A highly divergent gene cluster in honey bees encodes a novel silk family

Tara D. Sutherland,[1] Peter M. Campbell, Sarah Weisman, Holly E. Trueman, Alagacone Sriskantha, Wolfgang J. Wanjura, and Victoria S. Haritos

CSIRO Entomology, Canberra ACT 2601, Australia

The pupal cocoon of the domesticated silk moth *Bombyx mori* is the best known and most extensively studied insect silk. It is not widely known that *Apis mellifera* larvae also produce silk. We have used a combination of genomic and proteomic techniques to identify four honey bee fiber genes (*AmelFibroin*I–4) and two silk-associated genes (*AmelSA*1 and 2). The four fiber genes are small, comprise a single exon each, and are clustered on a short genomic region where the open reading frames are GC-rich amid low GC intergenic regions. The genes encode similar proteins that are highly helical and predicted to form unusually tight coiled coils. Despite the similarity in size, structure, and composition of the encoded proteins, the genes have low primary sequence identity. We propose that the four fiber genes have arisen from gene duplication events but have subsequently diverged significantly. The silk-associated genes encode proteins likely to act as a glue (*AmelSA*1) and involved in silk processing (*AmelSA*2). Although the silks of honey bees and silkmoths both originate in larval labial glands, the silk proteins are completely different in their primary, secondary, and tertiary structures as well as the genomic arrangement of the genes encoding them. This implies independent evolutionary origins for these functionally related proteins.

[Supplemental material is available online at www.genome.org. The sequence data from this study have been submitted to the honey bee BeeBase Official_Gene_Set_1 under accession nos. GB17818, GB19585, GB12184, GB12348, and GB15233.]

Many holometabolous insects produce a silken cocoon in the late larval stage for protection during pupal metamorphosis. The pupal cocoon of the domesticated silkmoth, *Bombyx mori,* is the best known and most extensively studied among insect silks; it is less widely known that honey bee larvae also construct silken cocoons within which they pupate. Successive layers of silk coat the brood cell walls, accumulating with each larval generation to comprise a significant proportion of the hive. As beeswax is a thermoplastic material which loses strength and stiffness with increasing temperature, it is thought that the accumulated silk gives the comb tensile strength and mechanical integrity (Hepburn and Kurstjens 1988).

B. mori silk is a composite of a 390-kDa protein (H-fibroin) associated with two smaller proteins, the 25-kDa L-fibroin and the 25-kDa P25 fibroin (Zhou et al. 2000). H-fibroin is composed of alternating hydrophobic and hydrophilic blocks. Each hydrophobic block is assembled from tandemly repeated motifs composed of glycine, alanine, and serine residues, which make up 88% of the amino acids in the protein (Zhou et al. 2000). These blocks form antiparallel pleated β-sheets which impart strength and toughness to the silk (Takahashi et al. 1999). The hydrophilic blocks are usually amorphous, providing elasticity to the fiber. *B. mori* silk also contains at least six sericin glues that provide structural integrity to the cocoon. The large (23 kbp) sericin 1 gene (*Ser1*) is made up of nine exons, which can be differentially spliced to produce five transcripts that produce protein variants ranging from 65 to 400 kDa (Garel et al. 1997). A comparable honey bee sericin has not been reported. In contrast to the growing body of information on silk genes of *B. mori* and related silkmoths, nothing is known of the genes encoding honey bee silks and associated glues.

Honey bee and *B. mori* silks are both produced from modified salivary glands known as labial glands. The labial glands of Hymenoptera and Lepidoptera are thought to be homologous, originating from the lateral ventral placodes of the labial ectoderm (Julien et al. 2004). The mechanism of silk formation within the labial glands has been shown to be similar in Lepidoptera and *Apis mellifera* (Akai et al. 1987; Silva-Zacarin et al. 2003). Polymerization of the silk protein occurs in the gland lumen followed by gradual dehydration of the polymerized silk during extrusion to form an insoluble silk filament.

X-ray diffraction patterns obtained from native silk fibers from honey bee larvae have an α-helical structure (Rudall 1962), in marked contrast to the antiparallel pleated sheets of *B. mori*. The fine structure of the honey bee silk X-ray diffraction patterns is indicative of a tetrameric coiled coil system, although it contains certain unexplained features (Atkins 1967). Amino acid analysis of honey bee silk has shown high levels of alanine, serine, aspartic, and glutamic acids and reduced amounts of glycine by comparison with the lepidopteran fibroins (Lucas and Rudall 1968). As the unusual secondary structure and amino acid profile of honey bee silk imply a novel primary silk sequence, we have investigated the silk proteins and genes and their genomic organization. This report describes the relatively compact genes that encode a remarkable and highly diverse family of proteins that comprise the silk present in the honey bee broodcomb.

Results and Discussion

Identification of novel silk proteins

Honey bee silk genes are expressed in the final instar specifically in the labial (modified salivary) gland. We constructed a cDNA

[1]**Corresponding author.**
E-mail Tara.Sutherland@csiro.au; fax +61 2 6246 4000.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.5052606. Freely available online through the *Genome Research* Open Access option.

16:1414–1421 ©2006 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/06; www.genome.org

library from the labial gland of late final instar larvae and obtained sequence information from 82 randomly selected clones. These cDNAs clustered into 46 groups, with 38 clusters represented by a single cDNA. The most abundant cluster comprised 13% of the analyzed library and other clusters, represented by more than a single cDNA, varied in abundance from 2% to 11%. A summary of the most abundant cDNAs is shown in Table 1. A full listing of the 102 ESTs and proteins identified in the larval labial gland can be found in Supplemental Table I.

The silk proteins were identified by matching the mass spectrometry peptide fragmentation patterns obtained after tryptic digestion of honey bee silk with in silico predictions from three data sets: cDNAs obtained from the labial gland, all protein sequences predicted by the honey bee genome sequencing project, and a database of translations in the six possible reading frames of each contiguous genomic DNA sequence provided by the bee genome project (Amel_3.0 release). Eight sequences matching tryptic peptides from silk were identified. Six of these proteins (and no others) were identified in silk fiber after treatment to remove nonfibrous proteins. Sequences corresponding to these six silk proteins were identified in the honey bee genome data sets—five corresponded to the existing automated protein annotations GB15233, GB12184, GB12348, GB17818, and GB19585. The sixth sequence matched a protein encoded within a large open reading frame spanning Contig2271 and Contig2272. The peptide hits to Contig2271 and 2272 are shown in Supplemental Figure I. As expected for extracellular proteins, all six proteins contained classical 19-residue secretory signal peptides, identified by SignalP 3.0 (Bendtsen et al. 2004) and Wolf-PSORT (http://wolfpsort.seq.cbrc.jp) algorithms.

The silk was not fully digested by the tryptic enzyme described above. Thus there was a possibility that remaining proteins in the undissolved silk were not identified. Since silk is produced by the labial gland, it was expected that the silk proteins would be abundant in gland extracts. This organ, completed with lumen contents, was dissected from final instar larvae, extracted with SDS, and proteins were separated by polyacrylamide gel electrophoresis (SDS-PAGE) (Fig. 1). Bands were excised from the electrophoresis gel and analyzed by in-gel trypsin digestion and mass spectrometry. Consistent with expectations, the major Coomassie blue-stained bands visible after SDS-PAGE were identical to the proteins found in the native silk. The mass spectral fragments obtained from single bands often matched more than one silk protein and all the silk proteins are identified in more than one band (Fig. 1). We suggest that this reflects partial polymerization of the silk proteins in the labial gland. Less intense gel electrophoresis bands were identified as typical "house-keeping" proteins that are abundant in most tissues (see Supplemental Table I for a full listing of proteins identified).

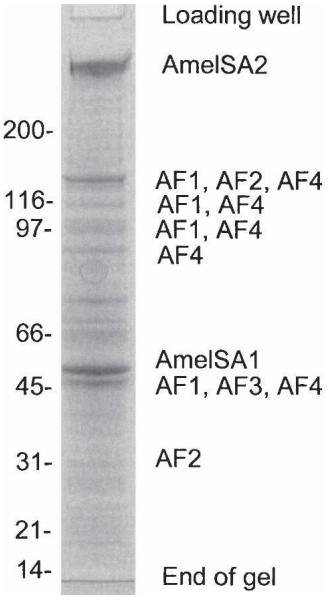

Figure 1. Analysis of proteins in the labial gland of late final instar *Apis mellifera* by SDS-PAGE showing that major protein bands correspond to identified silk proteins. Molecular weight markers (kDa) are shown alongside the gel. AF1, AmelFibroin1; AF2, AmelFibroin2; AF3, AmelFibroin3; AF4, AmelFibroin4. As Coomassie brilliant blue dye binds to basic and aromatic residues (Compton and Jones 1985), the relative intensity of the staining seen in this gel may not be linearly proportional to protein abundance. AmelSA1 and AmelSA2, the two most prominent bands, contain approximately twice the basic plus aromatic content of the AmelFibroin proteins.

Five of the silk proteins correspond to the most abundant clones from the labial gland cDNA library (GB15233, GB12184, GB12348, GB17818, GB19585). The predicted TATA boxes for these genes are close to the initiator methionine (55–57 nt) resulting in short untranslated regions. The energy cost of transcription has selected for short introns in highly expressed proteins (Castillo-Davis et al. 2002) and the same pressure can be expected to also select for short untranslated regions. The protein found in Contig 2271 and 2272 was represented by a single cDNA clone in the labial gland library, abundant in the labial gland protein extract and present at apparently trace levels in the silk (Table 1). The two proteins lost from silk after extraction in boiling alkali solution (GB12121, GB11658) were not encoded by analyzed cDNAs from the labial gland (albeit this was only 82 sequences) and were not identified as proteins from labial gland extracts. The honey bee larval cocoons include material from multiple origins including a secretion from the malpighian tubules and material from the mid-gut (Jay 1964), so these proteins possibly represent nonsilk cocoon components.

Overall, we have identified six genes encoding silk proteins. Structural analysis of the encoded proteins (see below) indicates that these genes fall into three groups: (1) four small genes corresponding to BeeBase annotations

Table 1. Identification of honey bee silk proteins

Protein synonyms	Abundance in labial gland cDNA library (%)	MS-MS confidence score (% sequence coverage) of trypsin digested proteins		
		Raw silk	Na_2CO_3-treated silk	Labial gland
AmelSA1	13	40 (9)	34 (5)	184 (34)
AmelFibroin3	11	89 (18)	108 (27)	189 (43)
AmelFibroin2	7	51 (12)	48 (13)	227 (45)
AmelFibroin4	7	86 (22)	88 (21)	163 (37)
AmelFibroin1	6	41 (9)	23 (6)	220 (49)
AmelSA2	1	18[a] (< 1)	23 (< 1)	821 (14)

[a]MS-MS scores <20 are not considered a confident identification.

GB17818, GB19585, GB12184, and GB12348 encoding fibrous proteins of 30–34 kDa that we have termed *AmelFibroin*1–4 (*Apis mellifera* fibroin), respectively; (2) a small gene corresponding to GB15233 that we have termed *AmelSA*1 (*Apis mellifera* silk-associated) that encodes a possible glue of 42 kDa; and (3) a large gene spanning Contig2271 and 2272 that we have named *AmelSA*2 (*Apis mellifera* silk-associated) encoding a high molecular weight (500 kDa) protein whose role in the silk, if any, is currently unclear.

Four of the honey bee silk genes are encoded in a high GC gene cluster with a similar, nonsilk gene

The four *AmelFibroin* genes are clustered sequentially in the *A. mellifera* genome (Fig. 2). Each gene is a single exon and the genes are separated by short intergenic regions (1659–1796 nt). The open reading frames of the genes have two features that distinguish them from the surrounding DNA: They have high GC content (*AmelFibroin*1–57%; *AmelFibroin*2–55%; *AmelFibroin*3–59%; *AmelFibroin*4 –62%), in contrast to the characteristically low GC content found in the honey bee genome, and they correspond to CpG islands (identified using newcpgseek). The intergenic regions do not have either high GC or higher-than-expected CpG.

Immediately upstream (1487 nt) of the silk gene cluster is another, slightly longer gene (1737 nt), with a somewhat elevated GC content (45%) and a short CpG island (Fig. 2). This gene is identified by a single cDNA in the labial gland library, and the protein (GB12085) is present at low levels in gland extracts (Supplemental Table I) but is not found in the cocoon silk. The protein has structural similarities to the AmelFibroin silk fiber proteins (see structural analysis below). We have named this open reading frame *AmelFibroin-rel* (and the protein AFrel, Amel-Fibroin-related) due to the probability that this gene has arisen through gene duplication events from a common ancestor shared with the *AmelFibroins*.

The high GC content corresponding to the open reading frames of this cluster is partially the result of the abundance (29%–33%) of the GCX codon, which encodes alanine. However, the alanine content of the gene product is not sufficient to completely account for the high GC content. Previously, genomic regions of GC bias have been associated with highly expressed genes in the human and *Drosophila* genomes (Versteeg et al. 2003; Stenoien and Stephan 2005), and the *AmelFibroin* genes appear to fit this pattern.

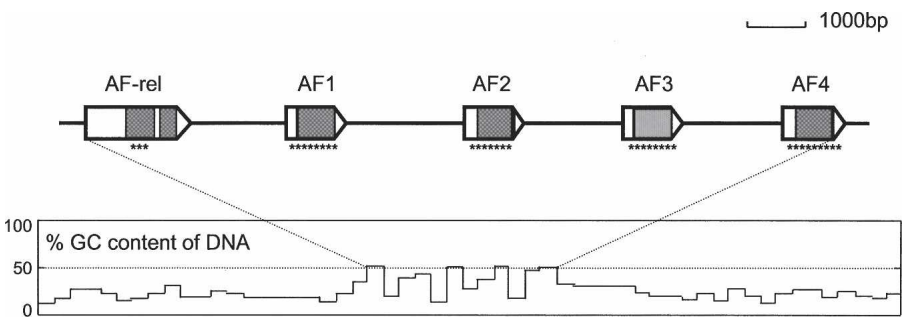

Figure 2. A schematic showing the genetic organization of *AmelFibroin*1–4 and *AmelFibroin-rel* genes in the *Apis mellifera* genome and the GC content and CpG islands in the corresponding genomic region. The open reading frames of the *AmelFibroin* genes are shown as block arrows with the regions encoding heptads shaded. Intergenic regions are shown to scale as lines. The GC content is redrawn from BeeBase (http://racerx00.tamu.edu/bee_resources.html). Regions corresponding to CpG islands (predicted by newcpgseek) are indicated by stars.

The high alanine content also contributes to a strong nucleotide bias with a high use of G (50%–57%) and a low use of C (9%–11%) at the first codon position and a high use of C (43%–46%) and a low use of G (8%–12%) in the second position. The same trends are found in *AmelFibroin-rel* although the bias is less (40% G and 11% C in first and 31% C and 9% G in second codon position). Despite the strong nucleotide bias in the first and second codon position, there is no GC bias in the third position, and this is reflected in an absence of strong bias in codon usage in the *AmelFibroin* genes.

The structure of the honey bee silk proteins is a novel coiled coil

As described in the introduction, previous analysis of honey bee silk X-ray diffraction patterns demonstrated that honey bee silk is predominantly α-helical (Rudall 1962), in a coiled coil form most likely involving four strands (Atkins 1967). Coiled coils are widely distributed in nature and are formed when helical strands interact to shield hydrophobic core residues from hydration. Coiled coils are characterized by a seven residue (heptad) repeat sequence conventionally denoted as $(abcdefg)_n$ with hydrophobic residues in position *a* and *d* and polar and charged residues filling the remaining positions. The regularity of character of the seven-residue repeat has allowed the development of numerous computational approaches to identify heptad periodicity in protein sequences (a comparison of algorithms can be found in Lupas and Gruber 2005).

The secondary structure algorithms PROFsec (Rost and Sander 1993) and NNPredict (McClelland and Rumelhart 1988; Kneller et al. 1990) predict that the AmelFibroin proteins are highly helical, with between 76% and 85% helical structure and insignificant β-sheet formation (Table 2). The MARCOIL program (Delorenzi and Speed 2002; http://www.isrec.isb-sib.ch/webmarcoil/webmarcoilC1.html) identifies numerous heptad repeats (27–30 at 90% confidence level) indicative of coiled coils in the helical regions of the four AmelFibroin proteins, with a single five-residue interruption in AmelFibroin1. Additional subsets of offset heptads are predicted in AmelFibroin1 and 2. The almost complete absence of discontinuities in the heptad repeat regions is uncommon in coiled coils and is consistent with these proteins adopting a highly regular structure. There is significant similarity between heptads, as evident in strong hatched patterns seen in dot plot matrices (Fig. 3) although exact heptad sequences are not conserved either within or between proteins. The lack of primary amino acid identity is characteristic of coiled coils, where residue character, in particular the positions *a* and *d*, is more important than residue identity. However, hydrophobic/hydrophilic character requirements limit the type of residue that allows coiled coil formation, resulting in the low complexity sequences evident in the dot plot analysis.

The amino acid composition of the *AmelFibroin* protein heptad repeats is quite distinctive compared with classical coiled coils. The silk proteins are particularly unusual in their high alanine content at positions *a* and *d* (between 43% and 71% alanine occupancy at either po-

Table 2. Properties of silk and related proteins from honey bee compared with the major silk proteins from silkworms

Protein synonyms	Secondary structure (%)[a]		Amino acids in mature protein	Predicted molecular weight (kDa)	Hydropathicity[b]	Heptad analysis[c]				
	Alpha	Beta				Number of heptads	Ala (%)	Ala in *a* (%)	Ala in *d* (%)	Ala in *a* and *d* (%)
Apis mellifera										
AmelFibroin1	76	—	315	33	−0.14	27	41	44	74	33
AmelFibroin2	88	—	290	30	0.04	28	37	71	57	36
AmelFibroin3	81	—	316	33	−0.20	30	37	63	43	27
AmelFibroin4	76	—	323	34	−0.11	29	48	79	58	38
AFrel	75	—	552	61	−0.38	37	27	35	43	5
AmelSA1	41	7	350	42	−0.64	—	n/a	n/a	n/a	n/a
AmelSA2	21	18	4262	490	−1.24	—	n/a	n/a	n/a	n/a
Bombyx mori										
Heavy fibroin	—	64	5242	392	0.21	—	n/a	n/a	n/a	n/a
Sericin 1B	—	22	1199	121	−1.08	—	n/a	n/a	n/a	n/a

[a]Secondary structure predicted by PROFsec.
[b]Hydropathicity predicted by GRAVY (Kyte and Doolittle 1982).
[c]Heptads predicted by MARCOIL.

sition, Table 2) and in the frequency of heptads with alanine occupancy in both *a* and *d* positions (27%–38% of heptads). In naturally occurring coiled coils, alanine is among the least favored hydrophobic residue in these positions (Woolfson 2005). One of the main driving forces for coiled coil formation is sequestration of the hydrophobic core from solvent exposure, and it is generally regarded that the relatively low hydrophobicity of alanine cannot stabilize an extended coiled coil conformation. We are very interested in the factors contributing to the stability of the alanine-rich cores of the honey bee silk coiled coils.

The high level of alanine in the core of the honey bee silk coiled coil regions may be expected to result in close spacing of the helices and the formation of very tight coiled coils. The X-ray diffraction pattern from honey bee silk fibrils was fitted to a tetrameric coiled coil model with an unprecedentedly short major helix radius R_0 = 5.2 Å (Atkins 1967). This is consistent with recent studies of a trimeric coiled coil formed by an engineered peptide containing exclusively alanine residues in the *a* and *d* heptad positions, which measured the major helix radius of the coiled coil as 5.1 Å (Liu and Lu 2002). For comparison, the radius for the well-known GCN4 leucine zipper trimeric structure is ~6.8 Å and tetrameric structure is ~7.1 Å.

The *AmelFibroin-rel* product, AFrel, was also predicted to be 72% α-helical with two regions of coiled coil, comprising 22 and 27 heptads separated by 25 residues. In contrast to the silk fiber proteins, the AFrel heptads have a lower overall alanine content and only 5% of heptads have alanine at both *a* and *d* position, suggesting a significant difference in the type of coiled coil formed by this protein (Table 2).

Generally positions other than *a* and *d* in the silk fiber protein heptads are populated by charged and polar residues, as expected for coiled coils. However, alanine is also quite abundant at these positions in the silk proteins (see Supplemental Table II). An increase in hydrophobicity outside positions *a* and *d*, in particular in positions *e* and *g*, is indicative of multistranded coiled coils, where more surface area per helix is buried and positions *e* and *g* contribute to core hydrophobic stability (Krammerer 1997). A high level of alanine in the noncore heptad positions may also be important for silk function in the waxen beehive environment. In contrast to previously described coiled coils, which are found in aqueous environments, the honey bee silk proteins function in a highly hydrophobic environment, and it is likely that the hydrophobicity of the noncore positions facilitates this interaction. We are investigating other Aculeate species that do not form wax nests to determine whether the silk proteins of these species have such distinctive noncore hydrophobicity.

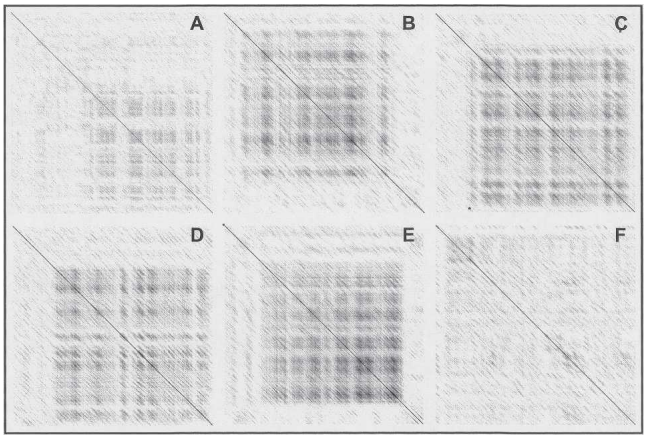

Figure 3. Dot matrix similarity analysis of honey bee silk and related proteins. The horizontal and vertical rows of dots seen as hatched patterns demonstrate repeats of the same sequence character. (*A*) AFrel; (*B*) AmelFibroin1; (*C*) AmelFibroin2; (*D*) AmelFibroin3; (*E*) AmelFibroin4; (*F*) AmelSA1.

Divergence of honey bee silk genes

BLAST or PSI-BLAST analyses using default settings on the NCBI Web page or as set in Predict Protein (http://cubic.bioc.columbia.edu/predictprotein/submit_def.html) of the *AmelFibroin* sequences against the nonredundant NCBI protein and DNA or EST databases, or databases of *A. mellifera*, *Anopheles gambiae* str. PEST, *B. mori*, *Drosophila melanogaster*, and *Tribolium castaneum*, do not identify any similar proteins (BLAST expectation values <0.001 over >10% of the query protein). Our inability to find orthologs in the available Dipteran, Lepidopteran, and Coleopteran whole genome data sets suggests that the honey bee silk fiber genes may be unique to the Hymenoptera lineage.

Table 3. Percentage of amino acid identity between AmelFibroin proteins

	AmelFibroin2	AmelFibroin3	AmelFibroin4
AmelFibroin1	25.1	28.8	30.6
AmelFibroin2		27.5	21.9
AmelFibroin3			24.0

The characteristics shared by the *AmelFibroin* genes (genetic location, length, nucleotide usage, and the proteins' amino acid composition, secondary and tertiary structure) suggest that the genes are paralogs. Despite the similarity in genetic and protein characteristics, the genes have low primary sequence similarity and encode proteins with low primary sequence similarity (Table 3). It was difficult to obtain convincing alignments using the conventional alignment algorithms. However, we were able to utilize the proteins' secondary structure predictions to manually align the genes as described in the Methods section. The protein sequence alignment (translated from the nucleotide alignment) is shown in Figure 4. The best estimate of phylogenetic relatedness between the four *AmelFibroin* genes is shown graphically in Figure 5.

It is likely that *AmelFibroin-rel* shares a common ancestor with the *AmelFibroin* genes due to its genomic colocation, genetic structure, and very similar protein amino acid composition and heptad substructure. The inability to align *AmelFibroin-rel* to the *AmelFibroin* genes suggests that *AmelFibroin-rel* is the most distantly related member of the gene family. Although it is found in the labial gland, AFrel is not found in the silk. Two ESTs from *AmelFibroin-rel* have been isolated from the honey bee brain (BB170026B10F06 and BB170026A20D06 from the adult bee brain library, BB17), an organ that does not produce silk. Coiled coils can form large regular structures (as proposed for the honey bee silks) or can mediate more dynamic interactions as transcriptional factors, receptors, and signaling molecules. The differential expression of *AmelFibroin-rel* suggests that AFrel may be playing a regulatory rather than structural role in silk production, consistent with its divergent sequence. Although AFrel is predicted to contain a 22-residue signal peptide, the *k*-NN algorithm suggests the protein is more likely associated with the mitochondria or targeted to the nucleus than secreted extracellularly.

One silk protein behaves like a glue but is unrelated to known sericins

*AmelSA*1 is represented by the most abundant labial gland cDNA, and the *AmelSA*1 protein (AmelSA1) is a significant component in honey bee silk (Table 1; Fig. 1). No sequences related to *AmelSA*1 were found by BLAST analysis against the NCBI protein, DNA or EST databases, or any of the genome databases, except that two *AmelSA*1 sequences were found in cDNA libraries synthesized from all five combined larval stages of Africanized *A. mellifera* (Nunes et al. 2004).

The mature silk is composed of fibrous threads that are glued together to form sheets, so we expect to find at least one silk protein that acts to glue the coiled coils together. The predicted structure of AmelSA1 is mainly amorphous (52%), and, as the α-helical regions do not contain coiled coil-forming heptads, it is not a fibrous silk protein. The protein is strongly hydrophilic, a property characteristic of *B. mori* sericins (Table 2). We attempted to remove glue components from the silk by boiling in sodium carbonate solution, a standard technique for degumming *B. mori* silk (Yamada et al. 2001). Amino acid analyses showed that the composition of silk before the treatment was a mixture of silk fiber proteins (AmelFibroin1–4) with AmelSA1 (and probably some contaminants), whereas the measurements after the treatment were shifted much closer to the composition of pure fiber proteins (data not shown). MS/MS analysis identified fewer AmelSA1 tryptic peptides after sodium carbonate treatment of silk (Table 1), indicating that this protein was largely but not completely removed. Although the composition, sequence, and genetic structure are quite unlike known sericins, the properties and partial solubility of AmelSA1 suggest it is the most likely candidate for a honey bee silk glue.

The genetic organization of *AmelSA*1 is much simpler than the *B. mori* sericin genes. The silkworm sericins are encoded by two genes composed of multiple exons that are differentially spliced to generate proteins with different characteristics (Garel et al. 1997). In contrast, *AmelSA*1 contains two exons, and no evidence of differential splicing was found among the 11 *AmelSA*1 cDNAs detected in our library or from MS data, which matched AmelSA1 to a single band after SDS-PAGE of labial gland proteins (Fig. 1).

The sericins are so named because their most abundant resi-

```
AmelFibroin 4   MKIPSILAVSLLIWGLASGAREEVETRDKTKTSTVVKSEKVEVVAPAKDELKLTSEPIFGRR
AmelFibroin 3   MQIPTFVAICLLTSGLVHAG---VEEFKSSATEEVIS----KNLEVDLLKNVDTSAKRRENG
AmelFibroin 2   MKIPAIFVTSLLVWGLAEGR---VINHESLKTSEDI---------QGGYSAGIVGDGSDALG
AmelFibroin 1   MKIPVLLATCLYLCGFASAG---LEGPGNSLPELVKG-------------------------

AmelFibroin 4   VGTGASEVASSSGEAIAISLGAGQSAAESQALAASQSKTAANAAIGASELTNKVAALVAGAT
AmelFibroin 3   APVLGKNTLQSLEKIKTSASVNAKAAAVVKASALALAEAYLRASALSAAASAKAAAALKNAQ
AmelFibroin 2   SSIENAQKVARAAENVGLNLELGAGARAASVAAAAQAKNTEAAEAGANAALAAAAIAKREEAI
AmelFibroin 1   ----SASATASTAVTARSGLRAGQVALASQKDAVLQAQAAASAASEARAAADLTAKLSQESA

AmelFibroin 4   GAQARATAASSSALKASLATEEAAEEAEAAVADAKAAAEKAESLAKNLASASARAALSSERA
AmelFibroin 3   QAQLNAQEKSLAALKAQSEEEAASARANAATAATQSALERAQASSRLATVAQNVASDLQKRT
AmelFibroin 2   KASEIANQLLTNAAKAAEATVSATKRAAQLTAAAKEATRASAAAAEAATEAQVKANADSIIT
AmelFibroin 1   SVQSQAAAKGKETEEAAVGQARAGLESVSMAASATSAAKEASTAAKAAASALSTAVVQAKIA

AmelFibroin 4   NELAQAESAAAAEAQAKTAAAAKAAEIALKVAEIAVKAEADAAAAAVAAAKARAVADAAAAR
AmelFibroin 3   STKAAAEAAATLRQLQDAERTKWSANAALEVSAAAAAAETKTTASSEAANAAAKKAAAIASD
AmelFibroin 2   KRAAIAEAQAAAEAQVKAAIARKSAANFLAKAQIAAAAESEATKLAAEAVVALTNAEVAVNQ
AmelFibroin 1   ERAAKAEAVASDEAKAKAIAAANLAAEASVAAEAALKAEKVAEEAIARAASAKAAARAAAAA

AmelFibroin 4   AAAVNAIAKAEEEASAQAENAAGVLQAAASAAAESRAAAAAAAATSEAAAEAGPLAGEMKPP
AmelFibroin 3   ADGAERSASTEAQSAAKIESVAAAEGSANSASEDSRAAQLEASTAARANVAAAVGDGAIIGL
AmelFibroin 2   ARNAQANASTQASMAVRVDSQAANAEAAAVAQAETLLVTAEAVAAAEAEVANKAATFAKQIV
AmelFibroin 1   LASSKEAATASARNAAESEARNEVAVLIAEIDKKSREIDAASSLNARAAAKASSRNVETATI

AmelFibroin 4   HWKWERIPVKKEEWKTSTKEEWKTTNEEWEVK----------------------
AmelFibroin 3   GEEAGAAAQLLAQAKALAEVSSKSENIEDKKF----------------------
AmelFibroin 2   NEKKIHVAKLE-------------------------------------------
AmelFibroin 1   GANINSSKQVVSIPVEIKKFSEPEVSTSWREDEEVTKEKKEHINLNDFDLKSNVF
```

Figure 4. Protein sequence alignment translated from the nucleotide alignment of the AmelFibroin proteins. The bulk of the nucleotide alignment (Supplemental Fig. II) was derived from aligning primary sequence and conserving heptad periodicity as described in the Methods. The alanine residues populating the hydrophobic heptad positions are marked (position *a* shaded and position *d* in bold). The predicted signal peptide is underlined. Regions shown in italic were aligned using MUSCLE.

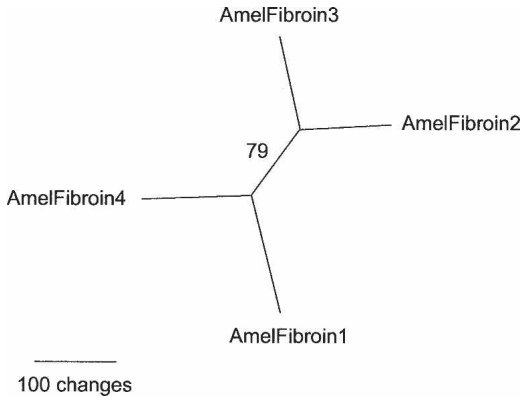

Figure 5. Maximum parsimony tree showing the relatedness of the four AmelFibroin proteins with branch length indicating the number of amino acid changes between the proteins. The bootstrap value separating the two gene clusters was calculated by carrying out 1000 replicates with PAUP* (Swofford 2002).

due is the nonessential amino acid serine. In contrast, the most abundant amino acids found in AmelSA1 are the essential amino acids leucine (17%) and lysine (17%). Use of essential amino acids above what is obtained in the diet in nonrecycled proteins involves high metabolic cost, so the high levels of essential residues in the apparently abundant AmelSA1 suggest that these residues are particularly functionally important. The silk fiber proteins incorporate 11%–16% acidic residues in the noncore positions of their heptads, so lysine residues in AmelSA1 may be involved in electrostatic interactions with the exposed surfaces of the coiled coils. Leucine is a bulky, hydrophobic amino acid, so the role of these residues may be to facilitate interactions of the silk sheets with the wax environment within the beehive.

AmelSA2

AmelSA2 encodes a high molecular weight protein (AmelSA2) prominent in a SDS-PAGE gel of labial gland extracts (Fig. 1) and strongly detected by LC-MS of the gland proteins (Table 1). In contrast, the LC-MS analysis barely identified the protein in the mature silk (less than 1% protein sequence coverage). This suggests that the protein is abundant in the dedicated silk-producing gland but found only at trace levels in the silk. This disparity suggests that the role of AmelSA2 may be in silk assembly rather than as a core component of the fiber.

PSI-BLAST analysis of the protein sequence identified a similarity to the protein Nestin (highest match to a Nestin fragment, EMBL:AF110498 with a BLAST expectation value of $7e-94$ over 355 residues, using default settings in Predict Protein). Nestin is a type VI intracellular intermediate filament (IF) protein that partially coassembles with other IF components to form heterodimer coiled coils leaving a long tail composed of highly charged peptide repeats in solution (Steinert et al. 1999). Similarly, AmelSA2 contains a short coiled coil region (seven heptads) toward its N terminus while the remainder of the protein is highly hydrophilic (Table 2), including 10 repeats of ~114 amino acids each that contain ~48% charged residues. If AmelSA2 functions similarly to Nestin, we speculate that the protein is partially incorporated into coiled coils within the gland. Its hydrophilic tail could serve the dual purposes of increasing silk solubility and sterically hindering coiled coil aggregation. It is unclear how AmelSA2 could be removed from the silk fiber prior to extrusion. Functional studies are required to elucidate the true role of this protein.

Conclusions

Six honey bee silk genes have been confidently identified by a combination of genomic and proteomic techniques. Five of these genes, encoding the four AmelFibroin proteins and the AmelSA1 glue protein, are completely novel, with no sequence similarity found to any known gene. The four *AmelFibroin* genes are physically clustered in the genome and are each composed of a single short exon. Although they encode proteins with similar amino acid composition, helical conformation, and heptad substructure, they share little primary sequence homology (Table 3). The four related but diverged genes may have slightly different roles in coiled coil formation. All four proteins might be required at fixed ratios for proper silk formation, or expression of the different genes at varying levels might allow honey bee silk to adapt rapidly to environmental changes. Alternatively, the four proteins might be functionally equivalent with gene duplication required to support the very high level of expression.

Methods

Tissue and silk preparation

A. mellifera larvae and brood comb were obtained from domestic hives. The labial gland was dissected from late fifth instar *A. mellifera* immersed in phosphate buffered saline. The posterior end of the dissected gland was immediately transferred to RNA*later* (Ambion) and stored at 4°C. The anterior end of the gland, the lumen of which contained silk proteins, was placed in LDS sample loading buffer (Invitrogen) including reductant (10% NuPage reducing agent, Invitrogen) and stored at −20°C for subsequent protein analysis.

Brood comb was washed extensively in warm water to remove water soluble contaminants and then extensively in three washes of chloroform to remove wax, producing raw silk. Further washing, analogous to treatment to remove sericins from *B. mori* silk cocoons, involved boiling the silk in aqueous 0.05% Na_2CO_3 for 30 min (Yamada et al. 2001). Approximately 35% weight of the raw silk was removed by the treatment, and a longer boiling time had no further effect. For comparison, this treatment removes about 25% weight of silkworm cocoons (Yamada et al. 2001).

cDNA library construction

Total RNA (35 µg) was isolated from the posterior end of 50 late larval labial glands using the RNAqueous for PCR kit from Ambion. mRNA was isolated from the total RNA using the Micro-FastTrack 2.0 mRNA Isolation kit from Invitrogen according to the manufacturers' directions. The cDNA library was constructed using the CloneMiner cDNA library construction kit of Invitrogen and comprised ~1,200,000 colony forming units with an average insert size of 1.3 ± 1.4 kbp. Eighty-two clones were randomly selected and sequenced using the GenomeLab DTCS Quick start kit (BeckmanCoulter) and a CEQ8000 Biorad sequencer.

Silk and labial gland protein analysis

Labial gland protein samples were separated by SDS-PAGE and stained with Coomassie blue. Major bands were digested with trypsin and the resultant peptides were analyzed by reversed phase liquid chromatography coupled by electrospray ionization

to ion trap tandem mass spectrometry. Bee silk was processed similarly (see Supplemental data for details). Mass spectral data sets from the entire experiment were analyzed using Agilent's Spectrum Mill software to match the data with predictions of protein sequences from the bee. Unless otherwise indicated, protein matches were based on more than one peptide match and scores >20 (the default setting for valid matches by the Spectrum Mill software).

Alignment of the *AmelFibroin* sequences

The *AmelFibroin* sequences show composition bias and the proteins have low complexity and obvious secondary structural features. Conventional alignment algorithms including CLUSTAL (Chenna et al. 2003), T-COFFEE (Notredame et al. 2000), POY (Wheeler et al. 2003), DIALIGN-T (Subramanian et al. 2005), POA (Lee et al. 2002), ProbCons (Do et al. 2005), and MUSCLE (Edgar 2004) gave unconvincing alignments in the heptad regions. For example, they either failed to maintain the protein heptad patterns, or failed to distinguish clearly among alternative alignments in which sequences were shifted relative to each other by whole heptads or, in the case of DNA-level alignments, failed to maintain the reading frames. We therefore developed a manual method, assisted by a MARCOIL heptad prediction for each sequence, on which alignments could be scored to identify the most likely alignment. In this method, indels which broke the reading frame or disrupted the heptad structure were not allowed except where MARCOIL identified alternative, interleaved heptads. When interleaved heptads were predicted, an additional class of indels was allowed, which shifted the alignment from one heptad substructure to another. Alignments were evaluated by counting the number of identical nucleotides in the first and second codon positions. Third position nucleotides were ignored on the grounds that they were likely to have changed too quickly across evolutionary time for meaningful comparisons. Pairwise comparison of random sequences with identical base composition give approximately three matches every 14 nt. We accepted alignments if, on average, four or more nucleotides over the same length matched.

Initially *AmelFibroin*3 and *AmelFibroin*4 were aligned as MARCOIL gave only one heptad prediction for the proteins encoded by these sequences. *AmelFibroin*2 was then brought into the alignment, followed by *AmelFibroin*1, choosing the highest-scoring alignment corresponding to any of the three subset heptad translations predictions in those sequences (Supplemental Table III). *AmelFibroin-rel* could not be aligned using this method.

Acknowledgments

We thank the Baylor College of Medicine Human Genome Sequencing Center for making the *Apis mellifera* and *Tribolium castaneum* gene sequences publicly available before publication. We acknowledge the financial support of the Grains Research and Development Corporation. We also thank Stephen Trowell for advice on the manuscript, Dennis Anderson for the supply of bees and fascinating discussions on their biology, and John Trueman for invaluable advice and help with alignment analysis.

References

Akai, H., Imai, T., and Tsubouchi, K. 1987. Fine-structural changes of liquid silk in silk gland during the spinning stage of *Bombyx* larvae. *J. Seric. Sci. Jpn.* **56:** 131–137.

Atkins, E.D.T. 1967. A four-strand coiled coil model for some insect fibrous proteins. *J. Mol. Biol.* **24:** 139–141.

Bendtsen, J.D., Nielsen, H., von Heijne, G., and Brunak, S. 2004. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **340:** 783–795.

Castillo-Davis, C.I., Mekhedow, S.L., Hartl, D.L., Koonin, E.V., and Kondrashow, F.A. 2002. Selection for short introns in highly expressed genes. *Nat. Genet.* **31:** 415–418.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., and Thompson, J.D. 2003. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* **31:** 3497–3500.

Compton, S.J. and Jones, C.G. 1985. Mechanism of dye response and interference in the Bradford protein assay. *Anal. Biochem.* **151:** 369–374.

Delorenzi, M. and Speed, T. 2002. An HMM model for coiled coil domains and a comparison with PSSM-based predictions. *Bioinformatics* **18:** 617–625.

Do, C.B., Mahabhashyam, M.S.P., Brudno, M., and Batzoglou, S. 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res.* **15:** 330–340.

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

Garel, A., Deleage, G., and Prudhomme, J.C. 1997. Structure and organization of the *Bombyx mori* Sericin 1 gene and of the Sericins 1 deduced from the sequence of the Ser 1B cDNA. *Insect Biochem. Mol. Biol.* **27:** 469–477.

Hepburn, H.R. and Kurstjens, S.P. 1988. The combs of honeybees as composite materials. *Apidologie* **19:** 25–36.

Jay, S.C. 1964. The cocoon of the honey bee, *Apis mellifera* L. *Can. Entomol.* **96:** 784–792.

Julien, E., Coulon-Bublex, M., Garel, A., Royer, C., Chavancy, G., Prudhomme, J.C., and Couble, P. 2004. Silk gland development and regulation of silk protein genes. In *Comprehensive molecular insect science* (eds. L. Gilbert et al.), Vol. 2, pp. 369–384. Pergamon Press, Oxford.

Kneller, D.G., Cohen, F.E., and Langridge, R. 1990. Improvements in protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol.* **214:** 171–182.

Krammerer, R.A. 1997. α-helical coiled coil oligomerization domains in extracellular proteins. *Matrix Biol.* **15:** 555–565.

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

Lee, C., Grasso, C., and Sharlow, M. 2002. Multiple sequence alignment using partial order graphs. *Bioinformatics* **18:** 452–464.

Liu, J. and Lu, M. 2002. An alanine-zipper structure determined by long range intermolecular interactions. *J. Biol. Chem.* **277:** 48708–48713.

Lucas, F. and Rudall, K.M. 1968. Extracellular fibrous proteins: The silks. In *Comprehensive biochemistry* (eds. M. Florkin and E.H. Stotz), Vol. 26B, pp. 475–558. Elsevier, Amsterdam.

Lupas, A.N. and Gruber, M. 2005. The structure of α-helical coiled coils. In *Fibrous proteins: Coiled coils, collagen and elastomers* (eds. D.A.D. Parry and J.M. Squire), pp. 37–78. Elsevier Academic Press, San Diego, California.

McClelland, J.L. and Rumelhart, D.E. 1988. *Explorations in parallel distributed processing*. MIT Press, Cambridge, MA.

Notredame, C., Higgins, D., and Heringa, J. 2000. T-Coffee: A novel method for multiple sequence alignments. *J. Mol. Biol.* **302:** 205–217.

Nunes, F.M.F., Valente, V., Sousa, J.F., Cunha, M.A.V., Pinheiro, D.G., Maia, R.M., Araujo, D.D., Costa, M.C.R., Martins, W.K., Carvalho, A.F., et al. 2004. The use of open reading frame ESTs (ORESTES) for analysis of the honey bee transcriptome. *BMC Genomics* **5:** 84.

Rost, B. and Sander, C. 1993. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232:** 584–599.

Rudall, K.M. 1962. Silk and other cocoon proteins. In *Comparative biochemistry* (eds. M. Florkin and H.S. Mason), pp. 397–433. Academic Press, New York.

Silva-Zacarin, E.C., Silva de Moraes, R.L., and Taboga, S.R. 2003. Silk formation mechanisms in the larval salivary glands of *Apis mellifera*. *J. Biosci.* **28:** 753–764.

Steinert, P.M., Chou, Y.H., Prahlad, V., Parry, D.A., Marekov, L.N., Wu, K.C., Jang, S.I., and Goldman, R.D. 1999. A high molecular weight intermediate filament-associated protein in BHK-21 cells is nestin, a type VI intermediate filament protein. Limited co-assembly in vitro to form heteropolymers with type III vimentin and type IV α-internexin. *J. Biol. Chem.* **274:** 9881–9890.

Stenoien, H.K. and Stephan, W. 2005. Codon mRNA stability is not associated with levels of gene expression in *Drosophila melanogaster* but shows a negative correlation with codon bias. *J. Mol. Evol.* **61:** 306–314.

Subramanian, A.R., Weyer-Menkhoff, J., Kaufmann, M., and Morgenstern, B. 2005. DIALIGN-T: An improved algorithm for segment-based multiple sequence alignment. *Bioinformatics* **6:** 66.

Swofford, D.L. 2002. *PAUP*. Phylogenetic analysis using parsimony (*and*

other methods). Sinauer Associates, Sunderland, Massachusetts.
Takahashi, Y., Gehoh, M., and Yuzuriha, K. 1999. Structure refinement and diffuse streak scattering of silk (*Bombyx mori*). *Int. J. Biol. Macromol.* **24:** 127–138.
Versteeg, R., van Schaik, B.D.C., van Batenburg, M.F., Roos, M., Monajemi, R., Caron, H., Harmen, J., Bussemaker, H.J., and van Kampen, A.H.C. 2003. The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res.* **13:** 1998–2004.
Wheeler, W.C., Gladstein, D.S., and De Laet, J. 2003. POY, Phylogeny reconstruction via direct optimization of DNA and other data. Version 3.0. http://research.amnh.org/scicomp/projects/poy.php
Woolfson, D.N. 2005. The design of coiled coil structures and assemblies. In *Fibrous proteins: Coiled coils, collagen and elastomers* (eds. D.A.D. Parry and J.M. Squire), pp. 79–112. Elsevier Academic Press, San Diego, California.
Yamada, H., Nakao, H., Takasu, Y., and Tsubouchi, K. 2001. Preparation of undegraded native molecular fibroin solution from silkworm cocoons. *Mater. Sci. Eng. C* **14:** 41–46.
Zhou, C.Z., Confalonieri, F., Medina, N., Zivanovic, Y., Esnault, C., Yang, T., Jacquet, M., Janin, J., Duguet, M., Perasso, R., et al. 2000. Fine organization of *Bombyx mori* fibroin heavy chain gene. *Nucleic Acids Res.* **28:** 2413–2419.

Received December 14, 2005; accepted in revised form February 23, 2006.