

Identification of four small molecular mass proteins in the silk of *Bombyx mori*

X. Nirmala,¹ K. Mita,² V. Vanisree,¹ M. Žurovec¹ and F. Sehnal¹

¹Institute of Entomology, Academy of Sciences, and the Faculty of Biological Sciences, University of South Bohemia, Branišovská, České Budějovice, Czech Republic; ²Genome Research Group, National Institute of Radiological Science, Anagawa, Inage-ku, Chiba-shi, Japan

Abstract

This paper describes cDNAs of four small-size proteins that occur in the cocoon silk of *Bombyx mori*. Two of them (9.9 and 10.3 kDa), which have the N-terminal sequences and the spacing of a few amino acids at C-termini similar to the seroin of *Galleria mellonella*, are called seroin 1 and seroin 2. The corresponding genes are expressed in the middle, and to a small extent also in the posterior silk gland sections. The seroin 1 and less conspicuously the seroin 2 mRNAs accumulate in the course of the last larval instar to a maximum in postspinning larvae. Two other proteins (6 kDa and 4.7 kDa) of *B. mori* cocoons were identified as a typical Kunitz-type and a somewhat unusual Kazal-type proteinase inhibitors, and named BmSPI 1 and BmSPI 2, respectively. Their genes are expressed in the middle, and the BmSPI 1 gene slightly also in the posterior silk gland sections. The expression ensues a few days after the last larval ecdysis and increases until the cocoon spinning. Post-spinning larvae still contain high amounts of the BmSPI 1 but no BmSPI 2 transcripts. It is assumed that seroins and proteinase inhibitors are involved in cocoon protection against predators and microbes.

Keywords: silk, seroin, proteinase inhibitor, protease inhibitor, antimicrobial agents, Lepidoptera.

Introduction

Fully grown larvae of some insects spin silky cocoons in which they pupate (Sehnal & Akai, 1990). Cocoons of a few

lepidopteran species are processed to obtain silk fibre that is woven into textiles. The composition of silk has been examined in detail in the major commercial silk producer, the domestic silkworm, *Bombyx mori* (suprafamily Bombycoidea). Spun silk filament consists of a firm, elastic core that is enveloped by sticky materials gluing the filaments together into the cocoon. The insoluble proteins, which make up the filament core and are processed to the textile fibre, are produced in the posterior section of the paired silk-producing glands. These proteins include heavy-chain and light-chain fibroins, which are linked together by disulphide bonds, and P25 glycoprotein that plays a role in assembling six fibroin dimers into an elementary silk unit (Inoue *et al.*, 2000). At least six sericin proteins derived from two genes (Michaille *et al.*, 1986) are produced in the middle section of the glands and provide the gluing substance enveloping the fibre core. They are dissolved in hot alkaline water during silk reeling.

The structure and expression patterns of major silk genes have also been elucidated in the waxmoth, *Galleria mellonella* (suprafamily Pyraloidea), a lepidopteran distantly related to *B. mori*. It was found that the overall silk composition and the silk gene structure are similar as in *B. mori* (Žurovec *et al.*, 1992). Amino acid sequence homologies between the two species reach about 50% in case of light-chain fibroin and P-25 (Žurovec *et al.*, 1995, 1998a), but are small and restricted to the peptide termini in the heavy-chain fibroin and sericins. These high molecular proteins are for the most part composed of regular amino acid repeats that appear to be species specific (M. Žurovec *et al.*, unpublished data).

Analysis of the dissolved cocoons of *B. mori* and *G. mellonella* revealed presence of components smaller than the structural silk proteins (Grzelak *et al.*, 1988; Kodrík, 1992). Three of the small molecular mass proteins were characterized in *G. mellonella*. Proteins of 22.5 and 23 kDa were identified as glycosylated products of a 16 kDa peptide, which was called seroin to emphasize its production in both the middle (sericin source) and the posterior (fibroin source) sections of the silk gland (Žurovec *et al.*, 1998b). Peptides of 6 kDa and 4 kDa, secreted exclusively in the middle silk gland section, were recognized as proteinase inhibitors of the Kunitz and Kazal families, respectively (Nirmala *et al.*, 2001).

Received 6 March 2001; accepted after revision 4 May 2001. Correspondence: Dr František Sehnal, Institute of Entomology AV CR, Branišovská 31, 370 05 České Budějovice, Czech Republic. Tel.: + 420 38 5300350; fax: + 420 38 5300354; e-mail: Sehnal@entu.cas.cz

There were indications that the small molecular mass components of *B. mori* silk are homologous to those of *G. mellonella*. Specifically, two proteinase inhibitors were found in *B. mori* cocoon extracts (Kurioka & Hirano, 1995; Kurioka *et al.*, 1999), and the N-terminal sequencing of another two cocoon proteins disclosed similarity to the N-terminus of *G. mellonella* seroin (Žurovec *et al.*, 1998b). However, the mass of the presumed Kazal-type inhibitor in *B. mori* silk was assessed as 11 kDa (Kurioka & Hirano, 1995), and the proteins exhibiting seroin homology were only 8 kDa and 13 kDa (Žurovec *et al.*, 1998b). The size discrepancies between *G. mellonella* and *B. mori* silk proteins contradicted the assumption of their mutual homologies and no conclusive statement could be made without additional information. We decided to identify the cDNAs and to examine the expression profiles of selected small molecular mass silk proteins in *B. mori*.

Results

Identification of two seroin cDNA sequences

Analysis of *B. mori* silk proteins disclosed two putative homologues of *G. mellonella* seroin: an 8 kDa peptide with N-terminus GFVXEDDNFPG and a 13 kDa peptide with similar N-terminus GFVXEDDDDLF (Žurovec *et al.*, 1998b). Screening of our EST database for these two partial sequences led to identification of two types of cDNA clones. Putative translation products of six clones included sequence GFVWEDDDDLF that matched the result of amino acid sequencing of the 13 kDa peptide and identified the uncertain residue as tryptophan. Two other cDNA clones in the EST database encoded sequence GFVWQDDNFPG, which was identical with the N-terminus of the 8 kDa peptide except that glutamine occurred in place of glutamate; the unknown residue turned out to be tryptophan. Differences in the N-termini and especially in the remaining amino acid sequences (see below) suggested that the two types of cDNAs were derived from two different genes. That with higher representation in the EST library was designated *seroin 1*, and the other *seroin 2*.

Characterization of seroin 1

Out of the six cDNA clones encoding the N-terminal sequence of seroin 1, those labelled msgV0150 and msgV0105 contained identical 5' UTR of 32 nt, an apparently complete ORF, and a truncated 3' tail (very short in msgV0150). The sequences were numbered from the first 5' nucleotide. All remaining *seroin 1* clones, msgV0001, msgV0243, msgV0403 and msgV0653, were 5' truncated and started between positions 57 and 73. The overlapping parts of all six sequences were identical except for position 128 where T was found in three, and C in the other three clones. This alternation in the third place of a codon did not alter its coding for Phe. To verify the 5' end and to complete

the 3' end of the cDNA, the clones msgV0105 and msgV0150 were isolated and sequenced independently of the EST database. Small discrepancies were resolved by repeated sequencing.

The complete cDNA sequence of *seroin 1* is 769 nt long (Fig. 1, top). The 5' end UTR includes 32 nt, ORF encompasses 327 nt, and the 3' UTR with two putative polyadenylation signals AATAAA extends for 398 nt prior to poly (A) tail. The ORF is initiated by the first ATG codon and encodes a preprotein of 108 amino acid residues, of which eighteen make up the signal peptide. Secreted seroin 1 begins with Gly¹⁹, is composed of ninety amino acids, and has a deduced molecular mass 9912.6 Da.

Characterization of seroin 2

The EST database sequences msgV0256 and msgV0508 encoded the N-terminus of seroin 2. Both sequences included an apparently complete ORF, which was preceded by 5' UTR of 74 nt in msgV0256 and of 56 nt in msgV0508. The clones msgV0256 and msgV0508 were isolated, subcloned, and analysed in detail. A primer 5' CAA GTC CTA GAA GAT AGA AGA 3', which corresponded to 403–424 nt of the cDNA, was synthesized and employed to sequence the middle portion of the cDNA. Sequences of the two clones proved to be identical.

The complete cDNA of *seroin 2* is 1413 nt long and includes 74 nt 5' UTR, 339 nt ORF, and 1000 nt 3' UTR, which contains two putative polyadenylation signals and a poly(A) terminal sequence (Fig. 1, bottom). The deduced translation product, starting from the first ATG codon, comprises 112 amino acids. Peptide sequencing revealed that the N-terminus of mature seroin 2 begins with Gly¹⁹, therefore the preprotein apparently consists of eighteen amino acid signal peptide followed by secreted peptide of 94 amino acids (10312.6 Da). Amino acid residues around positions 18–19 of the deduced translation product comply with the rules for signal peptide cleavage (Von Heijne, 1983).

Silk proteinase inhibitors

Using approaches specified below, we identified cDNAs of two proteinase inhibitors that were called BmSPI 1 and BmSPI 2 to emphasize their homologies with the corresponding inhibitors of *G. mellonella* (Nirmala *et al.*, 2001) and to indicate that *B. mori* was the source species. BmSPI 1 denotes a Kunitz-type inhibitor analysed by Kurioka *et al.* (1999), while BmSPI 2 is a Kazal-type inhibitor that was isolated and for the most part sequenced by Kurioka & Hirano (1995). The latter was independently identified also by our N-terminal peptide sequencing.

Kunitz type inhibitor BmSPI 1

No cDNA encoding BmSPI 1 (Kurioka *et al.*, 1999) was found in our EST database, but RT PCR yielded a product of about 200 bp, and subsequent 5' RACE led to a 230 bp

Seroin 1

[illegible]

Seroin 2

GTGTCCTACAGTTGGTTTCAACAGTGCTACACAGCTCGATTTTTTTATTGTTGAAGCCTTTTAACCGCCTGATA	ATGCGC	8
	MetAla	
TTCAAGCAAGTTCTTTTTATGTGTGCTAATAACCATCTCGGAGCGCTGGCTTGTGTTGGCAGCAGCATTA	ACTTTCGGGA	16
PheThrLysPheLeuPheMetLeuSerLeuIleThrIleAlaSerAlaGlyPheValTrpGlnAspAspAsnPheProGly		2
TTCCCAACGCGATATGTGGCCCTCTATACAGATCCCAACGATTCGCGCATTGATGCCAAAATTCGGA	ACTTTGCTTTTCG	24
PheProSerAspMetTrpProSerIleGlnIleProThrIleProProPheAspProLysIleProAsn	PheAlaPheSer	5
TTCCCTCTCTCTGACAAACATTAAGAAAACTAAACCAACCCGGCAAAACCTATAGCGGGCTCTAC	GTGAGCAGCAGTATGGA	32
PheProSerProAspAsnIleLysLysLysLysProGlnProGlyGlnThrLysGlyValTrpValSer	ValSerAsnGly	8
GGGAAGGGAACTATGGTTGCGAAACATCAATGGGGAAGTATCGAAAAAAGTTTGAGAGAGGAT	TCCAAAAATCTAAGAG	40
GlyLysGlyThrMetValAlaAsnIleAsnGlyGluValIleGluLysLysPheGlyGluAspSerLysLys	SerLysSer	11
----->		
AAGTCCTAGAAGATAGAAGATTTAAATTAAGTAACCGTCATTCTGACACCCACATTTTCGATTTAAT	TTTCATTTAAAAATA	48
LysSer		11
TTTTTATGTTAACAAATGAATACATTCATTATATAAACAATTTTAAGATAATTTAAATGTTATT	GATATAGTTTAAAT	56
GTGTTTTCCTGCTCCTCTCTTTTACGAGTGAATACATACACATAACAAACCCACCAATTAATTAAT	TTTATTTTATTT	64
TATCAACTTCAGGCTCTCTTCAAGATTTATTTATAAAATCAAATGTTTTCCTCCTGCAGCTAC	CGCGTAAAAATTAAC	72
GTAATCTTTTGATATTTATAAAACTGAATTCATGATTTTCATTAAGATACAATCTAAAGACT	TGAAAGCGATTTGAAT	80
GTGAACCAATTTAAACCTAATTTACAGCAAGAATCTTTTTCATAAACAATAATTTATTTT	TTCAGCAGTAGTACTAAC	88
CTTTGATCTGTGTTATGTAGATTGCCACAGGAATTCATAAATTTAACTGCTAGGTTTCATCT	GAGCCTCTCTCTAGCT	96
TGATTTCAATATCTACCAATATTTAGATATGAACAATGGTGACACAAACACAGCAATCAAGAT	CTGCTTCCATCATGG	104
CGATTCTGATGCTACTACGCCCGGTGAATTTAAAGTTTATTGATCGAGGAAAAGGTCCTAAT	GATGTACTACTATTAT	112
CAGTTCTCCTCACTAATGTATGGCCATCTCGAATAAAGCGTGATGCAATCTCGGAAAGTGC	GGGTACTATTTTGGATT	120
CTTTTAAACAGTCGGATTTTCGGCGGTATTTTCATTTGTAATCTTATATCTCTTTTATTT	TATGTCTTCAAAAATTAAG	128
AAAATTTAATGTTTATCATTTCAAATTCCAACGATTTTCATGTATGTAATTAAGTTC	CAAAAAGGTTTATTTG	136
ACATTTAACTAAATTTGATACATACCGGAAAAAATAAAAAA		141

Figure 1. Complete cDNA sequences and deduced amino acid sequences of seroin 1 (top) and seroin 2 (bottom). The initiation and termination codons are printed in bold and the putative polyadenylation signals are underlined. The double-underlined C nucleotide in *seroin 1* was identified in three cDNA clones, while another three clones contained T in this position. Horizontal arrow in *seroin 2* cDNA indicates position of the internal sequencing primer. Amino acid residues identified by N-terminal sequencing of isolated silk proteins are marked with asterisks, and the putative signal peptide cleavage sites with ↑.

DNA fragment, which was cloned and sequenced. The complete cDNA sequence of *BmSPI 1* is 394 nt long (Fig. 2, top). The 5' UTR consists of 25 nt, ORF encompasses 231 nt, and the 3' UTR of 138 nt includes two adjacent and partly overlapping polyadenylation signals and a terminal poly(A) tail. ORF is initiated by the first ATG codon and encodes a preprotein of seventy-six amino acid residues. Signal peptide cleavage, predicted to occur between Ala²¹ and Asn²² within a typical motif for the recognition by signal peptidases, apparently yields a signal peptide of twenty-one amino acids and the mature BmSPI 1 of fifty-five amino acids and 6000.75 Da mass. The deduced sequence of BmSPI 1 represents a typical single-domain serine proteinase inhibitor of the Kunitz type.

Kazal type inhibitor BmSPI 2

Clones msgV0972 and msgV0973 of the EST database encoded amino acid sequence TCICTEYRPVC that had been identified by our N-terminal sequencing of a 5 kDa silk peptide and corresponded to the Kazal-type inhibitor described by Kurioka & Hirano (1995). Both cDNA clones included an apparently complete ORF and were identical

except that the former lacked the poly(A) tail. Repeated sequencing of the two clones identified the complete 5' and 3' ends. The clones were identical save position 329 (3' UTR) where T was identified in one and G in the other clone.

The *BmSPI 2* cDNA is 411 nt long and includes 14 nt 5' UTR, 198 nt ORF, and 199 nt 3' UTR with three overlapping polyadenylation signals and a terminal poly(A) sequence (Fig. 2, bottom). The deduced translation product starts from the first ATG codon and consists of sixty-five amino acid residues. The mature BmSPI 2 obviously begins with Thr²² that was identified as the most N-terminal amino acid by Kurioka & Hirano (1995) and also by us. However, the most likely signal peptide cleavage site (Von Heijne, 1983) lies between Gly¹⁸ and Ala¹⁹ or between Ala¹⁹ and Glu²⁰, indicating that the preprotein is possibly further processed by an aminopeptidase. The secreted BmSPI 2 comprises forty-four amino acids and has a mass of 4767.56 Da. The first twenty-five of the deduced amino acid residues are identical with those found by direct peptide sequencing (Kurioka & Hirano, 1995). BmSPI 2 is clearly a single-domain proteinase inhibitor of the Kazal type.

BmSPI 1

```

AAGTGAATCTTCATTCTTCTGAGAATGTTGAAATACACGAGCATTTCGTTTTTATTAATAATCTTTATTTCTTCACT 81
MetLeuLysTyrThrSerIleSerPheLeuLeuIleIleLeuLeuPheSerPheThr 19
----->
AACGCTAATCCTGATTGTTTGTGTCCTAATAAGACGGGCCCTTGCAAGGGCAGTTTCCAGATATGCCTACGATTTCATCG 162
AsnAlaAsnProAspCysLeuLeuProIleLysThrGlyProCysLysGlySerPheProArgTyrAlaTyrAspSerSer 46
      ↑
GAGGACAAATCGCTTGAATTCATCTACGGAGGTTGCCAAGCGAATGCTAATAATTTGAGACTATAGAAGAGTGCGAAGCC 242
GluAspLysCysValGluPheIleTyrGlyGlyCysGlnAlaAsnAlaAsnAsnPheGluThrIleGluGluCysGluAla 73
<-----
GCCTGCCTTTAAAATACGTGGGAGTGTTCATGTCTCTCGTAGGAATGCGCTAGTGTGTGACCTAATAAATTATGGTAAT 324
AlaCysLeu . 76
GTTGGTTTATGGTGGTAAATTTTGAATTTATGAAATAAAATAAATGTGATAAGTAAAAA 394

```

BmSPI 2

```

CAAAATCGAAACAAATGAAGACTTCAATCGTTCTAATTTTCTCCTCGTCGCTGCACTTTGGGGGCTGAAAGTACC 80
MetLysThrSerIleValLeuIlePheLeuLeuValAlaCysCysThrLeuGlyAlaGluSerThr 22
      ↑ ↑
TGTATTTCACACCACGAATATAGCCTGTTTGGTACTAACGGTGTACTTACGGCAACAGGTGCCAACTGAGATGTGCG 161
CysIleCysThrThrGluTyrArgProValCysGlyThrAsnGlyValThrTyrGlyAsnArgCysGlnLeuArgCysAla 49
      * * * * *
AAAGCCATATTGCCTACGACGGCCGTGCTCGGTGGCATGCGAATTTAGGGCTGCCAGAGCTCTCGTATTCCCGGGAC 242
LysAlaIlePheAlaTyrAspGlyProCysCysGlyGlyMetArgIle . 66
GTCCCGTATTTAGGCTTAGTTTAAATGTCTCGCGGGGACTGACTCTGTTCCTGATTTCGAATGTTTCGATTACAATTGTT 323
TTGGTGTTTATAAATATGTTAAACTAAGTCTCGGAGAAATATCAAATAATAATAATAAAGAATGTCACATATAA 404
AAAAA 411

```

Figure 2. The cDNA and deduced peptide sequences of BmSPI 1 (top) and BmSPI 2 (bottom). The initiation and termination codons are printed in bold and the putative polyadenylation signals are underlined. Horizontal arrows in *BmSPI 1* cDNA indicate positions of one degenerate forward and two gene-specific reverse primers that were used in cDNA identification by PCR. The double-underlined nucleotide in *BmSPI 2* was identified as T in one, and as G in the other cDNA clone. The most probable signal peptide cleavage sites (Von Heijne, 1983) are marked with ↑. Amino acid sequence of BmSPI 1 reported by Kurioka *et al.* (1999) contained Glu in place of Asp. BmSPI 2 residues, which were identified by N-terminal sequencing of a 5 kDa silk protein, are marked with asterisks.

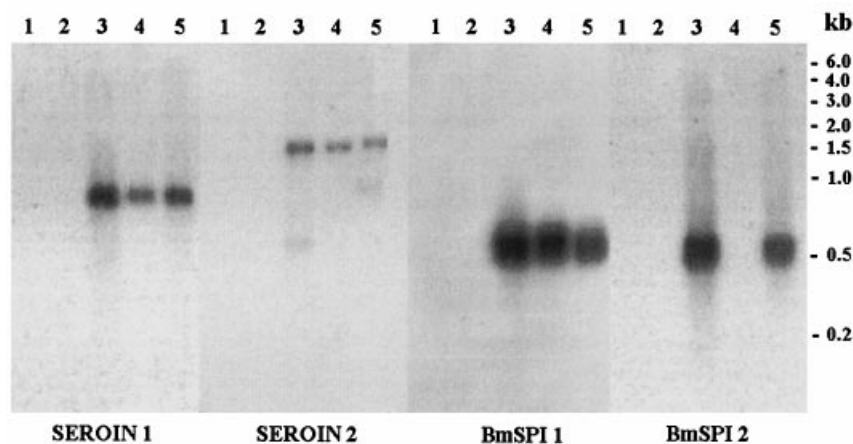


Figure 3. Expression of the *seroin 1*, *seroin 2*, *BmSPI 1* and *BmSPI 2* genes, respectively, in selected tissues. Ten micrograms total RNA from the integument (lane 1), fat body (lane 2), middle (lane 3) and posterior silk gland sections (lane 4), and 5 µg total RNA from the middle silk gland section (lane 5), were hybridized to the respective radiolabelled cDNA. Positions of standards in an RNA ladder are given on the right.

The expression pattern of seroins and proteinase inhibitors

Samples of total RNA from the fat body, integument, and the middle and posterior parts of silk glands of day 5 last instar larvae were probed with the radiolabelled cDNAs. Transcripts hybridizing with the *seroin 1*, *seroin 2* and *BmSPI 1* probes were found in the middle and, in much lower amounts, in the posterior parts of silk glands, whereas putative *BmSPI 2* mRNA occurred only in the middle part (Fig. 3). All examined genes were represented by single transcripts with approximate sizes 800 bp and 1500 bp for *seroin 1* and *seroin 2*, respectively, and 500 bp for either inhibitor. No hybridization signals were detected in tissues other than silk glands.

Developmental expression profiles of the genes in question were followed from the end of the penultimate larval instar until the end of cocoon spinning in the last larval instar (Fig. 4). The results confirmed that expression of both *seroin*

genes was intensive in the middle silk gland section. *Seroin 1* transcript was hardly detectable in the penultimate instar larvae, appeared in increasing amounts in the feeding last instar larvae, and accumulated dramatically between wandering (postfeeding larvae searching for pupation sites) and the end of cocoon spinning. By contrast, the expression of *seroin 2* gene increased slowly and rather gradually throughout the examined period. In the posterior silk gland section, the transcripts of both *seroin* genes were present in small amounts and increased little in the course of development.

The transcript of *BmSPI 1* was not detectable in the penultimate instar and at the beginning of the last larval instar. Its amount increased in later development, reaching a maximum in the middle silk gland section just before and during cocoon spinning (Fig. 4). The transcript was also present in the posterior silk gland section but its accumulation was not conspicuous. *BmSPI 2* was not expressed in the penultimate and at the beginning of the last instars, and

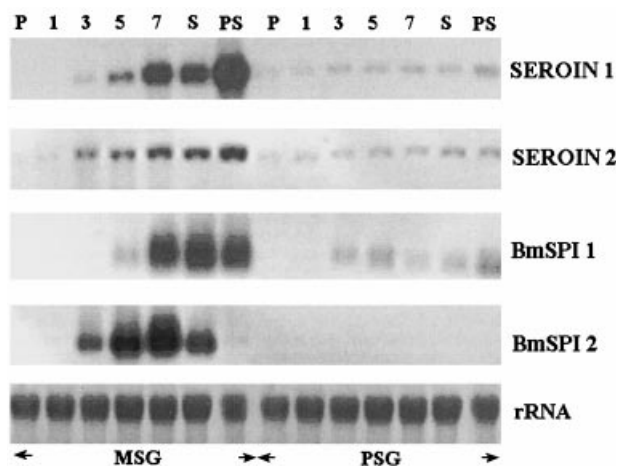


Figure 4. Developmental profiles of *seroin 1*, *seroin 2*, *BmSPI 1* and *BmSPI 2* mRNAs during the penultimate and last larval instars. Total RNA (5 µg) from the middle (MSG) and posterior (PSG) silk gland sections of the fully grown penultimate instar larvae (lane P), and of the last instar larvae of days 1, 3, 5 and 7, at spinning (S), and after spinning (PS), was blotted and probed with the respective radiolabelled cDNA. The bottom panel shows total RNA, in nylon membrane, stained with methylene blue.

its subsequent expression was limited to the middle silk gland region. The content of *BmSPI 2* mRNA grew rapidly in the feeding last instar larvae to a maximum in the fully grown, pre-spinning larvae. Spinning was accompanied by abrupt transcript decline that was no longer detectable in the post-spinning larvae (in contrast to the mRNAs of the *seroins* and *BmSPI 1*).

Discussion

The seroins

Data available a few years ago seemed to imply that lepidopteran silk consisted of just a handful of the fibre and glue proteins (Hui & Suzuki, 1995). The discovery of seroin in *G. mellonella* proved the existence of additional silk-specific proteins (Žurovec *et al.*, 1998b). The detection of two proteins with seroin-like N-termini in the silk of *B. mori* suggested that seroin homologues are regular silk components. Full structure of these proteins in *B. mori* silk has now been deduced from the cDNAs that are derived from

two separate genes, *seroin 1* and *seroin 2*. The genes are transcribed in the middle, and to small extent also in the posterior silk gland sections, similar to the *seroin* gene of *G. mellonella*. A shared feature of *seroin* transcripts in *B. mori* and *G. mellonella* is a relatively long 3' UTR with two AATAAA polyadenylation signals.

Putative translation products of *B. mori seroin* genes have similar signal peptides (fifteen positions out of eighteen residues are identical) resembling the signal peptide of *G. mellonella seroin* (Fig. 5). Signal peptides of the structural silk proteins are different (data not shown). The secreted seroins of *B. mori* exhibit homologies with the seroin of *G. mellonella* at the N-termini (six to seven out of the eight terminal amino acids are identical). Similarities in the relative positions of a few amino acid residues can be found close to the C-termini but major parts of the three seroins are dissimilar, indicating rapid gene diversification. The alignment pattern suggests that the *seroin* genes of *B. mori* originated by gene duplication followed by extensive deletions and point mutations (Fig. 5). Specifically, *Galleria seroin* contains two copies of a PPLPQPPL repeat that is missing in *B. mori seroins*. Another proline-rich region with three copies of a PPI motif in *G. mellonella seroin* is modified in *B. mori seroin 2* and cannot be recognized in seroin 1. In spite of these dissimilarities, the homologies at N-terminus, comparable short motifs at the C-terminus, analogous general peptide properties (see below), and identical tissue and developmental expression profiles justify our suggestion that all peptides in question belong to one seroin silk protein family.

All seroins possess a balanced ratio of hydrophobic and polar residues. C-termini include stretches of hydrophilic residues, such as ESKSSYSSSSST in seroin 1 and EDSKSKSKSKS in seroin 2 of *B. mori*, and EESRKESSSN in *G. mellonella seroin*. Basic and acidic residues are in equal proportions in seroin 2, but a prevalence of acidic residues in seroin 1 and in *G. mellonella seroin* lowers their pI values to 4.047 and 4.93, respectively. Glycosylation of *G. mellonella seroin* apparently accounts for the molecular weight increase from 16.3 kDa of the deduced peptide to the 22.5 and 23 kDa proteins present in the silk (Žurovec *et al.*, 1998b). The deduced masses of seroin 1 and seroin

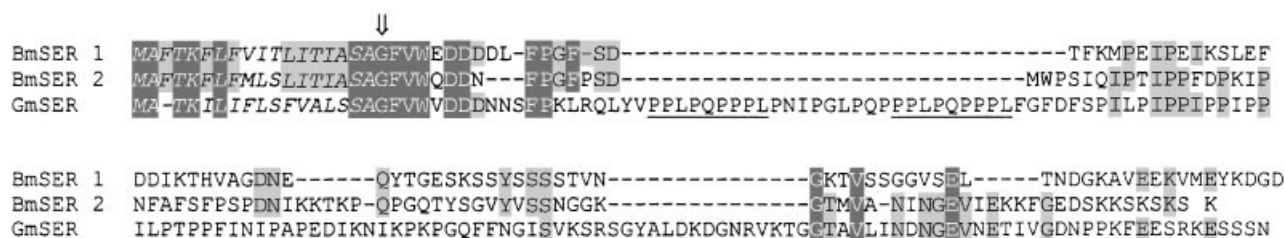


Figure 5. Tentative amino acid sequence alignment of *B. mori seroin 1* (BmSer 1) and seroin 2 (BmSer 2) with *G. mellonella seroin* (GmSer; Žurovec *et al.*, 1998b). The signal peptide is italicized and the first residue of the secreted peptides is marked with ↓. Amino acid residues conserved in all three proteins are shaded black and those conserved in two proteins are shaded grey. Two proline-rich motifs present only in GmSer are underlined.

BmSPI 1

		↓	
BmSPI 1	NPDCLLPK-KTGPCKGSFPRYAYDSSSEDKVEFIYGGCQANANNFETIEECEAAAL		% Overlap
GmSPI 1	DDDIQSLPL-KTGPCKRAAFQRYGY-VEGKGVLFYGGCQGNANNFETLEACKNAEN		55
<i>M. sexta</i> B	EDICSLPP-EVGPCKRAGFLKFAYSELNCKLFYGGCQGNENNFFETLQACQA??		49
<i>S. bullata</i>	VDKSAQLQPK-EVGPCKRSDVFVFYNADTKACEEFLYGGCGRGNDNRFTKEECKEL		43
SCI (1)	DKPTTKPIEQAFGNSGPKFAYIKLYSNQTKKCEEFIYGGCQGNDRFITLAECEQKCIK		40
BPTI	RPDFCLEPP-YTGPCKARIIRYFYNAGKGLQTFVYGGCRAKRNNFKSAEDCMRTGGAIGPWENL		37

BmSPI 2

		↓	
BmSPI 2	T-----G-----ICTTEYRPVCGTNGVTYGNRCQL--RCAKAIKAY-----DGPCCGGMRI		% Overlap
GmSPI 2	VCTTEWDPVCGKDGKTYSNLCWLNEA-----GVGLDHEGEC		44
<i>S. littor.</i>	LP-----PC-----TCTRNYPICGSGNETYSNQCLLD--CARYNTHQ-DIT??		41
OMPP (3)	VS----VDGSEYPQPTCTTEHRPVCSSNNETYGKNCN---CNAVVKSNGLTLVSHFGKC		35
BPSTI	NILGREAKCTNEV-NGCPRIYNPVCCTDGVTSNECLL---CMENLERQTPVLIQKSGPC		33

Figure 6. Amino acid sequence alignment and percentage overlap of BmSPI 1 and BmSPI 2 with the related members of the respective proteinase inhibitor family. The Kunitz family (BmSPI 1, top) is represented by GmSPI 1 (Nirmala *et al.*, 2001), trypsin inhibitors from the haemolymph of *Manduca sexta* (Ramesh *et al.*, 1988) and *Sarcophaga bullata* (Sugumaran *et al.*, 1985), *B. mori* haemolymph chymotrypsin inhibitor, SCI-I (Sasaki, 1978), and bovine pancreatic trypsin inhibitor, BPTI (Laskowski & Kato, 1980). The Kazal inhibitors (BmSPI 2, bottom) include GmSPI 2 (Nirmala *et al.*, 2001), inhibitor from the haemolymph of *Spodoptera littoralis* (Boigegrain *et al.*, 2000), third domain of egg white ovomucoid from peacock pheasant, OMPP (Laskowski *et al.*, 1987), and the bovine pancreatic secretory trypsin inhibitor, BPSTI (Green & Bartelt, 1969). Conserved Cys residues are shaded black and other conserved amino acid residues are shaded grey. Putative P1 residue is marked with ↓ and question marks indicate that the C-terminal sequence is incomplete (in this case, homology was established between the known sequence and the corresponding region of BmSPI 1 or BmSPI 2).

2 are 9912.6 Da and 10312.6 Da, whereas corresponding proteins extracted from the *B. mori* silk migrated in polyacrylamide gel electrophoresis as 13 kDa and 8 kDa moieties, respectively. We assume that this discrepancy is due to inaccurate size estimation from the electrophoretic mobility.

The silk proteinase inhibitors

The proteins inhibiting serine proteinases are typified by conserved spacing of 6 Cys and several other residues (Bode & Huber, 1992). Those identified in *G. mellonella* silk belong to the Kunitz and Kazal types (Nirmala *et al.*, 2001). They were originally designated SPI 1 and SPI 2, respectively, but these designations are changed in the present paper to GmSPI 1 and GmSPI 2 to indicate the source species, and appellations BmSPI 1 and BmSPI 2 are used for the homologous inhibitors of *B. mori*. Each inhibitor is derived from a separate gene. A notable feature of the *BmSPI 1* and *BmSPI 2* transcripts is the presence of overlapping double or triple polyadenylation signals (Fig. 2). Three dispersed polyadenylation signals occur also in the *GmSPI 1* transcript (Nirmala *et al.*, 2001).

The Kunitz-type inhibitor BmSPI 1 was isolated and its amino acid sequence ascertained by Kurioka *et al.* (1999). The protein was called cocoon specific trypsin inhibitor (CSTI) to stress that it had been isolated from the cocoons. Our deduced amino acid sequence of BmSPI 1 (Fig. 2, top) is identical to that reported for CSTI, except that Asp instead of Glu is present in position 27. The change is due to a single point mutation in the third position of a codon, and probably represents an allelic variation that does not affect the function. The BmSPI 1 deduced size 6000.75 Da is consistent with the CSTI mass determination 6026.7 Da by Kurioka *et al.* (1999).

BmSPI 1 is 55% identical with the corresponding inhibitor of *G. mellonella*, GmSPI 1 (Fig. 6, top). Among the silk components, a similarly high homology of 53.7% between *B. mori* and *G. mellonella* was found only in the mature P25 protein (Žurovec *et al.*, 1998a). BmSPI 1 and GmSPI 1 are related to the Kunitz-type inhibitors present in insect haemolymph. Such an inhibitor from the moth *Manduca sexta* (Ramesh *et al.*, 1988) exhibits 49% and 44%, that from the fly *Sarcophaga bullata* (Sugumaran *et al.*, 1985) 43% and 38%, and that from *B. mori* (Sasaki, 1978) 40% and 39% homology to BmSPI 1 and GmSPI 1, respectively. The BmSPI 1, however, differs from all Kunitz-type insect proteinase inhibitors by having Lys in the so-called P1 position after the second Cys residue. In this respect, BmSPI 1 resembles the bovine pancreatic trypsin inhibitor, BPTI (Laskowski & Kato, 1980). Amino acid residue in the P1 position confers specificity to the target proteinases, with Lys specifying activity against the trypsin. Inhibition of trypsin by BmSPI 1 has been demonstrated by Kurioka *et al.* (1999).

Expression studies revealed presence of *BmSPI 1* mRNA both in the middle and the posterior silk gland sections, whereas Kurioka *et al.* (1999) detected trypsin-inhibitory activity only in the middle section. This could imply that the transcript produced in the posterior section is not translated into the active inhibitor but we rather believe that the failure to detect trypsin inhibition was due to the low level of gene expression in this part of silk glands (cf. Fig. 4).

The N-terminal sequence TCICTTEYRPVC that we found in a 5 kDa silk protein, matches exactly the deduced N-terminus of BmSPI 2 after the signal peptide cleavage. Kurioka & Hirano (1995) identified twenty-eight amino acid residues of BmSPI 2. The first twenty-five are identical to

our deduced amino acid sequence and only the terminal Cys-Met-Ala are in our sequence replaced with Arg-Cys-Ala-Lys (Fig. 2, bottom). The deduced mass 4767.66 Da of BmSPI 2 is in accord with our electrophoretic assessment 5 kDa for the extracted silk protein. Kurioka & Hirano (1995) originally reported an 11 kDa size but Kurioka *et al.* (1999) later referred to a 4 kDa protein, which is close to our mass calculation.

BmSPI 2 includes a complete Kazal domain, in contrast to GmSPI 2 which is truncated at the N-terminus and lacks the 1st and 5th cysteines (Fig. 6, bottom). However, the structure of BmSPI 2 is also somewhat unusual by having CysV and CysVI only ten amino acid residues apart, while in typical Kazal inhibitors these cysteines are separated by seventeen residues. CysVI is usually the last residue in single domain Kazal inhibitors but in BmSPI 2 it is followed by five more amino acids. The overall homology between BmSPI 2 and GmSPI 2 amounts to 44% and is not much higher than the homology to the Kazal inhibitors from other insect tissues. Thr in P1 position, which is the second residue after Cys in the Kazal type inhibitors (Bode & Huber, 1992), is a shared feature of BmSPI 2 and GmSPI 2. As far as we know, substrate specificity of the inhibitors with P1 Thr has not been defined. Thr in P1 also occurs in the third domain of Kazal-type inhibitors of avian egg white, the ovomucoids (Laskowski *et al.*, 1987), and in a new family of fungal proteinase inhibitors isolated from the integument and haemolymph of *B. mori* (Pham *et al.*, 1996).

Possible function of low molecular silk components

B. mori genes for seroins and proteinase inhibitors are highly expressed just before and during cocoon spinning. Such developmental expression pattern and the presence of seroins and proteinase inhibitors in the cocoon silk indicate that the functions of these proteins are linked to cocoon formation and maintenance. It must be mentioned, however, that silk production in early larval instars of *B. mori* is negligible. In the larvae of *G. mellonella*, which spin intensively from the second larval instar to construct the silk tubes in which they live, the expression of the *seroin* gene begins in early larval instars, suggesting that seroin is important not only for the cocoon but for the silk in general. The high proline content of *G. mellonella* seroin invites suggestion that it may be a functional analogue of the proline-rich bacteriostatic peptides (Žurovec *et al.*, 1998b). The proteinase inhibitors, however, are produced in *G. mellonella* exclusively during cocoon spinning (Nirmala *et al.*, 2001). A possible explanation is that their early production would interfere with the ingestion of silk by *G. mellonella* larvae. The larvae consume the old tubes that cannot accommodate them any longer and obviously use the silk as a source of proteins that are scarce in their diet.

The conservation of seroins and silk proteinase inhibitors in the silk of distantly related lepidopteran species suggests

that they are important. They might provide some protection against microbial degradation but also against silk consumers and potential predators of the pupae. The environmental endurance of silk is well known (Akai, 1997).

Experimental procedures

Insects and tissues

Caterpillars of the Indian *B. mori* commercial hybrid Tamilnadu White \times NB₄D₂ were reared on mulberry leaves or artificial diet at 22–25 °C. Fat body, integument and silk glands were dissected from water-anaesthetized larvae of the final or the penultimate instars. The middle and posterior sections of the silk glands were separated as described by Couble *et al.* (1987). All tissues were immediately frozen in liquid nitrogen and stored at –80 °C until required.

Establishment and screening of the EST database

A cDNA library, which was based on poly(A)RNA from the middle silk glands of the last instar *B. mori* larvae and cloned in λ ZAP II vector (Mach *et al.*, 1995), was kindly provided by Dr V. Mach. The pBluescript plasmids containing cDNA inserts were excised using ExAssist™ helper phage (Stratagene), grown on agar plates, and 1000 randomly selected plasmid clones were isolated with the aid of the Kurabo PI-100 machine. The inserts of individual clones were amplified by PCR using primers complementary to flanking vector sequences. Electrophoresis was employed to verify that a single PCR product was obtained from each clone (clones generating more products were discarded). Suitable PCR products were treated with shrimp alkaline phosphatase and exonuclease 1 (Amersham Direct PCR Sequencing Kit), and taken for BigDye terminator sequencing reaction with 5' sequencing primer. Nucleotide sequences were established on the ABI Prism 377XL DNA sequencer (Applied Biosystems, USA), and stored in an EST database (Mita *et al.*, 1999). Identified cDNA sequences were analysed with the BLAST & BEAUTY software (Worley *et al.*, 1998). Separately we searched for the cDNAs whose translation products would match N-terminal sequences of several peptides we isolated from *B. mori* cocoons. In a later stage, we screened the database for the presence of proteinase inhibitors whose sequences were published by Kurioka & Hirano (1995) and Kurioka *et al.* (1999). Alignment of nucleotide sequences and amino acid residues was carried out using CLUSTAL as implemented by the Megalign program of the LASERGENE package (DNASTAR).

Reverse transcription PCR

Total RNA was extracted from the middle silk gland of a cocoon-spinning larva as described previously (Yang *et al.*, 1996). About 1 µg RNA was used as template in a 50 µl reaction mixture containing 1 \times AMV/Tfl reaction buffer (Promega), 200 µM dNTP, 1 mM MgSO₄, 0.1 units each of AMV reverse transcriptase and Tfl DNA polymerase, and 50 pmol of each primer. A degenerate forward primer 5' CCN ATH AAR ACN GGN CCN TG 3', which was based on the conserved region of a number of Kunitz-type inhibitors (Nirmala *et al.*, 2001), was combined with adapter primer 5' TGA GCA AGT TCA GCC TGG TTA TTT TTT TTT TTT TTT 3' designed in our lab. Cycling protocol included reverse transcription at 48 °C for 45 min followed by initial denaturation at 94 °C for 2 min, thirty cycles each including 30 s at 94 °C, 30 s at 60 °C and

1 min at 68 °C, and final extension at 68 °C for 7 min. The 5' end of the gene was obtained by Rapid Amplification of cDNA Ends (5' RACE) using two gene specific antisense reverse primers (5' TAC CAT AAT TAA TTA GGT CAC 3' and 5' TCG CAC TCT TCT ATA GTC TCA 3'). First strand products were purified using the Glass-MAX DNA isolation cartridge (Gibco BRL). Our adapter primer (see above) was used as the 5' anchor and dATP was employed to tail the 5' end of the first strand cDNA. PCR reactions were performed in the thermal cycler GeneAmp PCR System 2400 (Perkin Elmer). The products were separated by agarose electrophoresis, extracted with an extraction kit (Qiagen), ligated into the pGEM T easy vector (Promega), and sequenced using M13 forward and reverse primers in an ABI prism sequencer (Perkin Elmer model 310).

Northern blotting

Aliquots of 10 µg total RNA from various tissues were separated on 1.5% agarose gel containing formaldehyde and MOPS, and capillary blotted to Hybond-N⁺ nylon membrane (Amersham). The probes included complete cDNA sequences encoding seroin 1, seroin 2 and BmSPI 2. A 240 bp PCR product containing the entire 5' end but truncated close to the ORF termination was used as probe for BmSPI 1. The cDNA probes were labelled with $\alpha^{32}\text{P}$ (dATP) using the multi prime labelling kit (Amersham). Hybridization was done at 65 °C in 0.25 M Na₂PO₄, 1 mM EDTA, 1% BSA and 7% SDS for 16 h. Membranes were washed at the same temperature twice in 25 mM Na₂PO₄, 1 mM EDTA, 1% BSA and 5% SDS, and twice in 25 mM Na₂PO₄, 1 mM EDTA and 1% SDS.

GenBank accession numbers

Established cDNA sequences of *Seroin 1*, *Seroin 2*, *BmSPI 1* and *BmSPI 2* have been registered in the GenBank under accession numbers AF352584, AF352585, AF352583 and AF354456, respectively.

Acknowledgements

We thank our colleagues Dr Václav Mach for the gift of cDNA library and Dr Dalibor Kodrík for the help with N-terminal sequencing of the silk proteins. The research was supported by grants 204/00/0019 from the Grant Agency, and ME 204 from the Ministry of Education, Youth and Sports, of the Czech Republic. In Japan, the construction of EST database was partly financed by grant-in-aid from the Ministry of Education, Science, Sports and Culture, and by the 'Program for Promotion of Basic Research Activities for Innovative Biosciences' (PROBRAIN).

References

Akai, H. (1997) Anti-bacteria function of natural silk materials. *Int J Wild Silkworm Silk* **3**: 79–81.
 Bode, W. and Huber, R. (1992) Natural protein proteinase inhibitors and their interaction with proteinases. *Eur J Biochem* **204**: 433–451.
 Boigegrain, R., Pugniere, M., Paroutaud, P., Castro, B. and Brehelin, M. (2000) Low molecular weight serine protease

inhibitors from insects are proteins with highly conserved sequences. *Insect Biochem Mol Biol* **30**: 145–152.
 Couble, P., Michaille, J.J., Couble, M.L. and Prudhomme, J.C. (1987) Developmental switches of sericin mRNA splicing in individual cells of *Bombyx mori* silkgland. *Dev Biol* **124**: 431–440.
 Greene, L.J. and Bartelt, D.C. (1969) The structure of the bovine pancreatic secretory trypsin inhibitor – Kazal's inhibitor. II. The order of the tryptic peptides. *J Biol Chem* **244**: 2646–2657.
 Grzelak, K., Couble, P., Garel, A., Kludkiewicz, B. and Alrouz, H. (1988) Low molecular weight silk proteins in *Galleria mellonella*. *Insect Biochem* **18**: 223–228.
 Hui, C. and Suzuki, Y. (1995) Regulation of the silk protein genes and the homeobox genes in silk gland development. In: *Molecular Model Systems in the Lepidoptera* (eds M. R. Goldsmith and A. Wilkins), pp. 249–271. Cambridge University Press, UK.
 Inoue, S., Tanaka, K., Arisaka, F., Kimura, S., Ohtomo, K. and Mizuno, S. (2000) Silk fibroin of *Bombyx mori* is secreted, assembling a high molecular mass elementary unit consisting of H-chain, I-chain, and P25, with a 6:6:1 molar ratio. *J Biol Chem* **275**: 40517–40528.
 Kodrík, D. (1992) Small protein components of the cocoons in *Galleria mellonella* (Lepidoptera, Pyralidae) and *Bombyx mori* (Lepidoptera, Bombycidae). *Acta Ent Bohemoslov* **89**: 269–273.
 Kurioka, A. and Hirano, H. (1995) Partial amino acid sequence of a 11 kDa cocoon shell Kazal-type protein (in Japanese). *J Seric Sci Jpn* **64**: 392–394.
 Kurioka, A., Yamazaki, M. and Hirano, H. (1999) Primary structure and possible functions of a trypsin inhibitor of *Bombyx mori*. *Eur J Biochem* **259**: 120–126.
 Laskowski, M. and Kato, I. (1980) Protein inhibitors of proteinases. *Ann Rev Biochem* **49**: 593–626.
 Laskowski, M., Kato, I., Ardelt, W., Cook, J., Denton, A., Empie, M., Kohr, W., Park, S., Parks, K., Schatzley, B., Schoenberger, O., Tashiro, M., Vichot, G., Whatley, H., Wieczorek, A. and Wieczorek, M. (1987) Ovomucoid third domains from 100 avian species: Isolation, sequences, and hypervariability of enzyme-inhibitor contact residues. *Biochemistry* **26**: 202–221.
 Mach, V., Takyia, S., Ohno, K., Handa, H., Imai, T. and Suzuki, Y. (1995) Silk gland factor-1 involved in the regulation of *Bombyx* sericin-1 gene contains fork head motif. *J Biol Chem* **270**: 9340–9346.
 Michaille, J.J., Couble, P., Prudhomme, J.-C. and Garel, A. (1986) A single gene produces multiple sericin messenger RNAs in the silkgland of *Bombyx mori*. *Biochimie* **68**: 1165–1173.
 Mita, K., Morimyo, M., Okano, K., Shimada, T. and Maeda, S. (1999) The construction of EST database for genome analysis of *Bombyx mori*. *RIKEN Rev* **22**: 63–67.
 Nirmala, X., Kodrík, D., Žurovec, M. and Sehnal, F. (2001) Insect silk contains a Kunitz-type and a unique Kazal-type proteinase inhibitors. *Eur J Biochem* **268**: 1–10.
 Pham, T., Hayashi, K., Takano, R., Itoh, M., Eguchi, M., Shibata, H., Tanaka, T. and Hara, S. (1996) A new family of serine protease inhibitors (Bombyx family) as established from the unique topological relation between the positions of disulphide bridges and reactive site. *J Biochem* **119**: 428–434.
 Ramesh, N., Sugumaran, M. and Mole, J.E. (1988) Purification and characterization of two trypsin inhibitors from the hemolymph of *Manduca sexta* larvae. *J Biol Chem* **263**: 11523–11527.

- Sasaki, T. (1978) Chymotrypsin inhibitors from hemolymph of the silkworm, *Bombyx mori*. *J Biochem (Tokyo)* **84**: 267–276.
- Sehnal, F. and Akai, H. (1990) Insect silk glands: their types, development and function, and effects of environmental factors and morphogenetic hormones on them. *Int J Insect Morph Embryol* **19**: 79–132.
- Sugumaran, M., Saul, S.J. and Ramesh, N. (1985) Endogenous protease inhibitors prevent undesired activation of prophenolase in insect hemolymph. *Biochem Biophys Res Comm* **132**: 1124–1129.
- Von Heijne, G. (1983) Patterns of amino acids near signal-sequence cleavage sites. *Eur J Biochem* **133**: 17–21.
- Worley, K.C., Culpepper, P., Wiese, B.A. and Smith, R.F. (1998) BEAUTY-X: enhanced BLAST searches for DNA queries. *Bioinformatics* **14**: 890–891.
- Yang, C., Sehnal, F. and Scheller, K. (1996) Juvenile hormone restores larval pattern of sericin gene transcripts. *Arch Insect Biochem Physiol* **33**: 353–362.
- Žurovec, M., Kodrík, D., Yang, C., Sehnal, F. and Scheller, K. (1998a) The P25 component of *Galleria* silk. *Mol Gen Genet* **257**: 264–270.
- Žurovec, M., Sehnal, F., Scheller, K. and Kumaran, A.K. (1992) Silk gland specific cDNAs from *Galleria mellonella* L. *Insect Biochem Molec Biol* **22**: 55–67.
- Žurovec, M., Vašková, M., Kodrík, D., Sehnal, F. and Kumaran, A.K. (1995) Light-chain fibroin of *Galleria mellonella* L. *Mol Gen Genet* **247**: 1–6.
- Žurovec, M., Yang, C. and Sehnal, F. (1998b) Identification of a novel type of silk protein and regulation of its expression. *J Biol Chem* **273**: 15423–15428.