Insect Biochemistry and Molecular Biology 106 (2019) 28–38

Contents lists available at ScienceDirect

Insect Biochemistry and Molecular Biology

journal homepage: www.elsevier.com/locate/ibmb

The expansion of genes encoding soluble silk components in the greater wax moth, *Galleria mellonella*

Barbara Kludkiewicz[a], Lucie Kucerova[a], Tereza Konikova[a], Hynek Strnad[c], Miluse Hradilova[c], Anna Zaloudikova[a], Hana Sehadova[a], Peter Konik[b], Frantisek Sehnal[a,b,*], Michal Zurovec[a,b,**]

[a] *Biology Centre of the Czech Academy of Sciences, Institute of Entomology, Branisovska 31, 370 05, Ceske Budejovice, Czech Republic*
[b] *Faculty of Science, University of South Bohemia, Branisovska 31, 370 05, Ceske Budejovice, Czech Republic*
[c] *Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Videnska 1083, 142 20, Praha 4, Czech Republic*

ARTICLE INFO

Keywords:
Silk glands
Lepidoptera
Sericin
Mucin
Cocoon
B. mori

ABSTRACT

Lepidopteran silk is a complex assembly of proteins produced by a pair of highly specialized labial glands called silk glands. Silk composition has been examined only in a handful of species. Here we report on the analysis of silk gland-specific transcriptomes from three developmental stages of the greater wax moth, *Galleria mellonella,* combined with proteomics, Edman microsequencing and northern blot analysis. In addition to the genes known earlier, we identified twenty seven candidate cDNAs predicted to encode secretory proteins, which may represent novel silk components. Eight were verified by proteomic analysis or microsequencing, and several others were confirmed by similarity with known silk genes and their expression patterns. Our results revealed that most candidates encode abundant secreted proteins produced by middle silk glands including ten sericins, two seroins, one or more mucins, and several sequences without apparent similarity to known proteins. We did not detect any novel PSG-specific protein, confirming that there are only three fibroin subunits. Our data not only show that the number of sericin genes in the greater wax moth is higher than in other species thus far examined, but also the total content of soluble proteins in silk is twice as high in *G. mellonella* than in *B. mori or A. yamamai*. Our data will serve as a foundation for future identification and evolutionary analysis of silk proteins in the Lepidoptera.

1. Introduction

Silk is a proteinaceous polymer secreted by specialized exocrine glands in several groups of arthropods. Lepidopteran silk is generated in modified larval labial glands called "silk glands" (SG) and released through the spinneret as a continuous thread that hardens on contact with air. Silk components can be divided into four groups: fibroins that form two core filaments (one from each gland); sericin bioadhesives holding fibroin fibers together; seroins, with potential antimicrobial functions; and miscellaneous proteins (enzymes, enzyme inhibitors, etc.) with diverse functions. Lepidopteran silk genes are known for their tissue-specific expression; fibroins are produced in the posterior silk gland section (PSG) and sericins in the middle section (MSG). The sites of production of other proteins have not been precisely determined, but most of them seem to be derived from the MSG along with sericins. The assemblage of soluble proteins produced in the MSG is often referred to as "sericin," but sericins in *sensu stricto* are only specific fraction of the soluble silk proteins from the MSG. Available information on the respective genes suggests that they underwent fast evolutionary changes, including gene duplication and loss. The rapid divergence of most silk proteins makes it difficult to identify them on the basis of homology. Silk proteins are stored as a jelly dope in the MSG lumen, and new protein secretions are sequentially added to the surface of the dope column. The deposition sequence of different adhesive proteins around the fibroin column plays a role in the formation of solid silk (Sehnal and Sutherland, 2008). For example, the Ser3 protein of *B. mori* is produced in the anterior MSG section and is therefore placed in the outermost layer, to serve as a lubricant that reduces friction during the passage of the dope through the spinneret (Takasu et al., 2007).

Silk analyses were performed in detail for several species of the lepidopteran superfamily Bombycoidea ((Yukuhiro et al., 2016) for a recent review). *B. mori* silk contains three proteins in the fibroin fraction (BmFibH, BmFibL, and BmP25), several products of alternatively spliced transcripts of three sericin genes, and a few small seroins and other presumably nonstructural proteins. No orthologs of FibL and P25 were detected in *A. yamamai*, and further experiments have shown that

[*] Corresponding author. Biology Centre of the Czech Academy of Sciences, Institute of Entomology, Branisovska 31, 370 05, Ceske Budejovice, Czech Republic
[**] Corresponding author. Biology Centre of the Czech Academy of Sciences, Institute of Entomology, Branisovska 31, 370 05, Ceske Budejovice, Czech Republic.
E-mail addresses: sehnal@bc.cas.cz (F. Sehnal), zurovec@entu.cas.cz (M. Zurovec).

https://doi.org/10.1016/j.ibmb.2018.11.003
Received 22 October 2018; Accepted 14 November 2018
Available online 15 November 2018

two molecules of its FibH form homodimers (Tamura et al., 1987). As an exception to the fast evolutionary changes, the overall repetitive structure of fibroins, with small conserved regions in both ends, has been preserved through evolution even in Trichoptera (Yonemura and Sehnal, 2006). A closer examination of three *B. mori* sericin genes showed that they are distinct from each other and their protein products belong to separate classes (Takasu et al., 2007). In contrast, only one of five sericin genes predicted to encode *A. yamamai* sericins could be inferred to be orthologous with *B. mori* sericin 1, while four others cluster together and seem to be distantly related to *B. mori* sericin 3 (Zurovec et al., 2016). Similarly, one of six sericin genes detected in *S. cynthia ricini* seems to be related to *B. mori* sericin 1, while the other five are more similar to the *B. mori Ser3* gene (Tsubota et al., 2015).

In our previous studies, we isolated several silk genes in the greater wax moth, *Galleria mellonella*, by differential screening of a silk-gland-specific cDNA library with labeled cDNA probes or degenerate oligonucleotides based on N-terminal protein sequences. We identified and characterized cDNA sequences encoding GmFibH, GmFibL, and GmP25, as well as three closely related sericin-like genes, *GmMG1*, *GmMG2*, and *GmMG3* (Zurovec et al., 1995, 1998a, 2013). We also detected two protease inhibitors (GmSpi1 and 2) and seroin 1 (GmSn1) (Nirmala et al., 2001; Zurovec et al., 1998b). The three sericin genes identified in *G. mellonella* seem to be closely related and apparently belong to a sericin group represented by *B. mori Ser3* (Zurovec et al., 2013). We could not exclude the existence of proteins similar to other sericins.

The goal of the present study was the identification of all genes encoding *G. mellonella* silk components and elucidation of their evolutionary relationship to silk proteins of the Bombycoid species. We used *de novo* transcriptome sequencing of silk-gland-specific cDNA libraries and identified a set of candidate cDNAs encoding novel proteins with signal peptides. We also separated *G. mellonella* silk components by polyacrylamide gel electrophoresis (PAGE) and performed N-terminal sequencing in addition to identifying proteins and/or peptides by mass spectrometry (MS), thus confirming some of the novel proteins as silk components. By means of Northern blotting, we verified tissue-specific expression of more than 20 abundant cDNA candidates. Our results show that the number of sericin genes and the proportion of sericins in the cocoon silk are larger in *G. mellonella* compared to *B. mori*. Our study sheds light on the diversity and evolution of lepidopteran silk proteins.

2. Materials and methods

2.1. Insect dissection and histology

Larvae of the waxmoth *Galleria mellonella* L. (Lepidoptera, Pyralidae) were reared on a semi-artificial diet at 30 °C (Sehnal, 1966). The SG were dissected separately from the penultimate and the last-instar larvae whose age was measured in days from the preceding ecdysis. Entire SG were used for transcriptome preparation, while for northern analysis the SG were divided into four pieces: the posterior silk gland section (PSG) and the rear (R-MSG), central (C-MSG) and anterior (A-MSG) parts of the middle silk gland section. Subdivision of the MSG into three parts was performed at the position of the bends. Dissected tissues and body carcasses without SG were immediately frozen in liquid nitrogen and stored at −80 °C or used immediately after dissection.

2.2. Histology and electron microscopy

Hematoxylin-eosin staining: Dissected tissues were fixed overnight at 4 °C in Bouin-Hollande solution without acetic acid but supplemented with 0.7% mercuric chloride (Levine et al., 1995). Standard techniques were used for tissue dehydration, embedding in paraplast, sectioning to 7 µm, deparaffinization and rehydration. The sections were treated with Lugol's iodine followed by a 7.5% solution of sodium thiosulphate to remove residual heavy metal ions, washed in distilled water, stained with hematoxylin and eosin, dehydrated, mounted in DPX mounting medium (Fluka), and viewed and imaged under the microscope.

Semithin sections of cocoons: Pieces of freshly spun cocoon were prepared in phosphate buffer saline (PBS) and fixed in 2.5% glutaraldehyde in PBS for at least 4 h at room temperature (RT) or overnight at 4 °C. The samples were then washed in PBS supplemented with 4% glucose (three times for 15 min at RT) and subsequently treated with a 1:1 mixture of PBS and a 4% solution of OsO_4 (for 2 h at RT). The samples were washed again three times for 15 min at RT and then gradually dehydrated with acetone added to PBS at increasing concentrations of 30%, 50%, 70%, 80%, 90%, 95% and 100%, with 15 min of incubation for each. Dehydrated samples were transferred to acetone with an increasing amount of Epon resin (volume ratio of resin to acetone 1:2, 1:1 and 2:1, for 1 h at RT in each). The samples were then left in undiluted resin for 24 h at RT and then left to polymerize for 48 h at 62 °C. Semithin sections were cut with a glass knife and placed onto a droplet of 10% acetone on microscopic slide. Dried samples were stained with toluidine blue and imaged under a light microscope.

Scanning microscopy was performed as described earlier (Zurovec et al., 2016). Cocoon pieces were glued to aluminum holders, sputter-coated with gold, and analyzed with a Jeol 6300 scanning electron microscope.

2.3. Transcriptome preparation and analysis

RNA isolation, synthesis of cDNA libraries and RNA sequencing were performed as previously described (Zurovec et al., 2016). RNA was isolated using TRIzol Reagent (Invitrogen, Carlsbad, CA) and purified with a NucleoSpin RNA II kit (Macherey-Nagel, Duren, Germany). RNA integrity was checked, the concentration measured with a Bioanalyser 2100 (Agilent, Waldbronn, Germany) and the mRNA subsequently isolated with the aid of Oligo(dT)$_{25}$ Dynabeads (Ambion, Life Technologies). Reverse transcription was performed with a SMARTer PCR cDNA synthesis kit (Clontech). Roche GS-FLX 454 pyrosequencing was conducted according to the manufacturer's instructions. Independent transcriptomes were prepared from the SG of the penultimate-instar larvae (PI), post-feeding wandering last-instar larvae (WS), and apolyzing (initial phase of pupation) last-instar larvae (ECD). Sequencing produced 68,365, 76,492, and 81,015 reads, respectively. Software provided with the sequencer assembled the sequences into contigs. To analyze transcript abundance, sequence reads were calculated based on the number of hits received for local TBLASTN searches (Bioedit) in our transcriptome databases (using the sequence of 80 C-terminal amino acids as queries and a threshold e-value set to e^{-20}). Transcripts were manually annotated using a BLAST search and categorized as highly expressed (> 500), moderately expressed (25–500), and lowly expressed (< 25). We selected cDNAs encoding proteins with signal peptides for further analysis. Finally, we excluded from the candidate cDNAs those with close homologs in other species that had known functions unrelated to silk proteins. The sequences were deposited in GenBank**.**

2.4. Northern blotting

Total RNA was extracted with TRIzol Reagent (Invitrogen). RNA aliquots of 5 µg were taken for agarose electrophoresis, blotted onto a nylon membrane (Hybond N+, Sigma-Aldrich) and hybridized under high stringency conditions as previously described (Zurovec et al., 2016). Probes for northern blotting were synthesized by RT PCR using primers listed in Table S1 and labeled with α-^{32}P[dATP] using random priming with an oligo-labeling kit (Fermentas). Hybridization signals were detected by autoradiography using the storage phosphor screen of a STORM 860 Phosphorimager (Molecular Dynamics™).

2.5. Protein extraction

Freshly spun cocoons were cut into small pieces and submerged into a solution (10 mg silk per 250 μl) containing 8 M urea, 10 mM TrisCl pH 6.8, 2% SDS and 5% 2-mercaptoethanol for 48 h at RT, vortexed several times, and eventually centrifuged to obtain supernatants for SDS-PAGE. The dissolved silk proteins were separated on gradient gels (BioRad, 4–15%). Selected protein bands were cut out and used for MS. Alternatively, the PAGE gels were blotted to PVDF membrane and sent for N-terminal microsequencing to the commercial facility of the Medical College of Wisconsin (Milwaukee, WI) as described earlier (Zurovec et al., 1998b).

To determine the proportion of soluble silk components, 40 mg of dried *G. mellonella* and *B. mori* cocoons were cut into pieces and submerged into 8 M urea, 10 mM TrisCl pH 6.8, 2% SDS and 5% 2-mercaptoethanol for 48 h at RT. The samples were then centrifuged and the soluble fraction discarded. The undissolved silk remaining in the pellet was washed five times with water, vacuum dried and weighed.

2.6. Identification of protein fragments by mass spectroscopy

Mass spectroscopic analysis was performed as described previously (Zurovec et al., 2016). PAGE gels were stained with GelCode Blue Safe Stain (Thermo Scientific) and sliced to isolate distinct bands. The slices were incubated in 200 μl of 40% acetonitrile in 200 mM ammonium bicarbonate at 37 °C for 30 min. The solvent was discarded, the procedure repeated, and the snip dried in a SpeedVac for 30 min. A solution of proteomic-grade trypsin (20 μg/ml; Sigma-Aldrich) was added until the gel became saturated (10–20 μl). The samples were incubated for 45 min at 4 °C; the surplus of trypsin solution was removed and replaced with 20 μl of 9% acetonitrile in 40 mM ammonium bicarbonate. The protein digest was collected after overnight incubation at 37 °C. The solution was transferred to a new tube, and 20 μl of 9% acetonitrile in 40 mM ammonium bicarbonate was added to the gel pieces and incubated at 37 °C for 30 min. The solutions were then pooled and dried in a SpeedVac for 1 h. Immediately before analysis, the dried samples were resuspended in 0.1% formic acid, and a MS measurement was performed using a NanoAcquity Ultra Performance Liquid Chromatography (UPLC) coupled online to the ESI Q-TOF Premier mass spectrometer (the instrument, columns, and software were from Waters, UK). Peptides were separated by reverse-phase UPLC on a BEH300 C18 analytical column (75 mm i.d.; 150 mm length, particle size 1.7 mm) that was perfused at a 0.4 μl/min flow rate with 0.1% formic acid containing acetonitrile in a concentration increasing linearly from 3% (v/v, 1 min wash) to 40% for 30 min. Peptides eluted from the column flowed directly into the electrospray ionization source. Raw data for each sample was acquired in data-independent MSˆE mode and data-dependent survey analysis mode. In both modes, peptide and fragment spectra were acquired with a 2 and 5 ppm tolerance, respectively. PLGS2.3 software was used to match generated data with the entries in the transcriptome databases; identification of 2–3 consecutive y- or b-ions was required for a positive peptide match. A minimum of two peptides matched to a protein was considered a positive result.

2.7. Phylogenetic analysis

We used the following cDNA sequences for the construction of a phylogenetic tree: *Samia ricini* (LC001867, LC001868 partial sequence, LC001869 partial sequence, LC001870), *Antheraea yamamai* AySer1 (LC085887), AySer2 (LC085888), AySer3 (LC085889), AySer4 (LC085890), AySer5 (LC085891), *Bombyx mori* BmSer1 (NM_001044041), BmSer2l (NM_001172816), BmSer2s (NM_001172817), BmSer3 (NM_001114644), BmMucin4 (XM_021350007), *G. mellonella* GmMG1 (KC478777), GmMG2 (KC478778), together with novel proteins GmMG4-GmMG9, GmP150, GmP250 and GmMucin4-like. Coding sequences of the identified sericin genes were aligned with the MUSCLE program (Edgar, 2004) implemented in MEGA version 6 (Tamura et al., 2013). GTR + G was identified as the best model for calculating evolutionary distances in Smart Model Selection (Lefort et al., 2017) according to the lowest Bayesian information criterion (BIC) and Akaike information criterion (AIC) scores. Phylogenetic reconstruction with the maximum-likelihood (ML) method was performed in PhyML 3.0 (Guindon et al., 2010) and in MrBayes v3.2.6 for the Bayesian inference (BI) method (Ronquist and Huelsenbeck, 2003). Phylogenetic trees were visualized and finalized in MEGA6.

2.8. Prediction of protein secondary structures

The self-optimized prediction method with multiple alignments (SOPMA) tool (Geourjon and Deleage, 1995) was used for secondary structure prediction of candidate proteins (https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopma.html).

3. Results

3.1. Sections through G. mellonella larva

To examine the morphology of SG relative to their position in the body, we prepared a series of transverse sections through the last-instar larvae (Fig. 1A–D). The overall morphology of *G. mellonella* SG is similar to that of *B. mori*, but the PSG is relatively smaller and the MSG relatively larger. The narrow PSG tube (Fig. 1D) begins dorsally in the fifth abdominal segment, passes forward, and widens into the MSG, with a large lumen serving as a reservoir of synthesized silk proteins. Two characteristic bends delineate three MSG loops laid above one another (Fig. 1C). The ventrally located voluminous anterior part of the MSG narrows into the much thinner anterior silk gland section (ASG) that serves as a secretion channel ending in the spinneret on the distal labium (Fig. S1). Two distinct layers of secretory proteins can be distinguished in the lumen of the MSG with hematoxylin staining: the fibroins in the central part of the gland lumen stain light blue, while the dark blue/purple layer corresponds to the sericin coating. Successive sections through the MSG show that the proportion of sericins in the gland lumen increases toward the anterior end, due to the gradual synthesis of sericins in the MSG (Fig. S2).

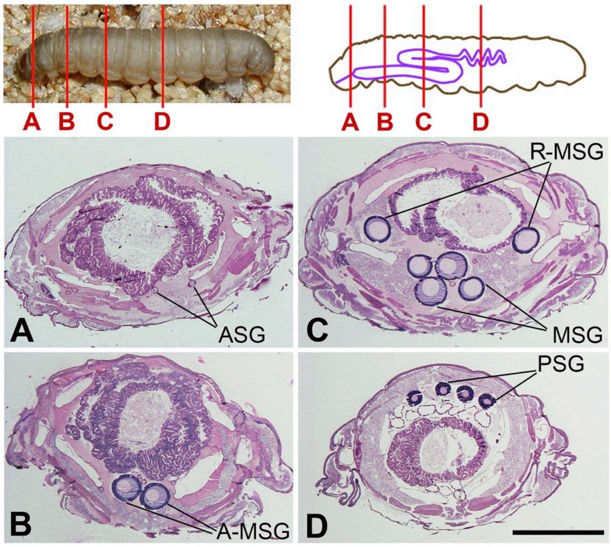

Fig. 1. Transverse sections through *G. mellonella* larva stained with hematoxylin-eosin. The order of successive sections and silk gland positions are schematically shown on the drawing above. ASG—anterior SG, MSG—middle SG, PSG—posterior SG, A-MSG—anterior part of middle SG; R-MSG—rear part of middle SG; hematoxylin-eosin staining; scale 1000 μm.

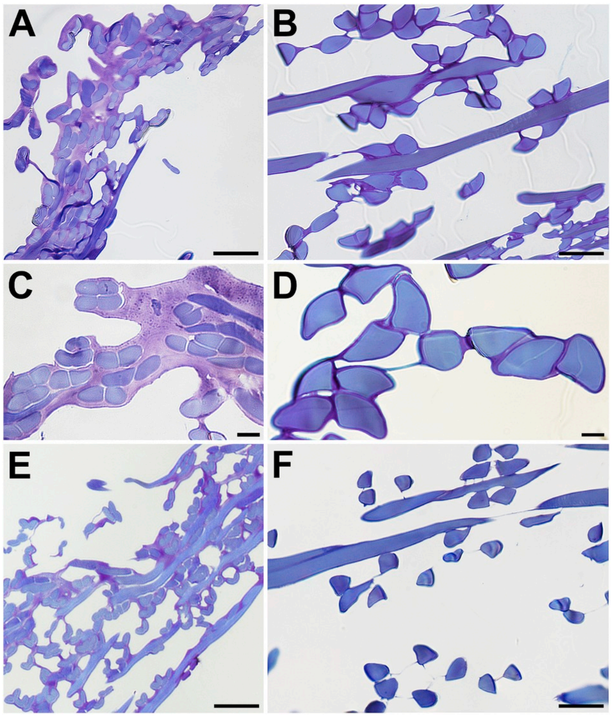

Fig. 2. Semithin sections through *G. mellonella* (A, C, and E) and *B. mori* (B, D, and F) cocoons. Sericins are reddish; fibroins appear blue. The sections were made before (A–D) and after degumming (E and F). Toluidine blue staining; scale A,B,E,F = 20 μm; C,D = 5 μm. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

3.2. The proportion of sericins in G. mellonella cocoons is larger than in those of B. mori

The cocoons of *G. mellonella* seem to be more compact than those of *B. mori*. To find out whether the proportion of fibroins and sericins differ between these species, we compared sections of *G. mellonella* and *B. mori* cocoons (Fig. 2). The sericins appear reddish when the toluidine blue staining technique is used. Sericins in both species seal the pair of fibroin filaments into a thread, but in *G. mellonella*, they also form a thick compact wrap that fills spaces between the threads, suggesting that the proportion of sericins in *G. mellonella* cocoons is higher than in those of *B. mori* (See Fig. 2). We also examined the outer and inner cocoon layers by scanning electron microscopy. The results show that the outer layer of a *G. mellonella* cocoon contains a higher proportion of sericin than the one from *B. mori* (Fig. 3).

In practical sericulture, silk reeling is possible after degumming; silk threads are liberated from the cocoon when sericins and other soluble proteins are dissolved in hot water and removed. To quantify the proportions of such proteins in silks, we dissolved cocoon proteins in 8 M urea that solubilized and removed most of the sericin layer. We weighed the undissolved silk and calculated the weight loss due to degumming (see Materials and Methods). As shown in Fig. 3f, sericins make up about 48% of the *G. mellonella* cocoon mass. By contrast, *B. mori* cocoons contained only about 26% soluble cocoon proteins.

The removal was verified by microscopic examination. As shown in Fig. 2e and f, the proportion of the soluble reddish mass was reduced dramatically, while the volume of the blue fiber mass remained similar. Sectioning of the native and degummed silk showed that the extraction procedure caused almost complete degumming of the *G. mellonella* cocoon and facilitated further analysis of the soluble silk proteins.

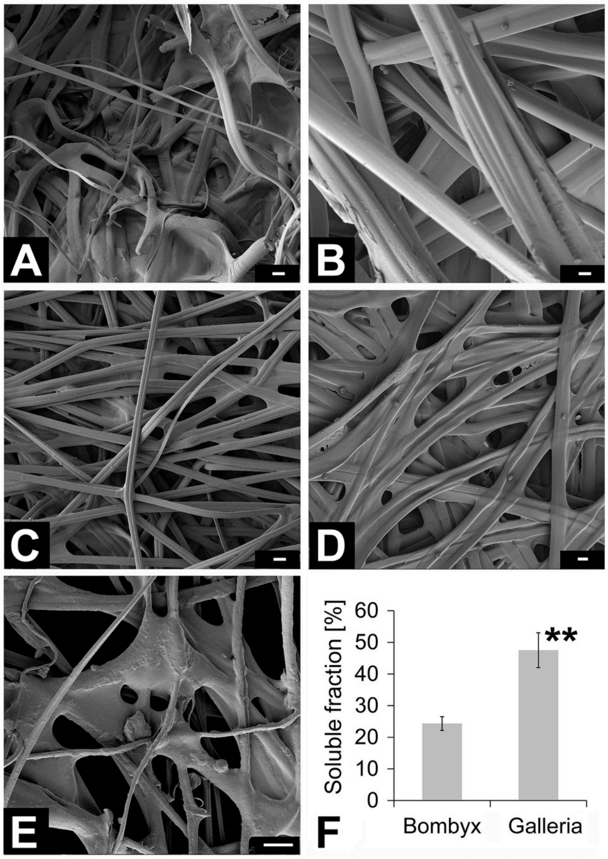

Fig. 3. Comparison of silks from *G. mellonella* and *B. mori* (A) Scanning microscopy of the outer side of *G. mellonella* cocoon, (B) outer side of *B. mori* cocoon, (C) inner side of *G. mellonella* cocoon, (D) inner side of *B. mori* cocoon, (E) silk mesh produced by *G. mellonella* PI larva, scale 10 μm, (F) chart showing average amounts of soluble proteins in *B. mori* (left) and *G. mellonella* (right) cocoons, measured as the loss of silk mass after degumming. **P < 0.01, *t*-test comparing means (n = 5). Values are given as mean ± SD.

3.3. Detection of candidate silk genes

Structural silk proteins secreted by SG cells are expected to be abundant and carry signal peptides at their N-terminal sequences. We sequenced three *G. mellonella* silk-gland-specific cDNA libraries and annotated more than 300 contigs. Most of the cDNAs encoded ribosomal proteins and proteins involved in peptide translation or trafficking. Potential secretory proteins accounted for approximately 24% of all annotated contigs. Among them, we identified cDNAs from the previously described silk genes *GmMG1*, *GmMG2*, *GmFibL*, *GmFibH*, *GmP25*, and *GmSn1* and the genes for protease inhibitors *GmSpi1* and *GmSpi2* (Table S2A). They all belong to highly or moderately expressed genes. From the remaining cDNAs, we excluded those encoding proteins with known functions unrelated to silk proteins. The remaining 27 cDNAs are shown in Tables 1A and 1B and Supplementary Fig. 3. The predicted proteins range in size from 8 to 300 kDa.

3.4. Expression of candidate silk genes

To estimate the abundance and temporal expression patterns of individual genes, we sequenced three transcriptomes based on the SG from penultimate instar (PI), wandering stage (WS), and prepupa (ECD) and obtained similar total numbers of reads in all cases. We divided the candidate genes into two groups based on the number of reads (Materials and methods). The first group consisted of 17 highly or moderately expressed genes, while the second group contained 10

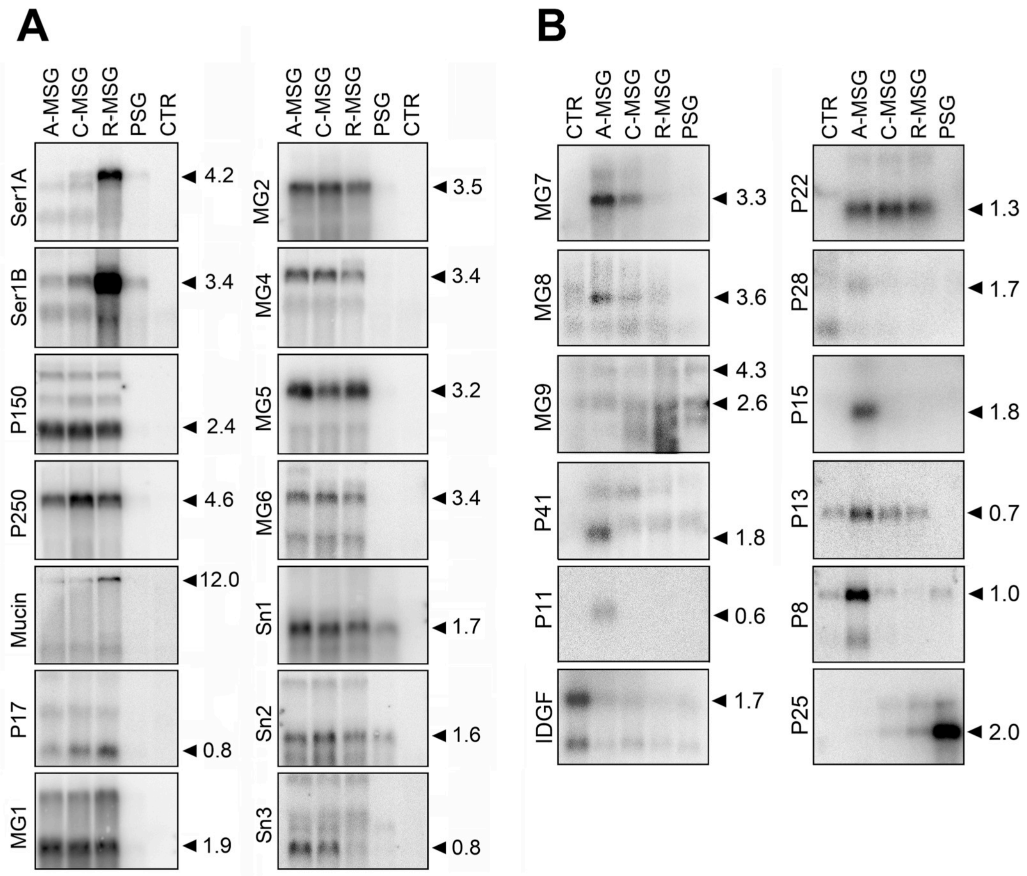

Fig. 4. Tissue-specific expression of genes encoding silk proteins. The genes were detected in the transcriptome study and the IDGF gene detected by proteomic analysis in indicated silk gland sections (GmMG1, GmMG2, GmSn1 and GmP25 were described earlier); A-MSG—anterior middle SG, C-MSG—central part of middle SG, R-MSG—rear middle SG, PSG—posterior SG, CTR—control larval carcass without SG). Total RNA (5 μg per lane) was resolved by electrophoresis and blotted on nylon membranes, which were then probed with ^{32}P-labeled cDNAs (designated on the left side). The estimated transcript size (in kb) is marked as a black arrowhead on the right, Mucin = GmMuc4-L.

lowly expressed ones.

Eleven cDNAs from the first group contain ORFs characterized by serine-rich repetitive sequences and may represent hitherto unknown sericin or mucin-like candidates. We named them *GmMuc4-like, GmSer1A, Gm Ser1B, GmP150, GmP250, GmMG4, GmMG5, GmMG6, GmMG7, GmMG8* and *GmMG9* (Table 1A). With the exception of GmMG8, the deduced sericin-like proteins are highly hydrophilic (Table 2A) having a grand average of hydropathy between −0.521 and −1.729 (Kyte and Doolittle, 1982). Two putative sericin candidates, GmSer1A and Gm Ser1B, include a short region of identity around the C-terminal Cys residues with the *B. mori* Ser1 protein (Fig. S4). Several other putative sericins, contain similar repeats with the GmMG1 and

Table 1A

GenBank accession numbers, number of reads and expression tissue specificity of silk components-encoding candidate genes with high and moderate levels of expression. "Muc" stands for mucins, "Ser" or "MG" for sericins, "Sn" for seroins, and "P" for other proteins distinguished by molecular weights. Number of reads is used for transcript-level abundance estimation and denotes numbers of 3' end-specific reads detected with TBLASTN using 80 C-terminal amino acids as a query sequence, with a threshold e-value set to −20. (PI)—the silk gland transcriptomes prepared from the penultimate-instar larvae; (WS)—post-feeding wandering last-instar larvae; (ECD)—apolyzing last-instar larvae. Tissue specificity of expression of the 18 mRNAs in parts of the SG was determined by Northern blot analysis (see Fig. 4). The predicted transcript sizes are compared with the approximate size of the transcripts as determined by hybridization; N.D. —the size or tissue specificity was not determined; nonspec – the transcript was not specific for the SG.

Candidate cDNA	GenBank number	Number of reads PI	WS	Ecd	Repetitive seq.	Tissue specificity	Predicted transcript	Northern transcript
GmMuc4-like	MG770312	4	7	15	Y	R-MSG	5.53	12.0
GmSer1A	MG770315	20	10	15	Y	R-MSG	3.36	4.2
GmSer1B	MG770316	118	96	339	Y	R-MSG	3.04	3.4
GmP150	MG770314	80	95	151	Y	MSG	1.93	2.4
GmP250	MG770313	835	3773	2594	Y	MSG	3.44	2.4
GmMG4	MG770317	1918	1907	4544	Y	MSG	3.13	3.4
GmMG5	MG770318	279	20	328	Y	MSG	2.20	3.2
GmMG6	MG770319	1	59	15	Y	MSG	2.78	3.4
GmMG7	MG846930	14	9	56	Y	A-MSG	2.41	3.3
GmMG8	MG846934	7	0	48	Y	A-MSG	1.84	3.4
GmMG9	MG992435	0	7	125	Y	R-MSG?	3.00	4.3; 2.6
GmP28	MG846875	44	1	9	N	A-MSG	1.14	1.7.
GmP15	MG846881	37	114	111	(Y)	A-MSG	0.45	1.8
GmSn2	MG604942	97	154	300	(Y)	MSG	0.78	1.6
GmP13	MG846890	23	25	112	N	Nonspec.	0.54	0.7
GmSn3	MG604945	3	12	40	(Y)	MSG	0.68	0.8
GmP8	MH464805	4	22	25	N	Nonspec.	0.34	1.0

Table 1B
GenBank accession numbers, number of reads and expression tissue specificity of silk components-encoding candidate genes with low levels of expression.

Candidate cDNA	GenBank number	Number of reads PI	WS	Ecd	Repetitive seq.	Tissue specificity	Predicted transcript	Northern transcript
GmP47	MG992430	0	1	0	N	N.D.	1.32	N.D.
GmP41	MG846870	12	0	4	N	A-MSG	1.33	1.8
GmP38	MG992414	0	0	1	N	N.D.	1.37	N.D.
GmP32	MG846903	0	0	1	N	N.D.	1.46	N.D.
GmP22	MG770325	2	3	1	Y	MSG	1.10	1.3
GmP17	MG770329	3	1	2	N	MSG	0.63	0.8
GmP14	MG770327	4	0	0	N	N.D.	0.59	N.D.
GmP12	MH464803	3	0	0	N	N.D.	0.43	N.D.
GmP11	MH464804	5	3	4	N	A-MSG	0.48	0.6
GmP7	MH464806	2	2	2	N	N.D.	0.44	N.D.

Table 2A
Properties of putative candidate silk proteins with high or moderate levels of expression. Major amino acids show the percentage of the two most abundant residues; N-terminal sequence shows first 10 N-terminal amino acid residues; predicted hydrophobicity and pI were determined using the ExPASy ProtParam tool (http://us.expasy.org/tools/prot param.html). % of predicted coil denotes percentage of random coil determined by SOPMA tool (https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopma.html; see Supplementary Fig. 1 for full SOPMA prediction.

Candidate cDNA	Major Amino acids	N-terminal sequence	Average Hydropathicity	Predicted pI	% of predicted coil (SOPMA)
GmMuc4-like	Ser 20.3%; Gln 12.0%	MKLYVTTVVA	−0.883	4.61	44.13
GmSer1A	Ser 22.3%; Gln 14.4%	MMGGFKTFVC	−0.521	3.95	61.65
GmSer1B	Ser 35.4%; Ala 21.4%	MRWLYVFASV	−0.848	5.92	57.57
GmP150	Ser 25.9%; Gln 19.5%	MKSPGFLTFT	−1.729	4.48	65.03
GmP250	Ser 50.2%; Asn 16.0%	MKFSILLVAA	−1.083	10.01	91.71
GmMG4	Ser 44.9%; Asn 22.1%	MSLKLIVLAA	−1.121	5.92	74.69
GmMG5	Ser 45.9%; Asn 13.6%	MRSSFVLVAL	−1.266	10.32	67.66
GmMG6	Ser 55.2%; Asn 15.9%	MKLSLVLLAF	−1.568	9.75	85.67
GmMG7	Ser 38.1%; Asn 21.6%	MISNMRLALL	−0.778	3.54	74.53
GmMG8	Ser 32.9%; Gly 22.8%	MIANMRFALL	0.045	3.77	64.45
GmMG9	Ser 51.4%; Ala 18.0%	MKFALLLVMA	−0.664	4.76	69.56
GmP28	Leu 10.0%; Ser 9.2%	MYFYKLVACI	−0.45	8.85	50.60
GmP15	Glu 18.4%; Pro 17.7%	MKIVLALFAC	−0.804	4.33	68.79
GmSn2	Gly 12.5%; Ser 11.1%	MGSVLSGALL	−0.343	5.82	51.75
GmP13	Pro 13.3%; Val 10.8%.	MKGAVIVLVA	−0.677	4.88	65.83
GmSn3	Asn 12.3%; Val 12.3%	MSRLTVVFVL	−0.058	7.86	47.17
GmP8	Lys 12.7%; Pro 11.4%	MRFAVICLVL	−0.381	9.00	59.49

Table 2B
Properties of putative candidate silk proteins with low level of expression.

Candidate cDNA	Major Amino acids	N-terminal sequence	Average Hydropathicity	Predicted pI	% of predicted coil (SOPMA)
GmP47	Asn 15.4%; Glu 13.7%	MELPLYIFLG	−1.124	4.11	56.73
GmP41	Ser 10.3%; Val 10.0%	MILNKIVISI	−0.175	4.47	55.78
GmP38	Pro 9.8%; Thr 8.6%	MNTVTILFGI	−0.798	6.53	63.50
GmP32	Ala 21.1%; Lys 15.3%	MKVLLLCLAF	−1.015	4.67	56.82
GmP22	Thr 18.0%; Ser 14.7%	MADIKCVLFI	−0.37	4.74	62.67
GmP17	Thr 16.6%; Pro 10.8%	MKKVLFVCLI	−0.859	7.56	73.55
GmP14	Pro 13.8%; Glu 10.8%	MYKSLVVLCV	−0.468	5.53	54.62
GmP12	Gly 31.4%; Ser 17.4%	MAFKIACLLL	−0.625	8.02	72.73
GmP11	Ser 12.1%; Leu 11.1%	MSSFGVALVF	−0.291	5.17	33.33
GmP7	Ile 16.1%; Asp 12.9%	MENNKIIIYL	0.055	3.73	27.42

GmMG2 proteins detected earlier. We did not detect any *GmMG3*-specific cDNA described previously (Zurovec et al., 2013), suggesting that it is a pseudogene or has a very low expression level. The remaining members of this group were characterized by lower predicted molecular weights and contained two putative seroin paralogs (*GmSn2* and *GmSn3*), as well as genes *GmP28*, *GmP15*, *GmP13*, and *GmP8* (see Table 2B).

The second group consists of ten genes with lower transcription level (Table 1B). They contain no or only a short region of repeats, except for *GmP17*, and *GmP22*. Several putative proteins of this group seem to be more conserved and show homology to some uncharacterized predicted proteins from *B. mori* or other lepidopterans, including GmP47, GmP41, GmP38, GmP32, GmP14.

To further characterize the identified cDNAs and map the expression of the respective genes within the SG (PSG, R-MSG, C-MSG, and A-MSG), we performed northern blots and examined the candidate cDNAs together with a cDNA encoding the imaginal disc factor (GmIDGF) (Kucerova et al., 2016) and several silk genes known earlier (*GmMG1, GmMG2, GmP25, GmSn1*). As shown in Fig. 4, almost all examined cDNAs showed MSG-specific transcription, except for GmP8, GmP13 and GmIDGF, which were also expressed in the carcass left after SG removal. Most other genes (*GmP150*, *GmP250*, *GmMG5*, *GmP22*, and *GmP17*) were expressed equally in all MSG parts. *GmSer1A* and *GmSer1B* genes seemed to be expressed in the rear MSG and also, to a lesser extent, in the PSG. The cDNAs of the high-MW *GmMucin4-like* appeared to be expressed mainly in the R-MSG. In contrast, the expression of several candidates, including *GmMG7*, *GmMG8*, *GmP41*, *GmP15*, and *GmP11* occurred predominantly in the A-MSG. We did not detect any novel PSG-specific cDNA.

The sizes of RNAs determined by northern blot analysis were sometimes larger than the sizes of the respective cDNAs determined by transcriptome sequencing (*GmMuc4-like*, *GmP22*), suggesting that in some cases we did not obtain full-length sequences, probably due to difficult assembling of the repetitive regions. The probe GmMG9 hybridizes with multiple bands, suggesting that there are multiple GmMG9 transcripts generated by alternative splicing or cross-reactivity with related transcripts from similar gene(s).

3.5. Analysis of cocoon silk components

G. mellonella larvae, from the second instar onward, produce thin protective silk tubes for their defense against bees. When they terminate feeding in the last instar, they spin dense cocoons. We collected silk from the tubes and cocoons, dissolved it in 8 M urea buffer, and separated protein fractions by SDS-PAGE. The fraction patterns of the two kinds of silk were similar, but the quantities of some proteins were different (Fig. 5) About 30 visible bands were excised for MS analysis. The translated protein sequences were used as targets for the peptide mappings. We used thorough extraction conditions, as shown in Fig. 2, to ensure that the vast majority of soluble components were subjected to PAGE separation.

As shown in Tables 3 and 4, the major PAGE bands were assigned to multiple proteins and some bands contained proteolytic products of high-MW proteins. The MS analysis identified 20 proteins. Six of them were candidates from transcriptome sequencing (GmP150, GmP250, GmMG4, GmMG5, GmP22, and GmP15), and five were known from our previous work (GmFibH, GmFibL, GmP25, GmSn1, and the protease inhibitor GmSpi1). The remaining nine proteins corresponded to the cDNAs not included among the candidates during transcriptome annotations (described above) because their homologs in other species had functions unrelated to silk. The identified genes encoded the protease inhibitor zonadhesin, imaginal growth factor (GmIDGF), several enzymes (apyrase, acetylcholine esterase), fatty acid binding protein, heat shock protein Hsp20, hexamerin, and GmP16 (Table S2B).

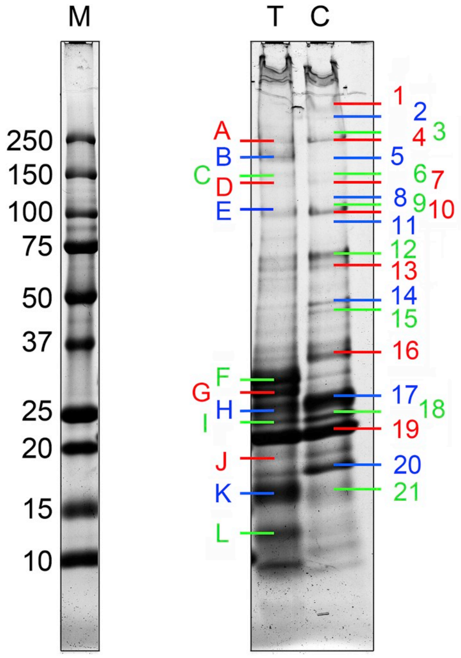

Fig. 5. Separation of proteins extracted from the silk tubes (lane T) and silk cocoons (lane C). The silk proteins were solubilized, separated by SDS-PAGE, and identified by tandem MS by searching against translated cDNA databases. The data are shown in Tables 3 and 4 and in the Supplementary Table S2. Letters (A–L) denote the isolated protein bands from protective tube silk; numbers (1–21) mark the bands in cocoon silk (the letters and numbers are colored to facilitate distinction of the protein bands). The left lane M shows molecular weight standards.

Table 3

Summary of extracted cocoon silk proteins identified using MS. Accession numbers (GenBank) of identified silk proteins and numbers of matching peptide fragments detected by MS with data-independent (Id) and data-dependent (dda) modes of analysis are shown (see Fig. 3 for a PAGE and Supporting Table S3 for complete MS data).

Band	MW (Da)	GenBank	Product name	Id(MSΛE)	dda(survey)
1	300	MG770312	GmMuc4-L	24	8
2	280	MG770312	GmMuc4-L	11	4
2	280	AH009792	GmFibH	3	0
3	260	AH009792	GmFibH	2	0
3	260	MG770312	GmMuc4-L	0	2
4	250	AH009792	GmFibH	4	1
4	250	MG770313	GmP250	5	0
5	200	MG770313	GmP250	10	1
5	200	AH009792	GmFibH	2	0
6	130	AH009792	GmFibH	4	1
6	130	MG770320	Zon-like	9	0
7	120	MG770320	Zon-like	4	1
7	120	MG770314	GmP150	2	3
7	120	AH009792	GmFibH	5	0
8	110	MG770314	GmP150	3	5
9	105	MG770324	GmHex2	3	9
10	100	MG770325	GmP22	9	3
10	100	MG770324	GmHex2	0	3
10	100	MG770323	Apy-like	13	2
10	100	M73793	GmHex2	2	3
10	100	AH009792	GmFibH	2	0
11	95	MG770325	GmP22	3	1
11	95	MG770323	Apy-like	2	3
12	70	MG770323	Apy-like	10	12
13	65	MG846884	Ach-like	5	4
13	65	MG770323	Apy-like	0	1
14	48	MG846880	GmIDGF	8	4
15	45	MG846880	GmIDGF	3	3
16	30	AF009828	GmSn1	2	4
17	25	AF009827	GmP25	18	9
18	23	AF009827	GmP25	8	2
18	23	MG846881	GmP15	2	2
18	23	S77817	GmFibL	2	2
19	22	S77817	GmFibL	28	5
19	22	MG846881	GmP15	6	1
19	22	AF009827	GmP25	8	1
20	17	AF009828	GmSn1	6	3
20	17	S77817	GmFibL	2	2
21	16	MG770326	GmP16	22	4
21	16	S77817	GmFibL	4	1
21	16	AF009827	GmP25	3	4
21	16	MG992392	Fatty ac. b.	3	2

Several cocoon silk proteins separated by PAGE were taken for N-terminal sequencing (Table 5). This analysis detected an N-terminus of a 43 kDa protein corresponding to the longest alternative splicing protein isoform of GmSn1. The N-terminus was also sequenced in the 22.5 kDa GmSn1 isoform and on bands identified as GmP15, GmP17, GmP16, and GmSpi1.

3.6. Evolutionary relationships among serine-rich proteins

To find a platform for the classification of complete sequences of serine-rich proteins, we aligned the putative *G. mellonella* sericin sequences with known sericins from *B. mori*, *A. yamamai*, and *S. cynthia ricini* (Fig. S6) (Dong et al., 2015; Kludkiewicz et al., 2009; Tsubota et al., 2015). We chose mucin 4 of *B. mori* and a similar serine-rich mucin found in *G. mellonella* to make the outgroup for rooting the

Table 4
Summary of extracted larval tube silk proteins identified using MS. Accession numbers (GenBank) of identified silk proteins and numbers of matching peptide fragments detected by MS with data-independent (Id) and data-dependent (dda) modes of analysis are shown (see Fig. 3 for a PAGE and Supporting Table S3 for complete MS data).

Band	MW (Da)	GenBank	Product name	Id(MSΛE)	dda(survey)
A	240	MG770313	GmP250	5	0
A	240	MG770317	GmMG4	5	0
A	240	AH009792	GmFibH	2	2
A	240	MG770312	GmMuc4-L	1	2
B	180	AH009792	GmFibH	5	1
B	180	MG770318	GmMG5	2	3
B	180	MG770313	GmP250	2	1
B	180	MG770317	GmMG4	4	0
B	180	MG770312	GmMuc4-L	3	4
C	145	MG770312	GmMuc4-L	3	1
C	145	AH009792	GmFibH	5	1
C	145	MG770318	GmMG5	3	1
D	140	AH009792	GmFibH	3	5
E	110	AH009792	GmFibH	4	0
E	110	MG770314	GmP150	2	4
E	110	MG770318	GmMG5	1	1
E	110	MG770312	GmMuc4-L	2	0
F	31	S77817	GmFibL	2	1
G	29	AF009827	GmP25	4	0
H	25	S77817	GmFibL	1	1
H	25	AF009827	GmP25	6	1
I	23	AF009827	GmP25	22	3
I	23	S77817	GmFibL	2	1
I	23	AH009792	GmFibH	8	0
I	23	MG992393	GmHsp20	6	1
I	23	MG770323	Apy-like	2	0
J	17	S77817	GmFibL	0	2
K	15	S77817	GmFibL	8	1
L	13	S77817	GmFibL	0	2

Table 5
N-terminal amino acid sequences of major *G. mellonella* silk proteins separated by PAGE and analyzed by microsequencing.

Band	Sequencing result	Matching sequence	Reference	Annotation
150 kDa	NDGDNQNGQVVT	NDGDNQNGQVVT	MG770314	GmP150
45 kDa	GFVWVDDDNNRF	GFVWVDDDNNRF	MG604941	GmSn1
22.5 kDa	GFVXVDDDNNRF	GFVWVDDDNNRF	MG604948	GmSn1
17.5 kDa	GKGCDDGGDGKG	GKGCDDGGDGKG	MG770329	GmP17
16.3 kDa	MNIMGYLITLAN	MNIMGYLITLAN	MG770326	GmP16
5.4 kDa	DDICSLPLKTGP	DDICSLPLKTGP	AF292098	GmSpi1

*N-terminal sequences were obtained only for major proteins of each band (Fig. 5). All sequences matched proteins deduced from the cDNAs deposited in our transcriptome database.

phylogenetic tree. The resulting phylogram shows that only two clades are well separated: mucins, and the sericins of type 1 (plus GmMG9 from *G. mellonella*) (Fig. 6a). The remaining sequences clustered mostly according to the species of their origin, suggesting recent duplication events or concerted evolution of proteins containing repetitive sequences (Elder and Turner, 1995). The incomplete sequence LC001866 from *S. cynthia ricini* (Tsubota et al., 2015) was not included in the analysis.

The presence of repetitive regions makes the alignment and phylogenetic analysis of proteins difficult. We also made a tree based only on the N— and C-terminal sequences, thus skipping most of the repetitive regions of these proteins. As shown in Fig. 6b, the clades of mucins and sericins of type 1 are well supported, and GmMG9 stands separated from other *G. mellonella* sericins; however, the position of mucins as a root has changed. The distribution of sericins according to the species of their origin was again evident.

The most diversified sericin sequences are *B. mori* type 2 sericin, *G. mellonella* sericin P150, and *S. cynthia ricini* LC001869 (compare branch lengths in Fig. 6a and b). When we exclude them from the analysis, the resulting phylogenetic tree becomes robust and distinguishes three highly supported clades—mucins, sericins of type 1, and sericins of type 3 (Fig. S7). Sericin 3 is represented by a single gene in *B. mori*, in contrast to other examined species in which there are several paralogs (Dong et al., 2015; Tsubota et al., 2015; Zurovec et al., 2016). Similarity among the intraspecific paralogs suggests that they represent the consequences of relatively recent duplication events.

4. Discussion

In attempt to detect all components of *G. mellonella* silk, we analyzed silk-gland-specific transcriptomes prepared from three developmental periods: PI larvae, WS larvae, and ECD larvae. Twenty-seven candidate cDNAs encoding putative secretory proteins were selected from silk gland transcriptomes and further tested for the spatial and developmental patterns of their expression, matching of deduced peptide sequences with proteins identified in the silk extracts, and similarity of the encoded proteins with previously identified silk components. Our results show that the candidates are produced almost exclusively in the MSG and represent soluble silk components.

Earlier analysis of *B. mori* silks by LC-MS identified 500 proteins that included fibroin and sericin components, several proteinase inhibitors, ten abundantly secreted proteins of unknown function, and a number of rare proteins (Dong et al., 2013). For *G. mellonella*, we used MS analysis of PAGE-separated protein bands and identified 20 abundant proteins. These included five proteins previously known to be parts of *G. mellonella* silk (GmFibH, GmFibL, GmP25, seroin 1, and protease inhibitor GmSpi1) together with eight abundantly transcribed candidate gene products denoted GmMuc4-like, GmP250, GmP150, GmMG4, GmMG5, GmP22, GmP15, and GmP17, and several others with functions unrelated to silk, as suggested by similarity to known proteins in other species. We did not detect the highly abundant proteins GmMG1 and GmMG2 described previously (Zurovec et al., 2013), likely because our MS method includes partial sample hydrolysis by trypsin that attacks cleavage sites containing Arg and Lys. The absence or low number of trypsin cleavage sites renders such proteins unsuitable for our method.

Among the novel proteins identified in an earlier analysis of *B. mori* silk, there is an 18 kDa glycine-rich cell wall structural protein 1.0-like (BGIBMGA001358-PA) detected as a highly abundant novel silk component (Dong et al., 2015). We found that *G. mellonella GmP17* cDNA encodes a similar 17 kDa protein that contains AspGlyLysGly repeats and is specifically transcribed in the MSG (Fig. 4). The full-length sequence of *G. mellonella* P17 is unknown because the highly repetitive character of this molecule causes problems with sequencing. Proteomic analysis of *G. mellonella* silk revealed some metabolic enzymes, including apyrase and acetylcholine esterase, whose significance for cocoon formation is not clear. *B. mori* silk was shown to contain a number of enzymes that may reflect the ancestral function of labial glands in digestion (Dong et al., 2013). Apyrase is known from the saliva of *Spodoptera frugiperda* (Acevedo et al., 2017), and acetyl choline esterases were detected in the salivary glands of the hemipteran bug *Cimex lectularius* (Seong et al., 2012). Interestingly, in *G. mellonella* silk, we found a significant amount of the imaginal disc growth factor (IDGF). IDGF is known from diverse species and tissues, but this is the first report of its occurrence in silk. *Drosophila* IDGF homologs were implicated in innate immune responses against microorganisms (Broz et al., 2017; Kucerova et al., 2016), and *G. mellonella* IDGF might have a similar antimicrobial function. Another abundant *G. mellonella* silk protein was identified as the proteinase inhibitor zonadhesin, known from the silk of *A. yamamai* (Zurovec et al., 2016).

The sericin family forms the largest group of the putative novel silk components (GmMG4, GmMG5, GmMG6, GmMG7, GmMG8, GmMG9, GmSer1A, GmSer1B, GmP150, and GmP250). They are characterized by a high proportion of serine (22–55%) as a major amino acid, in addition to the presence of repetitive sequences. *B. mori* sericins were

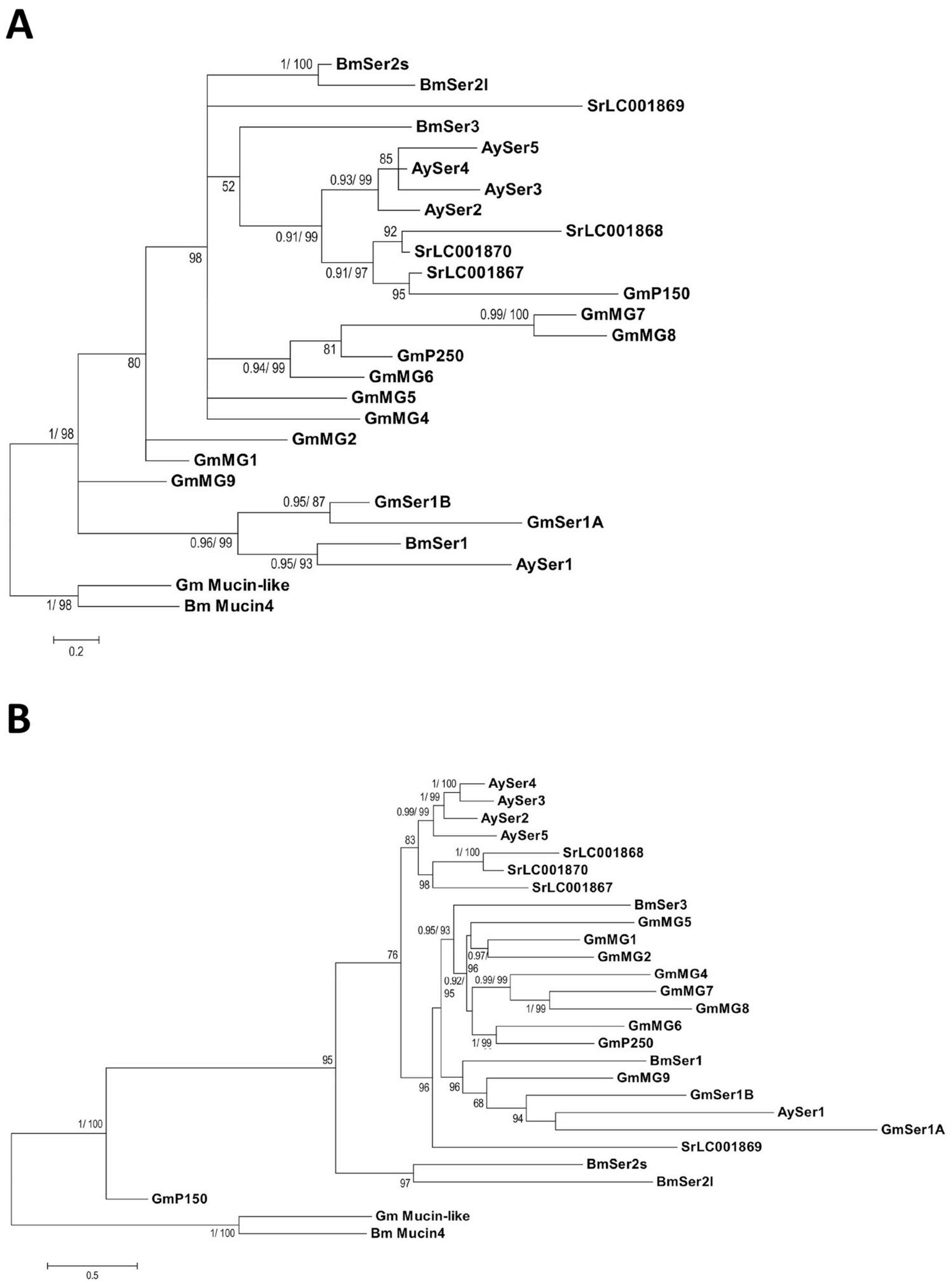

Fig. 6. Phylogenetic tree of sericin and mucin4 genes from *G. mellonella* (Gm), *B. mori* (Bm), *A. yamamai* (Ay), and *S. cynthia ricini* (Sr). (A) ML tree of complete cDNA sequences subtree pruning and regrafting (SPR search method, ten replicates, aBayes statistics higher than 50 shown in nodes). The same topology was found with the BI tree search (posterior probabilities higher than 0.9 are shown in well supported nodes). The alignment is shown in Fig. S6. **(B).** Phylogenetic tree based on the 5'- (approximately 72 bp) and 3'-ends of cDNAs (approximately 315 bp). BI tree search method, 5,000,000 generations (50% burn-in), nodes support is presented as posterior probabilities values higher than 0.9. In addition, aBayes statistic values are shown (higher than 50), which were calculated for the ML tree and exhibit similar topology, Gm Mucin-like = GmMuc4-L.

shown earlier not to form oriented structures and also Fourier transform infrared spectroscopy (FTIR) revealed the prevalence of random coils (Teramoto and Miyazawa, 2005; Wang et al., 2014). Secondary structure prediction of our *G. mellonella* candidate silk proteins suggested that putative sericins are also dominated by random coils (Table 3, Fig. S3). Some putative *G. mellonella* sericins seem to be closely related to each other (Fig. 7 and S4). The mutual similarity of sericins suggests that they arose by gene duplication. There is very low sequence conservation between sericins from different species. The exception is a short conserved region around two cysteine residues at the C-terminus of Ser1 in *B. mori* and *A. yamamai* Ser1 as well as in the Ser1A and Ser1B proteins of *G. mellonella* (Fig. S4).

Earlier studies in *B. mori* suggested that the SG is subdivided into three sections (PSG, MSG, and ASG) and functional subdivisions can be also recognized within the MSG. The individual parts differ by the expression patterns and by the presence of specific transcription factors (Li et al., 2015; Takiya et al., 2016). The expression of sericin genes in MSG compartments correlates with the deposition of their product to the fibroin core of the silk fiber, so that *B. mori* sericin 1 expressed in the R-MSG is deposited first and localized in the innermost sericin layer (Takasu et al., 2007). Similarly, the *A. yamamai Ser1* gene differs from other sericins by its expression in not only the MSG but also the PSG (Zurovec et al., 2016). We show here that the two *G. mellonella* paralogs of *Ser1* are expressed in the PSG and adjacent R-MSG regions, suggesting that the Ser1 protein is deposited onto the fibroin core as the first layer. Similarly, *G. mellonella* sericin MG-7 is expressed mainly in the A-MSG (Fig. 4), and its expression pattern resembles that of *B. mori* Ser3. *B. mori* Ser3 was suggested earlier as being localized at the very surface of the silk threads (Takasu et al., 2007; Wang and Zhang, 2011). Sericins are deposited around the fibroin core sequentially so that hydrophilic proteins, such as BmSer1, occur in the innermost layer and the less hydrophilic proteins, such as BmSer3, in the surface layer (Wang and Zhang, 2011). The general organization of sericin layers at the surface of the silk threads might be conserved in Lepidoptera.

Earlier studies also suggested that sericins can be classified into three subgroups represented by the *B. mori* genes *Ser1*, *Ser2*, and *Ser3*

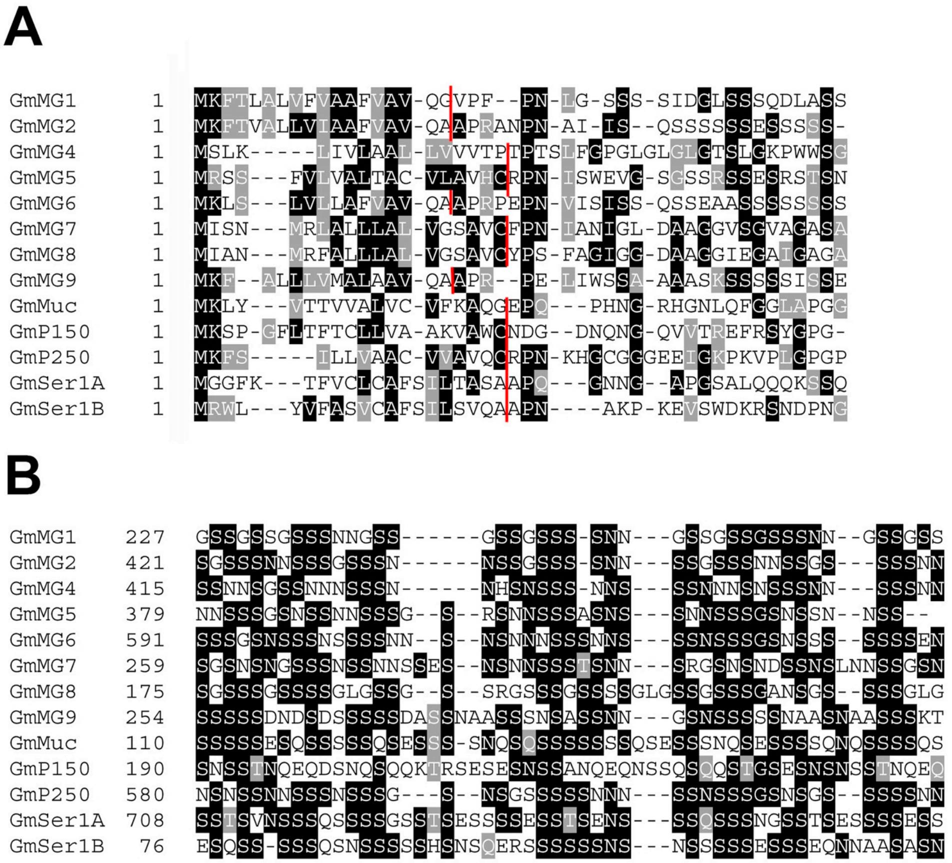
A
GmMG1 1 MKFTLALVFVAAFVAV-QGVPF--PN-LG-SSS-SIDGLSSSQDLASS
GmMG2 1 MKFTVALLVIAAFVAV-QAAPRANPN-AI-IS--QSSSSSSESSSSS-
GmMG4 1 MSLK-----LIVLAAL-LVVVTPTPTSLFGPGLGLGLGTSLGKPWWSG
GmMG5 1 MRSS---FVLVALTAC-VLAVHCRPN-ISWEVG-SGSSRSSESRSTSN
GmMG6 1 MKLS---LVLLAFVAV-QAAPRPEPN-VISISS-QSSEAASSSSSSSS
GmMG7 1 MISN---MRLALLLAL-VGSAVCFPN-IANIGL-DAAGGVSGVAGASA
GmMG8 1 MIAN---MRFALLLAL-VGSAVCYPS-FAGIGG-DAAGGIEGAIGAGA
GmMG9 1 MKF--ALLLVMALAAV-QAAPR--PE-LIWSSA-AAASKSSSSSISSE
GmMuc 1 MKLY---VTTVVALVC-VFKAQGEPQ---PHNG-RHGNLQFGGLAPGG
GmP150 1 MKSP-GFLTFTCLLVA-AKVAWCNDG--DNQNG-QVVTREFRSYGPG-
GmP250 1 MKFS-----ILLVAAC-VVAVQCRPN-KHGCGGGEEIGKPKVPLGPGP
GmSer1A 1 MGGFK---TFVCLCAFSILTASAAPQ---GNNG--APGSALQQQKSSQ
GmSer1B 1 MRWL---YVFASVCAFSILSVQAAPN----AKP-KEVSWDKRSNDPNG
B
GmMG1 227 GSSGSSGSSSNNGSS------GSSGSSS-SNN---GSSGSSGSSSNN--GSSGSS
GmMG2 421 SGSSSNNSSSGSSSN------NSSGSSS-SNN---SSGSSSNNSSGS---SSSNN
GmMG4 415 SSNNSGSSNNNSSSN------NHSNSSS-NNS---SSNNNSNSSSNN---SSSNN
GmMG5 379 NNSSSGSNSSNNSSSG--S--RSNNSSSASNS---SNNSSSGSNSSN--NSS--
GmMG6 591 SSSGSNSSSNSSSSNN--S--NSNNNSSSNNS---SSNSSSGSNSSS--SSSSEN
GmMG7 259 SGSNSNGSSSNSSNNSSES--NSNNSSSTSNN---SRGSNSNDSSNSLNNSSGSN
GmMG8 175 SGSSSGSSSSGLGSSG--S--SRGSSSGSSSSGLGSSGSSSGANSGS--SSSGLG
GmMG9 254 SSSSSDNDSDSSSSSDASSNAASSSNSASSNN---GSNSSSSSNAASNAASSSKT
GmMuc 110 SSSSSESQSSSSSSQSESSS-SNQSQSSSSSSSQSESSSNQSESSSSQNQSSSSQS
GmP150 190 SNSSTNQEQDSNQSQQKTRSESESNSSANQEQNSSQSQQSTGSESNSNSSTNQEQ
GmP250 580 NSNSSNNSSSNSSSG---S--NSGSSSSSNNN---SSNSSSGSNSGS--SSSSNN
GmSer1A 708 SSTSVNSSSQSSSSGSSTSESSSSSESSTSENS---SSQSSSNGSSTSESSSSESS
GmSer1B 76 ESQSS-SSSQSNSSSSSHSNSQERSSSSSNS---NSSSSSESSSEQNNAASASN

Fig. 7. Alignment of parts of amino acid sequences of putative serine-rich proteins. (**A**) The N-terminus—the signal peptide cleavage sites are indicated by a vertical red line. (**B**) Similarity of serine-rich tandem repeats. Protein sequence alignment was performed in Megalign (Lasergene) and presented using BOXSHADE version 3.2 at a threshold fraction of 0.3. Black boxes indicate identical and gray boxes similar residues, GmMuc = GmMuc4-L. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

(Yukuhiro et al., 2016). Our results suggest that the evolutionary relationships might be more complex. In our study, we prepared alignment and constructed phylogeny trees from serine-rich protein sequences of *G. mellonella*, *B. mori*, *A. yamamai*, and *S. cynthia ricini.* We created sequence alignments that either included or skipped the repetitive regions, which produced different results. However, in both cases, the phylogenetic trees showed well-supported separation of sericin-1-like proteins from other sericin molecules. The branching of the phylogenetic trees also supports the idea that sericins similar to *B. mori* Ser3 undergo expansion. The clustering seems to be rather species specific. The similarity of sericins within species thus sharply contrasts with the divergence between species. This result suggests that the paralogs within a species appear more similar due to concerted evolution (Elder and Turner, 1995) or that gene duplications occurred independently in different species and relatively recently. The species-specific branching of sericins rather seems to follow the birth-and-death multigene family evolution model (Nei et al., 1997). According to this model, new genes are formed by repeated gene duplication, and some duplicate genes are retained in the genome, while others are lost or become nonfunctional due to deleterious mutations. The rapid sericin gene turnover characteristic for this model is supported by the presence of two paralogs of the *Ser1* gene, the existence of the putative pseudogene *GmMG3*, and species-specific clustering of most sericins in the phylogenetic tree. More information is needed to determine the exact evolutionary relationships among sericin proteins.

Transcripts encoding the non-sericin candidate silk proteins are quite heterogeneous and produced in different MSG regions, suggesting their distinct positions in surface silk layers. Some of these putative proteins may have a repetitive structure. A serine-rich protein of more than 300 kDa containing more than 20% serine residues shows significant similarity to the *B. mori* mucin-4 protein and was annotated as mucin4-like (GmMuc4-L). This protein was also detected by proteomic analysis, showing that it is a genuine cocoon component. It contains several types of repeats, some of which are rich in serine or threonine. Similar mucin orthologs have previously been detected in the genomes of *S. frugiperda* and several other lepidopteran species, but this is the first time this protein has been detected in silk. There might be more mucin-like proteins in *G. mellonella* silk, including proteins GmP22 or GmP38 (Table 1B). The main characteristic of mucin proteins is the extended region of the tandemly repeated sequences with serine and/or threonine residues; however, unlike the sericins, they contain a relatively high number of proline residues (Syed et al., 2008). Such ProThrSer copies usually make up at least one third of the total mucin protein length (Perez-Vilar and Hill, 1999). Interestingly, both mucin4-like proteins and Ser1-types of sericins contain two cysteine amino acid residues separated by one non-conserved amino acid residue (i.e., CXC) located in the C-terminal region (Fig. S4). Sericins and mucins have a similar repetitive structure and are produced by homologous labial glands in other insects. Similar features of both protein groups suggest that they may share a common origin.

The products of three different seroin genes were detected in *G. mellonella* transcriptomes, and two or three protein isoforms of *G. mellonella* seroin 1 were found in the cocoon silk. Seroins were shown to form more than 5.5%–12.0% of the cocoon protein mass in *B. mori* (Dong et al., 2013). They were implicated in antimicrobial silk protection and can be produced in lower amounts in body tissues (Korayem et al., 2007).

Our results revealed that sericins and other soluble silk components in *G. mellonella* silk make up almost half the total silk protein mass (48 ± 2%), while soluble silk components in *B. mori* cocoons account for only 27 ± 2%. Earlier studies reported the total content of soluble sericins to be around 30% (Zhang et al., 2015) in *B. mori* cocoons and about 15.7% in *S. ricini* (Prasong et al., 2009). The expansion of soluble silk components in *G. mellonella* might contribute to the compact character of the cocoon, which provides protection against parasitoids and bees. Sericins and other soluble proteins are absent in the convergently evolved spider silks, which typically consist of two different types of spider fibroins (spidroins) (Romer and Scheibel, 2008). Some glue-like spidroin proteins, AgSF2, with repeats containing high amount of proline, glycine, and threonine (55%), were described as products of an aggregate gland in the black widow spider (Collin et al., 2016). Earlier transcriptome sequencing also revealed that some spider silks contain nonspidroin proteins, including peptidase inhibitors, of which the preproteins contain signal peptides (Clarke et al., 2014).

In conclusion, we identified transcripts encoding 27 novel secretory proteins *in G. mellonella* SG. The putative protein products represent soluble silk components, including ten sericins, two or three mucins, two seroins, and several proteins with unknown functions. Lepidopteran mucins were known from previous studies, but this is the first time they were discovered as silk components. The cocoons of *G. mellonella* are more compact and the total content of soluble proteins in silk is twice as high as in *B. mori*. Moreover, the number of different types of sericins is also much higher. Such differences are most likely related to the unique physiology of this insect. Our study deepens the understanding of lepidopteran silk structure, and contributes information relevant to research on antimicrobial peptides and bioinspired glues. This analysis will also improve annotations of silk genes in the Lepidoptera.

Acknowledgements

We thank to Mrs. Jitka Pflegerová for the help with microscopy. This research was supported by European Community's Programme Interreg Bayern Tschechische Republik Ziel ETZ 2014–2020 no. 123 and by the Czech Republic Ministry of Education, Youth and Sports under the Funding Programme INTER-COST no. LTC17073 (Subprogram of Interexcellence).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ibmb.2018.11.003.

References

Acevedo, F.E., Stanley, B.A., Stanley, A., Peiffer, M., Luthe, D.S., Felton, G.W., 2017. Quantitative proteomic analysis of the fall armyworm saliva. Insect Biochem. Mol. Biol. 86, 81–92.

Broz, V., Kucerova, L., Rouhova, L., Fleischmannova, J., Strnad, H., Bryant, P.J., Zurovec, M., 2017. Drosophila imaginal disc growth factor 2 is a trophic factor involved in energy balance, detoxification, and innate immunity. Sci. Rep. 7, 43273.

Clarke, T.H., Garb, J.E., Hayashi, C.Y., Haney, R.A., Lancaster, A.K., Corbett, S., Ayoub, N.A., 2014. Multi-tissue transcriptomics of the black widow spider reveals expansions, co-options, and functional processes of the silk gland gene toolkit. BMC Genomics 15, 365.

Collin, M.A., Clarke 3rd, T.H., Ayoub, N.A., Hayashi, C.Y., 2016. Evidence from multiple species that spider silk glue component ASG2 is a spidroin. Sci. Rep. 6, 21589.

Dong, Y., Dai, F., Ren, Y., Liu, H., Chen, L., Yang, P., Liu, Y., Li, X., Wang, W., Xiang, H., 2015. Comparative transcriptome analyses on silk glands of six silkmoths imply the genetic basis of silk structure and coloration. BMC Genomics 16, 203.

Dong, Z., Zhao, P., Wang, C., Zhang, Y., Chen, J., Wang, X., Lin, Y., Xia, Q., 2013. Comparative proteomics reveal diverse functions and dynamic changes of Bombyx mori silk proteins spun from different development stages. J. Proteome Res. 12, 5213–5222.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.

Elder Jr., J.F., Turner, B.J., 1995. Concerted evolution of repetitive DNA sequences in eukaryotes. Q. Rev. Biol. 70, 297–320.

Geourjon, C., Deleage, G., 1995. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Comput. Appl. Biosci. 11, 681–684.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59, 307–321.

Kludkiewicz, B., Takasu, Y., Fedic, R., Tamura, T., Sehnal, F., Zurovec, M., 2009. Structure and expression of the silk adhesive protein Ser2 in Bombyx mori. Insect Biochem. Mol. Biol. 39, 938–946.

Korayem, A.M., Hauling, T., Lesch, C., Fabbri, M., Lindgren, M., Loseva, O., Schmidt, O., Dushay, M.S., Theopold, U., 2007. Evidence for an immune function of lepidopteran silk proteins. Biochem. Biophys. Res. Commun. 352, 317–322.

Kucerova, L., Broz, V., Arefin, B., Maaroufi, H.O., Hurychova, J., Strnad, H., Zurovec, M., Theopold, U., 2016. The Drosophila chitinase-like protein IDGF3 is involved in protection against nematodes and in wound healing. J Innate Immun 8, 199–210.

Kyte, J., Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105–132.

Lefort, V., Longueville, J.E., Gascuel, O., 2017. SMS: Smart model selection in PhyML. Mol. Biol. Evol. 34, 2422–2424.

Levine, J.D., Sauman, I., Imbalzano, M., Reppert, S.M., Jackson, F.R., 1995. Period protein from the giant silkmoth Antheraea pernyi functions as a circadian clock element in Drosophila melanogaster. Neuron 15, 147–157.

Li, J.Y., Ye, L.P., Che, J.Q., Song, J., You, Z.Y., Yun, K.C., Wang, S.H., Zhong, B.X., 2015. Comparative proteomic analysis of the silkworm middle silk gland reveals the importance of ribosome biogenesis in silk protein production. J Proteomics 126, 109–120.

Nei, M., Gu, X., Sitnikova, T., 1997. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc. Natl. Acad. Sci. U. S. A. 94, 7799–7806.

Nirmala, X., Kodrik, D., Zurovec, M., Sehnal, F., 2001. Insect silk contains both a Kunitz-type and a unique Kazal-type proteinase inhibitor. Eur. J. Biochem. 268, 2064–2073.

Perez-Vilar, J., Hill, R.L., 1999. The structure and assembly of secreted mucins. J. Biol. Chem. 274, 31751–31754.

Prasong, S., Yawalak, S., Wilaiwan, S., 2009. Characteristics of silk fiber with and without sericin component: a comparison between Bombyx mori and Philosamia ricini silks. Pakistan J. Biol. Sci. 12, 872–876.

Romer, L., Scheibel, T., 2008. The elaborate structure of spider silk: structure and function of a natural high performance fiber. Prion 2, 154–161.

Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.

Sehnal, F., 1966. Kritisches studium der bionomie und biometrik der in verschiedenen lebensbedingungen gezuchteten wachsmotte Galleria mellonella L (lepidopera). Z Wiss Zool Abt A 174, 53.

Sehnal, F., Sutherland, T., 2008. Silks produced by insect labial glands. Prion 2, 145–153.

Seong, K.M., Kim, Y.H., Kwon, D.H., Lee, S.H., 2012. Identification and characterization of three cholinesterases from the common bed bug, Cimex lectularius. Insect Mol. Biol. 21, 149–159.

Syed, Z.A., Hard, T., Uv, A., van Dijk-Hard, I.F., 2008. A potential role for Drosophila mucins in development and physiology. PloS One 3, e3041.

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.

Takiya, S., Tsubota, T., Kimoto, M., 2016, 19. Regulation of silk genes by hox and homeodomain proteins in the terminal differentiated silk gland of the silkworm Bombyx mori. J. Dev. Biol. 4.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30, 2725–2729.

Tamura, T., Inoue, H., Suzuki, Y., 1987. The fibroin genes of Antheraea-yamamai and Bombyx-mori are different in their core regions but reveal a striking sequence similarity in their 5' ends and 5' flanking regions. Mol. Gen. Genet. 207, 189–195.

Teramoto, H., Miyazawa, M., 2005. Molecular orientation behavior of silk sericin film as revealed by ATR infrared spectroscopy. Biomacromolecules 6, 2049–2057.

Tsubota, T., Yamamoto, K., Mita, K., Sezutsu, H., 2015. Gene expression analysis in the larval silk gland of the eri silkworm Samia ricini. Insect Sci. 6, 791–804.

Wang, Y.J., Zhang, Y.Q., 2011. Three-layered sericins around the silk fibroin fiber from Bombyx mori cocoon and their amino acid composition. Adv Mater Res-Switz 175–176, 158–163.

Wang, Z., Zhang, Y., Zhang, J., Huang, L., Liu, J., Li, Y., Zhang, G., Kundu, S.C., Wang, L., 2014. Exploring natural silk protein sericin for regenerative medicine: an injectable, photoluminescent, cell-adhesive 3D hydrogel. Sci. Rep. 4, 7064.

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

Yukuhiro, K., Sezutsu, H., Tsubota, K., Takasu, Y., Kameda, T., Yonemura, N., 2016. Insect silks and cocoons: structural and molecular aspects. In: Cohen, E., Moussian, B. (Eds.), Extracellular Composite Matrices in Arthropods, 1 st ed. Springer, pp. 515–555.

Zhang, Y., Zhao, P., Dong, Z., Wang, D., Guo, P., Guo, X., Song, Q., Zhang, W., Xia, Q., 2015. Comparative proteome analysis of multi-layer cocoon of the silkworm, Bombyx mori. PloS One 10, e0123403.

Zurovec, M., Kludkiewicz, B., Fedic, R., Sulitkova, J., Mach, V., Kucerova, L., Sehnal, F., 2013. Functional conservation and structural diversification of silk sericins in two moth species. Biomacromolecules 14, 1859–1866.

Zurovec, M., Kodrik, D., Yang, C., Sehnal, F., Scheller, K., 1998a. The P25 component of Galleria silk. Mol. Gen. Genet. 257, 264–270.

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.

Zurovec, M., Yang, C., Kodrik, D., Sehnal, F., 1998b. Identification of a novel type of silk protein and regulation of its expression. J. Biol. Chem. 273, 15423–15428.

Zurovec, M., Yonemura, N., Kludkiewicz, B., Sehnal, F., Kodrik, D., Vieira, L.C., Kucerova, L., Strnad, H., Konik, P., Sehadova, H., 2016. Sericin composition in the silk of Antheraea yamamai. Biomacromolecules 17, 1776–1787.