yamada, H., Tsubouchi, K., 2002. Isolation of three main sericin components from the cocoon of the silkworm, *Bombyx mori*. *Biosci. Biotechnol. Biochem.* 66, 2715–2718.
- Takasu, Y., Yamada, H., Tamura, T., Sezutsu, H., Mita, K., Tsubouchi, K., 2007. Identification and characterization of a novel sericin gene expressed in the anterior middle silk gland of the silkworm *Bombyx mori*. *Insect Biochem. Mol. Biol.* 37, 1234–1240.
- Von Heijne, G., 1986. A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res.* 14, 4683–4690.
- Zurovec, M., Vaskova, M., Kodrik, D., Sehnal, F., Kumaran, A.K., 1995. Light-chain fibroin of *Galleria mellonella* L. *Mol. Gen. Genet.* 247, 1–6.