Ancient Properties of Spider Silks Revealed by the Complete Gene Sequence of the Prey-Wrapping Silk Protein (AcSp1)

Nadia A. Ayoub,*[,1,2] Jessica E. Garb,[3] Amanda Kuelbs,[2] and Cheryl Y. Hayashi[2]

[1]Department of Biology, Washington and Lee University

[2]Biology Department, University of California, Riverside

[3]Department of Biological Sciences, University of Massachusetts, Lowell

***Corresponding author:** E-mail: ayoubn@wlu.edu.

Associate editor: Willie Swanson

Sequences generated in this study were submitted to GenBank (accession nos. JX978170-JX978182).

Abstract

Spider silk fibers have impressive mechanical properties and are primarily composed of highly repetitive structural proteins (termed spidroins) encoded by a single gene family. Most characterized spidroin genes are incompletely known because of their extreme size (typically >9 kb) and repetitiveness, limiting understanding of the evolutionary processes that gave rise to their unusual gene architectures. The only complete spidroin genes characterized thus far form the dragline in the Western black widow, *Latrodectus hesperus*. Here, we describe the first complete gene sequence encoding the aciniform spidroin AcSp1, the primary component of spider prey-wrapping fibers. *L. hesperus AcSp1* contains a single enormous (~19 kb) exon. The AcSp1 repeat sequence is exceptionally conserved between two widow species (~94% identity) and between widows and distantly related orb-weavers (~30% identity), consistent with a history of strong purifying selection on its amino acid sequence. Furthermore, the 16 repeats (each 371–375 amino acids long) found in black widow AcSp1 are, on average, >99% identical at the nucleotide level. A combination of stabilizing selection on amino acid sequence, selection on silent sites, and intragenic recombination likely explains the extreme homogenization of *AcSp1* repeats. In addition, phylogenetic analyses of spidroin paralogs support a gene duplication event occurring concomitantly with specialization of the aciniform glands and the tubuliform glands, which synthesize egg-case silk. With repeats that are dramatically different in length and amino acid composition from dragline spidroins, our *L. hesperus* AcSp1 expands the knowledge base for developing silk-based biomimetic technologies.

Key words: aciniform silk, concerted evolution, full-length gene, *Latrodectus hesperus*, spidroin, Western black widow.

Introduction

Spiders (Araneae) rely on silk throughout their lifetime and are unparalleled in the diversity of silks they can synthesize. A single orb-web weaving spider (Orbiculariae, fig. 1) possesses seven types of specialized abdominal silk glands. Each gland type produces a different silk fiber or glue that has a unique function (Foelix 2010). For example, major ampullate glands produce the dragline, tubuliform glands synthesize large diameter egg-case silk fibers, capture spiral threads of the orb-web originate in the flagelliform glands, and prey-wrapping silk is synthesized in aciniform glands. Diversity of silk function is paralleled by diversity of silk mechanical properties (Gosline et al. 1999; Blackledge and Hayashi 2006). Dragline silk approaches the tensile strength of steel and capture-spiral fibers can stretch more than one and a half times their original length, an order of magnitude greater than dragline fibers (Denny 1976; Gosline et al. 1999; Blackledge and Hayashi 2006). In garden orb-weavers (*Argiope argentata* and *A. trifasciata*), prey-wrapping silk combines high extensibility with tensile strength to form a fiber that is twice as tough as the dragline of those species (Hayashi et al. 2004; Blackledge and Hayashi 2006) and is one of the toughest silks measured thus far for any species (Agnarrson et al. 2010).

Differences among silks in mechanical properties derive in large part from the differences in protein composition of each fiber type (Hayashi and Lewis 1998, 2001; Gosline et al. 1999; Hayashi et al. 1999). Spider silk fibers are primarily composed of one or more unique structural proteins termed spidroins (a contraction of "spider fibroin"), which are members of a single gene family (Guerette et al. 1996; Gatesy et al. 2001). In orb-weaving spiders, different gland types secrete different spidroins. For instance, major ampullate glands express the dragline spidroins, MaSp1 and MaSp2 (Xu and Lewis 1990; Hinman and Lewis 1992; Sponner et al. 2005) and aciniform glands synthesize AcSp1 (Hayashi et al. 2004). Thus, spider silks could be used in a plethora of biomimetic applications that capitalize on transgenic technology (Sponner 2007). Furthermore, spider silks represent a spectacular example of functional diversification via gene duplication followed by sequence and expression divergence.

Efforts to understand the molecular evolution of spidroins and create recombinant spider silks have been hampered by the difficulty in characterizing complete spidroin-encoding sequences. Thus far, only two complete spidroin genes have been described: *MaSp1* and *MaSp2* of the Western black widow, *Latrodectus hesperus* (Ayoub, Garb, Tinghitella, et al. 2007). Spidroins are extremely large proteins (200–350 kDa;

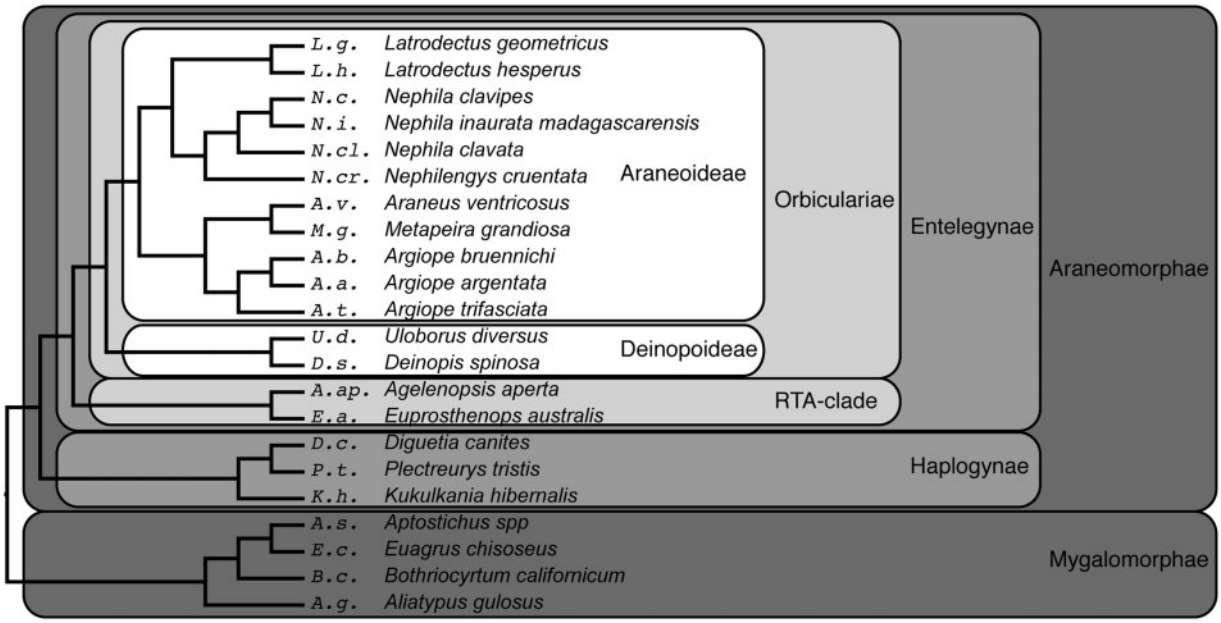

Fig. 1. Relationships among spider species used in this study based on hypotheses of Scharff and Coddington (1997), Coddington (2005), Ayoub, Garb, Hedin, et al. (2007), Kuntner et al. (2008), and Elices et al. (2009). Species abbreviations precede the species names. Major taxonomic groups are bracketed.

e.g., Hayashi et al. 1999; Sponner et al. 2005; Ayoub, Garb, Tinghitella, et al. 2007; Vasanthavada et al. 2007) made up almost entirely of repeated blocks of amino acids (aa) flanked by short (<150 aa) nonrepetitive amino (N)- and carboxy (C)-terminal domains. The N- and C-terminal domains of spidroins are conserved in length and share amino acid signatures across gene family members (e.g., Guerette et al. 1996; Gatesy et al. 2001, Motriuk-Smith et al. 2005; Garb et al. 2010). Although these terminal domains are involved in functions general to spidroins (e.g., fiber assembly, Ittah et al. 2006; Askarieh et al. 2010; Hagn et al. 2010; Eisoldt et al. 2012), the repetitive regions are thought to be responsible for the variation in mechanical properties of different fiber types (Gosline et al. 1999; Hayashi et al. 1999). Secondary structures, such as beta-pleated sheets or beta-turns, have been predicted for a few simple amino acid sequence motifs that are common to a subset of spidroins (Hayashi et al. 1999; Holland, Creager, et al. 2008; Holland, Jenkins, et al. 2008; Jenkins, Creager, Butler, et al. 2010; Jenkins, Creager, Lewis, et al. 2010). Some spidroins, such as MaSp1, MaSp2, and Flag (capture-spiral spidroin), string together a subset of these simple motifs to form a unit called an ensemble repeat, which can then be repeated in the tens to hundreds of times within one spidroin molecule (e.g., Gatesy et al. 2001; Ayoub, Garb, Tinghitella, et al. 2007). Other spidroins, however, such as AcSp1, have longer and more complex repeat units with only a few of these simple motifs (Gatesy et al. 2001; Hayashi et al. 2004; Garb and Hayashi 2005; Garb et al. 2007; Starrett et al. 2012).

Although much work has focused on spidroins with simple repeats, few studies have investigated spidroins with long, complex repeats like AcSp1 (Rising et al. 2011). AcSp1 of *A. trifasciata* possesses over fourteen 200 aa long repeats that are virtually identical to each other (Hayashi et al. 2004). Although all spidroins show high levels of identity among ensemble repeats (e.g., Gatesy et al. 2001; Garb and Hayashi 2005; Ayoub, Garb, Tinghitella, et al. 2007; Garb et al. 2007; Starrett et al. 2012), *A. trifasciata AcSp1* is exceptional with an average of 99.9% pairwise identity among repeats at the nucleotide level (Hayashi et al. 2004). Although partial *AcSp1* cDNAs have been described from the cob-web weaving Western black widow (Vasanthavada et al. 2007), and the feather-legged orb-weaver *Uloborus diversus* (Garb et al. 2006), these cDNAs were too short to evaluate the generality of extreme homogeneity among repeats. Intragenic concerted evolution has been cited as the process that homogenizes repeats of various spidroins, but it is unclear why *A. trifasciata* AcSp1 is dramatically more homogenized than other spidroin paralogs. Characterization of a complete *AcSp1* gene from another species would address if extreme homogeneity of intragenic repeats is a general property of AcSp1 and provide insight about the molecular evolution of the complex repeat unit.

Aciniform spidroins may prove instrumental in deciphering the history of spider silk gene duplications. The number of functionally specialized silk gland types is positively correlated with number of spidroin paralogs (Garb et al. 2007). This association is consistent with the hypothesis that gland types and spidroins have co-evolved (Hayashi and Lewis 1998). Virtually all spiders possess spherical to pear-shaped glands that are similar in structure to the aciniform or pyriform glands of the superfamily Orbiculariae (fig. 1, Shultz 1987). If these simple acinous-shaped structures represent the ancestral gland type, as proposed by Shultz (1987), we predict aciniform spidroins would be recovered in a basal phylogenetic position relative to other spidroins. Shultz (1987)

also proposed that tubuliform glands are specialized aciniform glands. We thus predict that spidroins expressed in aciniform and tubuliform glands will be closely related. Testing these predictions requires reconstructing spidroin gene trees from the nonrepetitive terminal domains due to challenges associated with determining positional homology among repetitive sequences (e.g., Gatesy et al. 2001). Complete spidroin encoding sequences double the amount of phylogenetic information relative to partial N- or C-terminal sequences and ensure that the two domains are accurately associated (Garb et al. 2010).

In this study, we report a complete *AcSp1* gene from the Western black widow, *L. hesperus*, identified from a fully sequenced 39 kb region of genomic DNA and partial *AcSp1* gene sequences from the brown widow, *L. geometricus* (Theridiidae). These data were used to address three related goals: 1) to test whether extreme homogeneity of intragenic sequence repeats is a general property of AcSp1, 2) to evaluate whether spidroins co-evolved with glandular specialization, specifically testing the hypotheses that AcSp1 has a basal position in spidroin gene trees and groups with tubuliform spidroins, and 3) to determine whether the gene structure of *AcSp1* is similar to *MaSp1* and *MaSp2* in lacking introns, making it an expedient template for the production of recombinant silk proteins. Our results show that black widow *AcSp1* is composed of a single enormous exon (18,999 bases of coding sequence). We demonstrate that extreme homogeneity of intragenic repeats is a general feature of AcSp1, with sequence repeats being exceptionally conserved between *Latrodectus* species and among orbicularians. This level of conservation and intragenic homogenization may be explained by the combined forces of stabilizing selection and intragenic concerted evolution having acted on the *AcSp1* gene. Finally, our phylogenetic analyses of spidroin paralogs demonstrate novel support for co-evolution of gene duplications with glandular specialization by uncovering a close relationship between AcSp1 and tubuliform spidroins. In addition to these contributions to understanding the molecular evolution of spidroins, our full-length AcSp1 provides a complete genetic blueprint for biomimetic applications that capitalize on the extreme toughness of aciniform silk fibers.

Results

We sequenced and assembled 39,269 base pairs (bp) of *L. hesperus* genomic DNA (JX978171), including a complete open reading frame (ORF) that is 18,999 bp in length and predicted to encode a 6,332 aa AcSp1. This is the longest coding region described for any spidroin. The most abundant amino acids in AcSp1 are alanine (15.1%), serine (13.1%), and glycine (11.3%) (fig. 2). Despite the prevalence of alanine and glycine, the amino acid motifs poly-A, GGX, GPG, and poly-GA that dominate the MaSp1 and MaSp2 dragline spidroins (Ayoub, Garb, Tinghitella, et al. 2007), are absent or rare in AcSp1 (supplementary fig. S1, Supplementary Material online). Furthermore, the amino acid composition of AcSp1 is much more evenly distributed than in MaSp1 or MaSp2, for which more than 60% of the protein is made

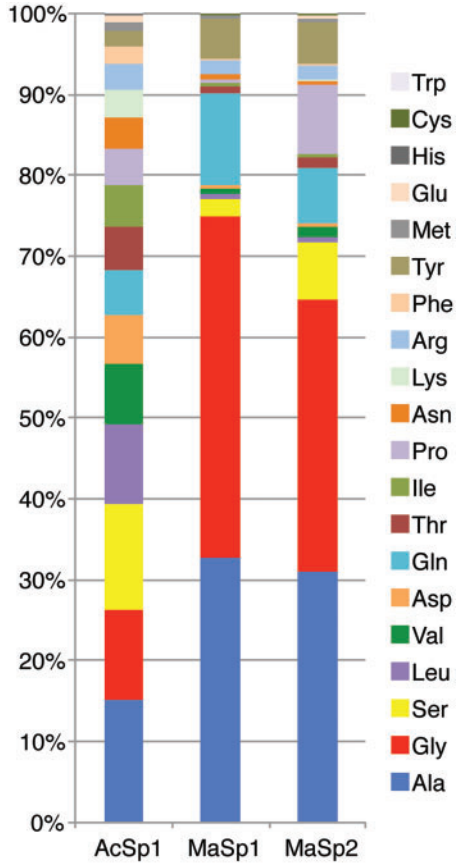

FIG. 2. Amino acid compositions of complete *Latrodectus hesperus* AcSp1, MaSp1, and MaSp2. Three letter amino acid abbreviations are used.

up of alanine or glycine (fig. 2). As has been observed with other spidroin genes, codon usage in *AcSp1* for alanine, glycine, threonine, and proline is skewed toward codons that end in adenine or thymine (77.2% of alanine codons, 82.1% of glycine codons, 72% of threonine codons, 82.4% of proline codons; supplementary table S1, Supplementary Material online).

AcSp1 alternates between hydrophobic and hydrophilic regions (range of Kyte–Doolittle hydrophilicity = −3.4 to 3.3) and on average is slightly hydrophobic (average Kyte–Doolittle hydrophilicity = −0.16). The shifts from hydrophilic to hydrophobic regions in the N- and C-termini are very similar to those seen in *L. hesperus* MaSp1, but the repetitive region of AcSp1 does not display the consistent shift seen in MaSp1 between hydrophilic glycine-rich and hydrophobic alanine-rich sequences (supplementary fig. S2, Supplementary Material online).

Partial sequencing of four additional genomic clones containing *AcSp1* (JX978172–JX978175) revealed that they were very similar to each other and to sequences amplified from four individual spiders (JX978176–JX978179). Average uncorrected *p*-distance across the region sequenced from all clones and individuals was 1.3% (0.077–2.9%). The previously described *L. hesperus* "*AcSp1*-like" cDNA sequence was only 1.2% different from our completely sequenced clone.

Homogenization and Conservation of AcSp1

Our *L. hesperus AcSp1* includes 15 iterations of a 1,125 bp repeat and one, slightly shorter repeat variant of 1,113 bp. There is also very little sequence variation among repeats (supplementary fig. S1, Supplementary Material online). At the amino acid level, the average intragenic pairwise difference among repeats was just 0.68% for *L. hesperus* (range = 0–2.4%). AcSp1 intragenic repeats are similarly homongenized in three species of Araneidae: 0.62% for *A. trifasciata* (range = 0–1.9%), 2.1% for *A. amoena* (range = 0–3.3%), and 2.4% for *Araneus ventricosus* (range = 0–6.1%). Differences between repeats from closely related species were always higher than intragenic differences. The lowest pairwise difference between species was for the comparison of *L. hesperus* repeats to the *L. geometricus* consensus repeat (average = 5.9%, range = 5.6–7.0%; JX978182). Average pairwise differences for the araneid repeats were 26.0% between *Aran. ventricosus* and *A. amoena*, 23.0% between *Aran. ventricosus* and *A. trifasciata*, and 18.2% between *A. amoena* and *A. trifasciata*.

Homogeneity of *Latrodectus* AcSp1 repeats is maintained at the nucleotide level (average pairwise difference = 0.58%, range = 0–2.2%). This level of intragenic nucleotide divergence among repeats is even lower than observed in other *L. hesperus* spidroins, such as *TuSp1* (average = 2.9%, range = 0.18–7.1%) and *MaSp1* (average = 2.4%, range = 0.28–6.3%). We tested whether homogeneity of *L. hesperus* AcSp1 intragenic repeats could result solely from constraints on amino acid sequence by calculating the pairwise number of synonymous substitutions per synonymous sites (Ks) and nonsynonymous substitutions per nonsynonymous sites (Ka) between intragenic repeats. The extremely low divergence among *L. hesperus AcSp1* repeats is due to both very few nonsynonymous and synonymous substitutions (average Ka = 0.0022, average Ks = 0.0081; fig. 3). Both types of substitutions are five to nine times lower for intragenic *AcSp1* repeats compared with intragenic *MaSp1* (average Ka = 0.011, average Ks = 0.057) and *TuSp1* (average Ka = 0.017, average Ks = 0.076; fig. 3).

The AcSp1 repeats are more conserved than the adjacent nonrepetitive C-terminus in *Latrodectus* (9.1% difference between C-termini vs. 5.9% average pairwise difference between repeats of *L. hesperus* and *L. geometricus*) but less conserved in the araneids (14.1% difference between C-termini vs. 23.0% average pairwise difference between repeats of *Aran. ventricosus* and *A. trifasciata*). At the nucleotide level, *AcSp1* repeats are also highly conserved between *L. hesperus* and *L. geometricus* relative to interspecific comparisons of adjacent N- and C-termini and *TuSp1* (fig. 3). We compared selective pressures on *AcSp1* repeats with adjacent N- and C-terminal encoding regions and repeats of paralogous spidroins. Nonsynonymous substitutions between *L. hesperus* and *L. geometricus* are similarly low for *AcSp1* repeats (Ka = 0.026), *TuSp1* repeats (Ka = 0.033), N-termini of *AcSp1* (Ka = 0.0095), and C-termini of *AcSp1* (Ka = 0.036) and *TuSp1* (Ka = 0.066). However, synonymous substitutions in *AcSp1* repeats appear to be severely suppressed (fig. 3). Average interspecific

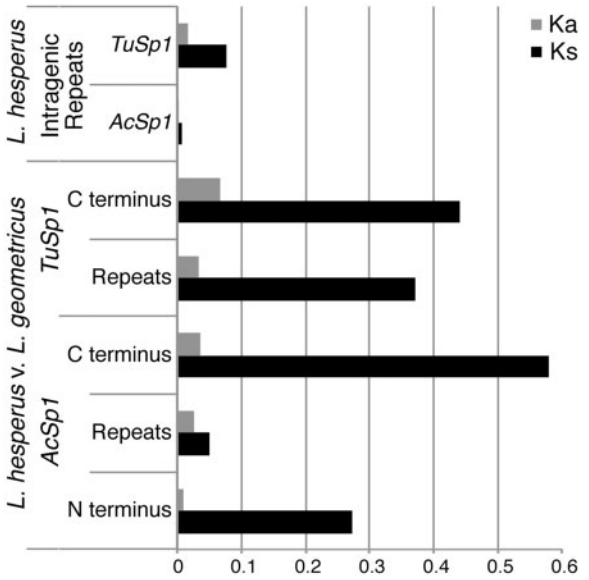

Fig. 3. Pairwise values of nonsynonymous substitutions per nonsynonymous sites (Ka; gray) and synonymous substitutions per synonymous sites (Ks; black) for intragenic comparisons among *Latrodectus hesperus* spidroin repeats or intergenic comparisons between *L. hesperus* and *L. geometricus* for corresponding regions of spidroins. Values shown for repeats are averaged across all pairwise comparisons.

synonymous substitutions between *AcSp1* repeats (Ks = 0.050) is 11 times lower than the adjacent C-termini (Ks = 0.58), five times lower than the adjacent N-termini (Ks = 0.27), and seven times lower than the repeats of *TuSp1* (Ks = 0.37).

AcSp1 repeats are also conserved among more distantly related species (supplementary figs. S3 and S4, Supplementary Material online). The repeat length of *L. hesperus* (375 aa) is similar to the feather-legged spider, *U. diversus* (357 aa) but almost twice as long as the araneid AcSp1 repeats (200–215 aa). BLASTP produced significant alignments between the araneid AcSp1 repeats and the first and second halves of the *L. hesperus* and *U. diversus* AcSp1 repeats. We thus split the *L. hesperus* and *U. diversus* consensus repeats into halves and aligned the two parts with the araneid AcSp1 repeats (supplementary figs. S3 and S4, Supplementary Material online). Phenetic (neighbor joining) and phylogenetic (MP) clustering with mid-point rooting indicated that the first and second parts of the *Latrodectus* repeats are more similar to each other than to the repeats of other species. In contrast, the second half of *U. diversus* is more similar to araneid repeats than to the first part of the *U. diversus* repeat (supplementary fig. S3*B*, Supplementary Material online).

BLASTP searches of the NCBI nr protein database additionally recognized significant similarity of the *L. hesperus* AcSp1 repeat with other spidroin paralogs ($E < 10^{-10}$ compared with $E < 10^{-26}$ for AcSp1 orthologs). These included TuSp1 (egg-case) spidroins as well as other spidroins with complex repeats characterized from araneomorphs and mygalomorphs. However, based on the nonrepetitive terminal regions, these spidroins do not form a monophyletic group (figs. 4 and 5).

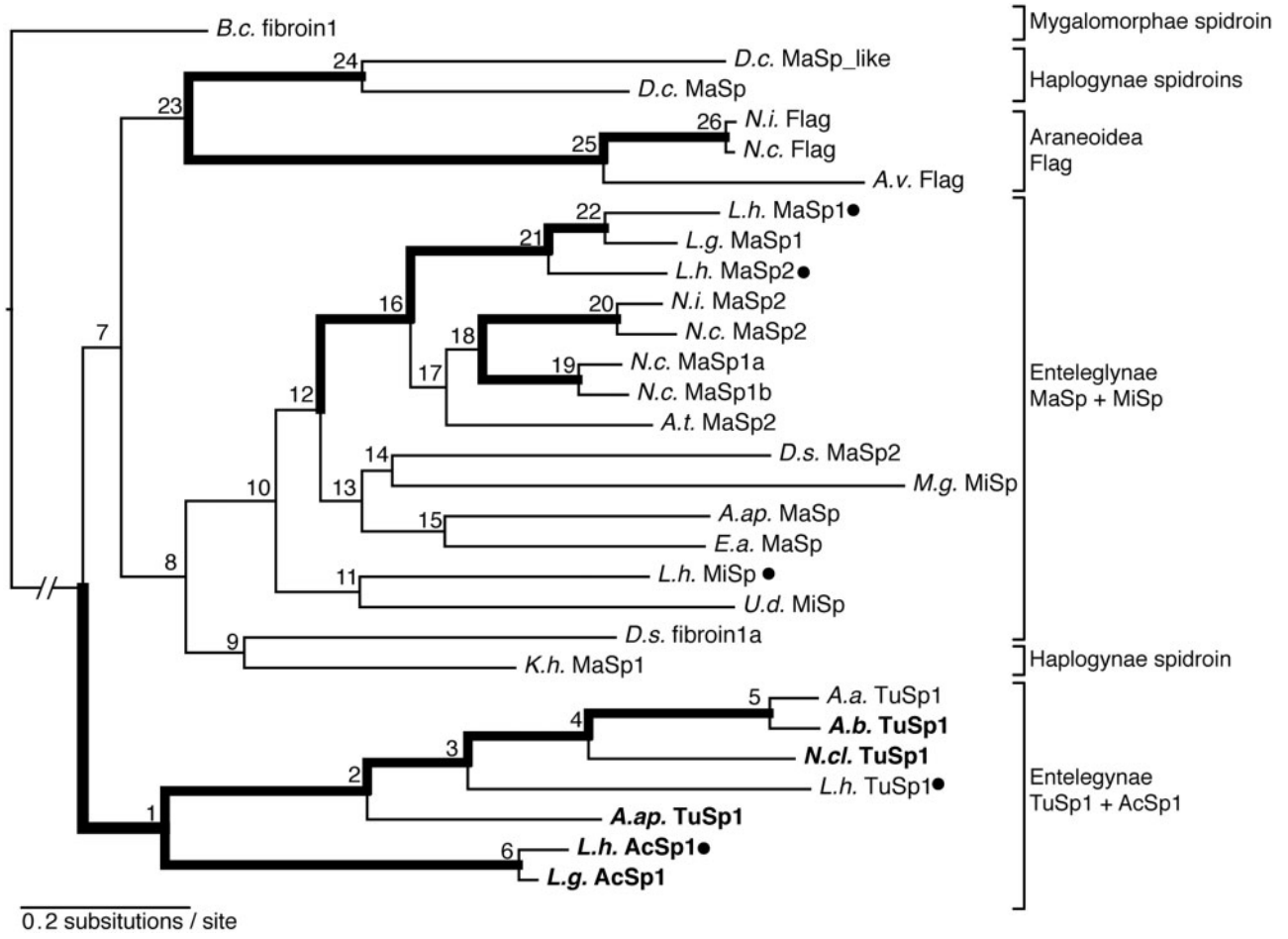

Fig. 4. Maximum likelihood tree of combined spidroin N- and C-terminal encoding regions. Species abbreviations are defined in figure 1 and supplementary table S2, Supplementary Material online. Thickened branches indicate relationships supported by >50% MP bootstrap replicates and >0.95 Bayesian posterior probability for both amino acids and nucleotides. Support values for numbered nodes are shown in supplementary table S3, Supplementary Material online. Rooting with the mygalomorph spidroin, *B.c.* fibroin1, resulted in the fewest inferred duplications and losses. Slashed lines indicate that the branch to *B.c.* fibroin1 was arbitrarily shortened. Dots indicate *L. hesperus* paralogs. Spidroins in bold had significant similarity to the *L. hesperus* AcSp1 repeat according to BLASTP. Major clades of spidroins are bracketed.

We identified only a few conserved sequences flanking *AcSp1*, which included a TATA box and a four base motif (CACG) identified upstream of other spidroin genes (Motriuk-Smith et al. 2005; Ayoub, Garb, Tinghitella, et al. 2007). Other conserved sequences were part of transposable elements (supplementary methods and results and supplementary fig. S5, Supplementary Material online).

Relationship of AcSp1 to Other Spidroins

Consistent with the hypothesis that spidroins coevolved with glandular specialization, we found phylogenetic support for a sister relationship between AcSp1 and the spidroin expressed in tubuliform glands, TuSp1 (figs. 4 and 5). Parsimony, likelihood, and Bayesian analyses of combined N- and C-terminal amino acids and encoding nucleotides for 29 spidroins (supplementary table S2, Supplementary Material online) grouped *Latrodectus* AcSp1 with a clade of TuSp1s with strong support (fig. 4; supplementary fig. S6 and table S3, Supplementary Material online). This result contrasts with the weakly supported grouping of Flag with TuSp1 in the most comprehensive phylogenetic analysis of spidroin paralogs prior to this one, which did not include AcSp1 (Garb et al. 2010). In our analyses, Flag consistently grouped with *Diguetia canities* MaSp-like sequences, but with poor support (fig. 4 and supplementary fig. S6, Supplementary Material online). Support for grouping AcSp1 with TuSp1 largely derived from N-terminal amino acids and nucleotides (partitioned decay index = 5 for N-terminal amino acids or nucleotides, partitioned decay index = 0 or −3 for C-terminal amino acids and nucleotides, respectively; Baker and DeSalle 1997).

Other relationships among spidroins were consistent with patterns described by Garb et al. (2010); the 14 well-supported nodes (e.g., >0.95 posterior probability, *pp*) out of a total of 23 nodes from Garb et al. (2010) were also recovered with strong support in our analyses. For instance, we found strong support for a monophyletic group of TuSp1 present in orbicularians and a member of the RTA-clade, monophyletic araneoid Flag, and monophyletic araneoid MaSp1 and MaSp2 (fig. 4; supplementary fig. S6 and table S3, Supplementary Material online).

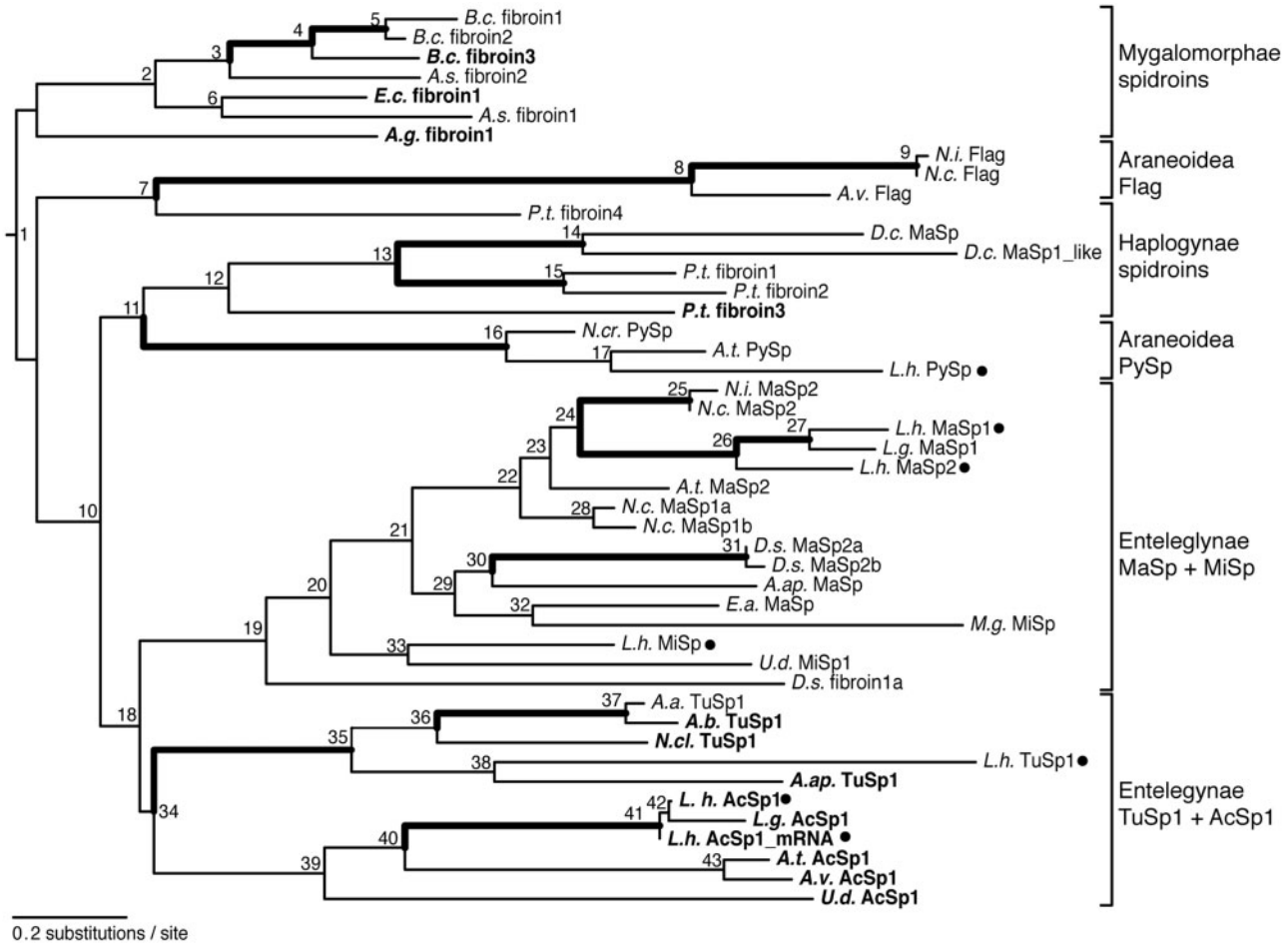

Fig. 5. Maximum likelihood tree of expanded dataset of C-terminal region encoding nucleotides. Species abbreviations are defined in figure 1 and supplementary table S2, Supplementary Material online. Thickened branches indicate that the node is supported by >50% MP bootstrap replicates and >0.95 Bayesian posterior probability for both amino acids and nucleotides. Support values for numbered nodes are shown in supplementary table S4, Supplementary Material online. Rooting with the mygalomorph spidroins resulted in the fewest inferred duplications and losses. Dots indicate *L. hesperus* paralogs. Spidroins in bold had significant similarity to the *L. hesperus* AcSp1 repeat according to BLASTP. Major clades of spidroins are bracketed.

Expanding the C-terminal data set to include 46 spidroins (supplementary table S2, Supplementary Material online) recovered a monophyletic group of orbicularian AcSp1 (>0.95 *pp*) in all analyses except parsimony searches of amino acids (fig. 5; supplementary table S4 and fig. S7, Supplementary Material online). Grouping AcSp1 with TuSp1 was additionally recovered in nucleotide ML and amino acid Bayesian analyses, albeit with low support (<0.95 *pp*). Well-supported relationships described earlier were additionally recovered with high support (>0.95 *pp* or >75% of MP bootstrap replicates) by the expanded data set, including monophyletic TuSp1, monophyletic araneoid MaSp1 and MaSp2, and monophyletic Flag. C-terminal sequences also strongly supported monophyletic mygalomorph spidroins (excluding *Aliatypus gulosus* fibroin1) and monophyletic araneoid PySp (spidroin in attachment cement fibers). In contrast to the poorly supported grouping of Flag with *D. canities* MaSp-like sequences in the smaller (29 spidroin) data set, the larger (46 spidroin) C-terminal dataset grouped *D. canities* MaSp-like sequences with two spidroins described from *Plectreurys tristis* with strong support (>0.95 *pp*, >60% of MP bootstrap replicates of nucleotides), except that this relationship was not recovered by MP analysis of amino acids (fig. 5; supplementary table S4 and fig. S7, Supplementary Material online). *D. canities* and *P. tristis* are representatives of the araneomorph clade Haplogynae, which is divergent from the orbicularians (fig. 1).

Species tree—gene tree reconciliation supported rooting with the mygalomorph spidroin, *B.c.* fibroin1, for the phylogenetic trees based on both N- and C-terminal data. This rooting resulted in the lowest duplication/loss scores (ML: 14 duplications, 51 losses; amino acid MP: 13 duplications, 44 or 47 losses depending on MP tree; nucleotide MP: 13 duplications, 47 losses; amino acid Bayes: 14 duplications, 43 losses; nucleotide Bayes: 13 duplications, 44 losses). Root placement for the expanded C-terminal data set varied among trees. For the ML and nucleotide Bayes trees, rooting with all the mygalomorph spidroins resulted in the lowest duplication/loss scores (ML: 21 duplications and 61 losses; nucleotide Bayes: 20 duplications, 49 losses). However, the

MP trees based on amino acids possessed eight equal root placements (supplementary fig. S7, Supplementary Material online, 23 duplications and 74 losses). The nucleotide MP tree rooted with a clade of spidroins found in divergent taxa (*A.g.* fibroin1 from a mygalomorph, *P.t.* fibroin4 from a haplogyne araneomomorph, and Flag from an orbicularian araneomorph) resulted in the lowest score (supplementary fig. S7, Supplementary Material online; 25 duplications and 76 losses). Finally, the Bayesian tree based on amino acids rooted with a pairing of a mygalomorph and a haplogyne spidroin resulted in fewest duplications (23) and losses (59) (supplementary fig. S7, Supplementary Material online).

Discussion

Our complete black widow *AcSp1* contains the longest spidroin coding sequence (18,999 bp) described thus far. Only four other complete spidroin coding sequences have been previously reported: *CySp1* (9.1 kb) and *CySp2* (9.2 kb; *CySp* is a synonym for *TuSp*) cDNAs from the wasp spider, *A. bruennichi* (Zhao et al. 2006); and *MaSp1* (9.4 kb) and *MaSp2* (11.3 kb) genes from the Western black widow (Ayoub, Garb, Tinghitella, et al. 2007). Our black widow *AcSp1* surpasses all of these at 19 kb, encoding a 6,332 amino acid protein with a remarkably large, predicted molecular weight of 630 kDa. In contrast, gel electrophoresis indicated that black widow AcSp1 is a 300 kDa protein (Vasanthavada et al. 2007). The discrepancy between predicted and observed sizes can be explained by extreme allelic length variation of *AcSp1* in black widows. The near identical (>98% pairwise nucleotide identities) tandem repeats found in *AcSp1* could have facilitated unequal crossing over that resulted in the rapid loss or gain of repeats. There is precedent for huge size variation in a spider silk gene. Chinali et al. (2010) documented that *MaSp1* from 100 individual golden orb-weavers (*Nephila clavipes*) can range from 10 to 17.5 kb. The discrepancy in observed and predicted sizes could also be attributed to various posttranscriptional or translational modifications (e.g., Tran et al. 2011) that have yet to be observed with AcSp1.

Similar to Western black widow *MaSp1* and *MaSp2*, *AcSp1* has a peculiar gene structure because it lacks introns. *AcSp1* stands out with an exon size more than double that of *MaSp1* and *MaSp2* and exceeding the longest exons known in human (17 kb), chimpanzee (11.6 kb), mouse (17.1 kb), zebrafish (12 kb), and the roundworm *Caenorhabditis elegans* (15 kb), and approaching the longest known exon in *Drosophila melanogaster* (27.7 kb) (Peng et al. 2009). Although limited genomic information is available for other species, single exon silk genes may be the rule for widow spiders. Partial sequences of Western black widow and brown widow *MaSp1*, *MaSp2*, *MiSp*, *TuSp1*, and *AcSp1* contain no evidence of introns (Garb and Hayashi 2005; Motriuk-Smith et al. 2005; Ayoub and Hayashi 2008; this study, unpublished data). Partial sequences of *Nephila MaSp2* also lack introns (Motriuk-Smith et al. 2005). In contrast, *Argiope MaSp2* and *Nephila Flag* have multiple introns that are nearly identical within a single gene (Hayashi and Lewis 2000; Motriuk-Smith et al. 2005). Ayoub, Garb, Tinghitella, et al. (2007) noted that single exon genes could reflect a process of gene duplication involving retrotransposition of mRNA transcripts that would necessarily give rise to intronless paralogs. This process of gene duplication could be the dominant mode for the spidroin gene family. However, retrotransposition often results in pseudogenes since the necessary regulatory sequences are not simultaneously duplicated (Zhang 2003). Instead, lack of introns may be the ancestral condition for the spidroin gene family and *Argiope MaSp2* and *Nephila Flag* independently gained introns. Characterization of complete spidroin genes from divergent spider species would clarify whether spider silk genes have experienced multiple gains or losses of introns.

Homogenization and Conservation of AcSp1 Repeats

Our complete gene sequence demonstrates that black widow *AcSp1* repeats are as highly homogenized as *Argiope* and other araneid *AcSp1* repeats. The near identity of intragenic *AcSp1* repeats is unusual even in comparison with other spidroins (e.g., fig. 3). Purifying selection could maintain identity among repeats. Indeed, the rate of silent substitutions exceeds that of amino acid replacements in black widow *AcSp1* (Ka/Ks = 0.22), but similar low values of Ka/Ks were found for black widow *TuSp1* (fig. 3) and *MaSp1* intragenic repeats (Ayoub, Garb, Tinghitella, et al. 2007). Thus, purifying selection on amino acid sequence or stabilizing selection to maintain similar amino acid repeats within a single polypeptide alone cannot explain near identity among repeats. Gene conversion and unequal crossing-over, facilitated by iterated repeats, have been frequently cited as processes that lead to intragenic concerted evolution among spidroin repeats (e.g., Beckwitt et al. 1998; Gatesy et al. 2001; Hayashi et al. 2004; Garb and Hayashi 2005; Ayoub, Garb, Tinghitella, et al. 2007). Here, we discuss how selective constraints, mutation, and concerted evolution could act differently on *AcSp1* compared with other spidroin-encoding genes.

AcSp1 intragenic repeats could be more homogenized than the repeats of paralogous spidroins because they experience lower mutations rates. In spidroins with simple repeats, such as *MaSp1* and *MaSp2*, the tandem repetition of codons for amino acid sequence motifs such as contiguous stretches of alanines, can lead to slip-strand mispairing that results in a higher mutation rate within the repeats (to the extent that repeats cannot be reliably aligned between species as distantly related as *L. hesperus* and *L. geometricus*) than in terminal regions of the genes (Ayoub, Garb, Tinghitella, et al. 2007; similar pattern in *Flag* exons vs. introns, Hayashi and Lewis 2000). *Latrodectus* AcSp1 and TuSp1 repeats have complex amino acid sequences that do not have a high proportion of these simple motifs, and thus are expected to have less localized slip-strand mispairing. In fact, the synonymous substitution rate between *L. hesperus* and *L. geometricus TuSp1* repeats is similar to the adjacent C-terminal encoding region (fig. 3), suggesting similar mutation rates and/or selective constraints across the gene. In contrast, *AcSp1* repeats experienced far fewer interspecific synonymous substitutions than the adjacent N- or C-termini (fig. 3).

A dramatically lower mutation rate in the repetitive versus the terminal-region encoding sequences of AcSp1 seems unlikely considering that the adjacent gene regions are in the same genomic location, have the same base composition (~55% AT), and that both lack the strings of GC-rich sequence associated with higher mutation rates (e.g., Sved and Bird 1990). Codon bias in *AcSp1* could constrain synonymous substitutions, but the tendency toward A or T in the third position of codons is similarly high in black widow *MaSp1* and *MaSp2* (Ayoub, Garb, Tinghitella, et al. 2007) and *TuSp1* (e.g., 100% of Gly and 75% of Ala codons end in A or T; Garb and Hayashi 2005). Instead, stabilizing selection on *AcSp1* repeats for specific mRNA secondary structures that increase mRNA stability, control translation, or prevent degradation could constrain synonymous substitutions (e.g., Katz and Burge 2003; Chamary and Hurst 2005; Meyer and Miklós 2005).

Stochasticity in the process of intragenic concerted evolution probably also contributes to the homogeneity of intragenic *AcSp1* repeats and could explain different patterns of conservation of *AcSp1* repeats among taxonomic groups. In the absence of selective constraints, concerted evolution should increase the apparent rate of interspecific divergence because a single mutation can rapidly proliferate to each of the repeats. Consistent with this hypothesis, araneid (*A. amoena*, *A. trifasciata*, and *Aran. ventricosus*) AcSp1 repeats are less conserved between species than are the adjacent C-terminal regions, a pattern also seen in MaSp1, MaSp2, and Flag of multiple species comparisons (e.g., Beckwitt et al. 1998, Hayashi and Lewis 2000, Ayoub and Hayashi 2008). However, it is also possible for new mutations to be replaced by the ancestral repeat sequence during concerted evolution. By chance, only the latter may have happened in *Latrodectus*. Denser taxonomic sampling of *AcSp1* orthologs both above and below the species level is needed to evaluate the relative roles of selective constraints on synonymous sites, mutation rates, and concerted evolution on the unusual patterns of conservation in *AcSp1* and its extreme intragenic homogeneity.

History of Silk Gene Duplications

Our phylogenetic results are consistent with the hypothesis that spidroins co-evolved with gland specialization, or the glandular affiliation hypothesis proposed by Hayashi and Lewis (1998). The first prediction of this hypothesis is that spidroins expressed in the same type of differentiated silk glands should be orthologous. Within Entelegynae (fig. 1), monophyly of TuSp, Flag, and PySp (fig. 5) and their expression in tubuliform glands, flagelliform glands, and pyriform glands, respectively, lend support to orthology of spidroins with gland-specific expression. C-terminal domains also supported the monophyly of AcSp1 (fig. 5). Furthermore, relationships among AcSp1 C-termini reflect putative species relationships (fig. 1). The deinopoid, *U. diversus*, is sister to a clade of araneoids including monophyletic *Latrodectus* and monophyletic araneid sequences (fig. 5). In the report of a partial *AcSp1* cDNA from *L. hesperus*, Vasanthavada et al. (2007) suggested that their "*AcSp1*-like" might not be orthologous to araneid *AcSp1* and predicted the presence of a second copy of *AcSp1*. The "*AcSp1*-like" cDNA, however, was nearly identical to our completely sequenced *AcSp1* gene and partial *AcSp1* sequences from four additional genomic clones and four individual spiders in both coding and non-coding regions. Inspection of chromatograms from directly sequenced PCR-amplifications of *AcSp1* from individual spiders revealed that double peaks, which can be interpreted as variation within a single genome that corresponds to either two alleles at a single locus (heterozygosity) or more than one locus, were never found at the same site among all four individuals. Thus, we failed to find evidence for a second copy of *AcSp1* in the *L. hesperus* genome. Instead, each of the genomic clones and the previously described cDNA likely represent allelic variants of the same locus.

The glandular affiliation hypothesis also predicts that relationships among spidroin paralogs mirror relationships among gland types. Schultz (1987) suggested that aciniform-shaped glands are the ancestral type for both spider infra-orders, Mygalomorphae and Araneomorphae (fig. 1). Each silk gland is connected to its own spigot that is visible on the external anatomy of a spider. Spigots vary in size, shape, and sculpturing according to the type of silk gland to which they are connected. Based on the broad taxonomic distribution of morphologically distinguishable aciniform, pyriform, and major ampullate-shaped glands or their diagnostic spigots in all examined members of Araneomorphae, their common ancestor is thought to have possessed aciniform, pyriform, and major ampullate glands (Kovoor 1987; Platnick et al. 1991; Griswold et al. 2005). Minor ampullate glands are also widely distributed among araneomorphs and were likely present in the common ancestor of Haplogynae and Entelegynae (fig. 1, Griswold et al. 2005). Tubuliform glands, which are distinguished by their presence in adult females but not adult males, are present in most representatives of Entelegynae that have been examined but are only found in two families of Haplogynae (Kovoor 1987). Tubuliform glands of Entelegynae are thus considered to have an independent derivation from those in Haplogynae (Platnick et al. 1991). Intriguingly, tubuliform glands in entelegynes are virtually indistinguishable from aciniform glands during early development (Richter 1970; Shultz 1987). Furthermore, the number of aciniform glands in *Peucetia* and *Oxyopes* (RTA-clade, Oxyopidae) in adult males is equal to the number of aciniform plus tubuliform glands in adult females (Kovoor and Muñoz-Cuevas 1998), suggesting that tubuliform glands are specialized aciniform glands (Shultz 1987).

Consistent with the glandular affiliation hypothesis, we found a sister relationship between AcSp1 and TuSp1 (fig. 4). The gene duplication event that gave rise to the AcSp1 and TuSp1 paralogs is at least as old as the divergence of orbicularian and RTA-clade spiders (fig. 1), ~240 Ma (Ayoub and Hayashi 2009). This divergence could have happened in the common ancestor of all Entelegynae families (Griswold et al. 2005), or more recently

if Orbiculariae and the RTA-clade are united to the exclusion of other Entelegynae families (Griswold et al. 1999; Coddington 2005). In either case, we predict that Haplogynae should possess spidroins (expressed in aciniform glands) that are orthologous to the Entelegynae AcSp1 plus TuSp1 clade.

Gene duplication events leading to other functionally distinct paralogs are not well supported by our phylogenetic results, but a basal position of the AcSp1 plus TuSp1 clade among araneomorph spidroins (fig. 4) is consistent with aciniform glands representing a gland type that differentiated early in the history of spiders. Additionally, the grouping of araneoid PySp with spidroins from Haplogynae in multiple analyses is consistent with very ancient origins of PySp, perhaps concomitant with the origin of pyriform glands in the common ancestor of Araneomorphae (fig. 5, supplementary fig. S7, Supplementary Material online). More intensive taxonomic sampling of both N- and C-terminal spidroin domains should clarify the history of these ancient duplication events.

Evolution of Repeat Length and Complexity

Spidroin repeat units are extremely variable in sequence among paralogs but can be grouped into three basic categories (referred to here as 1, 2, and 3). (Category 1) The first category includes spidroins with long (e.g., >150 aa) repeat units that have complex amino acid compositions and little internal repetitions. AcSp1, TuSp1, mygalomorph spidroins, and some Haplogynae spidroins (e.g., *P.t.* fibroins 3 and 4) fall into this category (Gatesy et al. 2001; Hayashi et al. 2004; Garb and Hayashi 2005; Garb et al. 2007; Starrett et al. 2012). (Category 2) In contrast, MaSp1 and MaSp2 of Entelegynae have short ensemble repeats that are almost entirely composed of internal repetitions of simple amino acid motifs (e.g., Gatesy et al. 2001; Ayoub, Garb, Tinghitella, et al. 2007). (Category 3) Other spidroins do not fit neatly into either of these two categories but combine elements of both. For instance, the Flag ensemble repeat is almost entirely composed of iterations of GPGXX, but the region that is repeated is very long (e.g., >200 aa) and punctuated by short spacers (e.g., 27 aa) that have a complex amino acid sequence (Hayashi and Lewis 1998). PySp, MiSp and some Haplogynae spidroins (e.g., *P.t.* fibroin1) also fit into category 3 (Gatesy et al. 2001; Blasingame et al. 2009). The broad taxonomic distribution of the first category (long complex repeats) and our rooted spidroin trees suggest that a long repeat unit with complex amino acid composition is the ancestral spidroin condition (figs. 4 and 5, Starrett et al. 2012). Within Orbiculariae, the group of spiders with the most diverse glands and spidroin paralogs, AcSp1 and TuSp1 have retained these features. Retention of these characters may be related to the ecological functions of AcSp1 and TuSp1, which include prey-wrapping and protecting eggs. The functions of spidroins containing simple amino acid motifs (e.g., MaSp1, MaSp2, Flag) include aspects of aerial web building that have demanding tensile requirements, which may have selected for multiple independent shortening and simplification events of spidroin repeats (Garb et al. 2010).

Among spidroins with category 1 (long) repeats, *L. hesperus* and *U. diversus* AcSp1 stand out as exceptionally long (375 and 357 aa, respectively). Araneid AcSp1 repeats are 200–215 aa, TuSp1 repeats are 180–294 aa, and mygalomorph spidroin repeats are 169–181 aa except for fibroins 1 of *Euagrus chisoseus* and *Megahexura fulva*, which are 342 and 365 aa, respectively (Garb et al. 2007; Starrett et al. 2012). Intriguingly, most of the spidroins with repeats longer than 340 aa can be divided into two approximately equal length subrepeats (e.g., supplementary fig. S3, Supplementary Material online), suggesting that ancestral spidroin repeat length is slightly less than 200 aa. Shifts in the periodicity of intragenic concerted evolution events from ~600 to ~1,200 bp could have led to the doubling in size of each of these spidroin repeats. In the case of AcSp1, this shift could have taken place in an orbicularian ancestor or earlier, with a reversal to the smaller repeat unit in araneids. Alternatively, *Latrodectus* and *U. diversus* may have independently evolved longer periodicity while araneids retained the ancestral condition.

Conclusions

Spider aciniform silk is unique in terms of function (prey wrapping), mechanical properties (one of the toughest), and molecular structure. Our complete black widow *AcSp1* gene is the longest coding sequence described for a spider gene and includes 16 iterations of repeating units that are near identical at the amino acid and nucleotide levels. Each repeat encodes a complex amino acid sequence that is conserved across AcSp1 repeats of other species and, at a lower level of similarity, with other spidroin types, such as TuSp1. The homogeneity and complexity of repeats likely contribute to the mechanical properties of aciniform silk and its proper function during prey wrapping. However, stabilizing selection among repeats or purifying selection on amino acid sequence cannot explain the extreme homogenization of *AcSp1* repeats, suggesting that other forms of selection and/or concerted evolution contribute to its molecular structure. AcSp1 possesses many features that are presumed ancestral in spidroins, such as a complex amino acid repeat sequence that lacks extensive subrepeats and a long repeat length. Our phylogenetic results are consistent with the coevolution of spidroin gene duplication events and gland specialization. Specifically, tubuliform glands and aciniform glands are likely derived from an aciniform-like ancestral gland. The sister relationship between TuSp1 and AcSp1 suggests that gene duplication and expression divergence of *TuSp1* and *AcSp1* occurred concomitant with gland differentiation. Finally, our complete *AcSp1* can be used as a template for synthesis of recombinant aciniform silk via transgenic technology. As we increase our understanding of the role of non-coding flanking sequences in the regulation of spidroins, we will be able to capitalize on these regions of genomic sequences for increasing the artificial production of spider silks.

Materials and Methods

Sequencing

We screened a *L. hesperus* genomic library with PCR for clones containing *AcSp1* (see Ayoub, Garb, Tinghitella, et al. 2007 for library construction and screening protocols). Primers used in screening were designed from the repetitive region of a partial-length cDNA (EU025854; primers listed in supplementary table S5, Supplementary Material online). An ~48 kb *AcSp1* containing clone was shotgun sequenced and assembled to 8X coverage by Qiagen (Hilden, Germany), resulting in three contiguous sequences (contigs). One contig contained the 5′ portion of *AcSp1*, a second contig included the 3′ portion, and the third contig consisted of the cloning vector and some insert sequence. All three contigs contained AT microsatellites at one end; the exact number of base pairs in the AT microsatellite regions was difficult to confirm. Comparisons of predicted restriction enzyme recognition sites of the contigs and restriction enzyme digests of the clone indicated a ~500 bp gap within the noncoding region of the insert and a ~500 bp gap within *AcSp1*. We used primer walking to close the gap in the noncoding region but this approach was not possible for the gap within *AcSp1*, due to the extreme similarity among repeats (see Results). Instead, we digested the clone with EcoRV (NEB) and ligated three fragments (2.1, 7, and 7.8 kb) that contained only *AcSp1* coding sequences into pZErO^{TM}-2 plasmids and then transformed TOP10 electrocompetent *Escherichia coli* cells (Invitrogen). We sequenced the ends of each of these subclones and performed a de novo assembly of sequences generated from primer walking, EcoRV subcloning, and shotgun sequencing using PREDPHRAP (Ewing and Green 1998; Ewing et al. 1998). We then manually edited the assembly using CONSED v.19 (Gordon et al. 1998, 2001; Gordon 2004) so that there was agreement between the predicted restriction enzyme digest and the observed digest pattern of the clone.

We amplified *AcSp1* from *L. geometricus* genomic DNA using primers designed from the *L. hesperus AcSp1* genomic sequence (primers in supplementary table S5, Supplementary Material online). We sequenced the following coding regions: 779 bp of N-terminus and adjacent repeat, 850 bp of C-terminus and adjacent repeat, and three fragments of AcSp1 repeats. We assembled the three fragments from the repetitive region (those not adjacent to N- or C-termini) into a single 1,200 bp fragment using Sequencher v.4.9 (GeneCodes) and considered this sequence to represent the consensus *AcSp1* repeat in *L. geometricus*. Double peaks in these chromatograms were scored as polymorphic positions using the IUPAC ambiguity code. Polymorphic positions could represent allelic variation (heterozygosity) or intragenic repeat variation.

We assessed *L. hesperus* allelic variation by directly sequencing four additional *AcSp1* containing genomic clones with C-terminal primers (the genomic library was constructed from multiple individuals). We also amplified C-terminal and adjacent repetitive encoding sequence or adjacent downstream noncoding regions from four individual *L. hesperus* spiders (three collected in Riverside, CA, and one collected in Tucson, AZ).

Homogenization and Conservation of AcSp1

We identified all ORFs greater than 300 bp in the *L. hesperus* genomic clone using ORF Finder (NCBI). We identified *AcSp1* using conceptual translations and BLASTX (Altschul et al. 1990; universal genetic code) comparisons with other spidroins. We considered the first in frame Met to begin AcSp1. Amino acid content and codon usage were determined with CODONW (http://codonw.sourceforge.net/, last accessed November 20, 2012). Hydrophobicity plots were constructed with MacVector 12.5 (MacVector, Inc.) using a window size of eight amino acids. For comparison, hydrophobicity was also plotted for *L. hesperus* MaSp1 (EF595246).

L. hesperus AcSp1 amino acid repeats were identified by eye, separated, and manually aligned. This alignment was used to identify and align nucleotide repeats. We checked every polymorphic position among *AcSp1* repeats in the assembly to ensure that single chromatograms contained all polymorphic positions within a single repeat. This was done to confirm that each of the repeats reported (see Results) existed in the genomic clone.

We searched the NCBI nr protein database using BLASTP with a *L. hesperus* AcSp1 repeat to determine whether the repetitive region of AcSp1 was significantly conserved among species (E value $< 10^{-5}$). We then manually aligned *L. hesperus* AcSp1 repeats to our *L. geometricus* sequence and those identified by BLAST, including sequences from three species in the family Araneidae: *Aran. ventricosus* (ADM35668), *A. trifasciata* (AAR83925), and *A. amoena* (ADM35669); and one in Uloboridae: *U. diversus* (ABD61598) using BLASTP alignments as a preliminary guide.

Pairwise number of synonymous substitutions per synonymous sites (Ks) and nonsynonymous substitutions per nonsynonymous sites (Ka) between repeats were calculated using DNASP v.5 (Librado and Rozas 2009). We compared substitution patterns within the repeats with adjacent terminal-encoding regions by calculating interspecific Ks and Ka between *L. hesperus* and *L. geometricus AcSp1* repeats, N-termini, and C-termini. For comparison with paralogous spidroins, we similarly calculated intragenic and interspecific Ks and Ka values for a partial length cDNA of an *L. hesperus TuSp1* (AY953070), and a complete genomic copy of an *L. hesperus MaSp1* (EF595246). Intragenic comparisons of *MaSp1* focused on the 20 "aggregate repeats" identified in Ayoub, Garb, Tinghitella, et al. (2007).

We also searched for conserved sequences in the flanking regions of *AcSp1* using a variety of methods (see supplementary methods and results, Supplementary Material online). These flanking regions should contain elements involved in gland-specific regulation of AcSp1.

Relationship of AcSp1 to Other Spidroins

We added our *L. hesperus* and *L. geometricus* AcSp1 N- and C-terminal coding sequences to an alignment of 26 spidroin termini generated by Garb et al. (2010). We also added N- and C-terminal coding sequence for "fibroin 1a" from *Deinopis*

spinosa (Deinopidae, see supplementary table S2, Supplementary Material online). N-terminal sequences were not available for many spidroins and thus we added 17 more spidroin sequences to the C-terminal alignment to more comprehensively represent spidroin gene family diversity in our analyses (supplementary table S2, Supplementary Material online). The expanded data sets were aligned with MUSCLE (Edgar 2004) implemented in SEAVIEW v.4.1 (Galtier et al. 1996) and manually edited. Amino acid alignments were used to guide nucleotide alignments in SE-AL v.2.0a11 (http://tree.bio.ed.ac.uk/software/seal/, last accessed November 20, 2012).

We conducted heuristic searches for maximum parsimony (MP) and maximum likelihood (ML) trees based on amino acid and nucleotide alignments in PAUP* v4.0b10 (Swofford 2002) using tree bisection reconnection branch swapping and 1,000 (MP) or 10 (ML) replicates of random stepwise addition of taxa. Support for clades recovered in MP analyses was evaluated with 1,000 bootstrap pseudoreplicates and 10 random addition sequences per pseudoreplicate. Support for clades was further evaluated by calculating decay indices (Bremer 1988; Baker and DeSalle 1997) with the assistance of TREEROT v.3 (Sorenson and Franzosa 2007). Bayesian analyses were carried out with MRBAYES v.3.1.2 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). Optimal models of evolution were determined for nucleotide sequences with JMODELTEST v0.1.1 (Posada 2008) and for protein sequences with PROTTEST v2.4 (Abascal et al. 2005) for N- and C-termini separately. Combined analysis of nucleotides employed a model partitioned by N- and C-termini. Combined analysis of amino acids employed a mixed model, which allowed estimation of the optimal model of protein evolution during the Bayesian analysis. Default priors and Metropolis coupled, Markov-chain, Monte Carlo sampling procedures were executed for two independent runs, sampled every 100th generation, carried out simultaneously. Convergence was assessed every 1,000th generation and the posterior distribution was considered adequately sampled when the standard deviation of split frequencies of these two runs dropped below 0.01 (1–5 million generations depending on data set).

We determined the root of spidroin trees by gene tree-species tree reconciliation, which minimizes the number of inferred gene duplications and losses given a species tree, using NOTUNG v.2.6 (Durand et al. 2006; Vernot et al. 2008) and default cost parameters. Our species tree (fig. 1) is based on a number of previously developed phylogenetic hypotheses for spiders. Family level relationships were based on Coddington (2005) for Araneomorphae and Ayoub, Garb, Hedin, et al. (2007) for Mygalomorphae. Lower level relationships followed those described by Kuntner et al. (2008) for *Nephila*, Scharff and Coddington (1997) for Araneidae, and Elices et al. (2009) for *Argiope*.

Supplementary Material

Supplementary methods and results, figures S1–S7, and tables S1–S5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Patrick Oley and the students in BIOL 221, Winter 2010, at Washington and Lee University for assistance collecting *AcSp1* sequences for *L. geometricus*. They additionally thank Crystal Chaw, James Starrett, and three anonymous reviewers for insightful comments on previous drafts of this manuscript. This work was supported by the National Science Foundation grants (IOS-0951886) to N.A.A. and (IOS-0951061) to C.Y.H.; National Institutes of Health grant (F32 GM78875-1A) to N.A.A.; Army Research Office grant (W911NF-06-1-0455) to C.Y.H.; and Washington and Lee University through Lenfest Summer Fellowships to N.A.A.

References

Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* 21:2104–2105.

Agnarsson I, Kuntner M, Blackledge TA. 2010. Bioprospecting finds the toughest biological material: extraordinary silk from a giant riverine orb spider. *PLoS One* 5:e11234.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol.* 215:403–410.

Askarieh G, Hedhammar M, Nordling K, Saenz A, Casals C, Rising A, Johansson J, Knight SD. 2010. Self-assembly of spider silk proteins is controlled by a pH-sensitive relay. *Nature* 465:236–238.

Ayoub NA, Hayashi CY. 2008. Multiple recombining loci encode MaSp1, the primary constituent of dragline silk, in widow spiders (*Latrodectus*: Theridiidae). *Mol Biol Evol.* 25:277–286.

Ayoub NA, Hayashi CY. 2009. Spiders (Araneae). In: Hedges SB, Kumar S, editors. The timetree of life. New York: Oxford University Press. p. 255–259.

Ayoub NA, Garb JE, Hedin M, Hayashi CY. 2007. Utility of the nuclear protein-coding gene, elongation factor-1 gamma (EF-1γ), for spider systematics, emphasizing family level relationships of tarantulas and their kin (Araneae: Mygalomorphae). *Mol Phylogenet Evol.* 42: 394–409.

Ayoub NA, Garb JE, Tinghitella RM, Collin MA, Hayashi CY. 2007. Blueprint for a high-performance biomaterial: full-length spider dragline silk genes. *PLoS One* 2:e514.

Baker RH, DeSalle R. 1997. Multiple sources of character information and the phylogeny of Hawaiian drosophilids. *Syst Biol.* 46:654–673.

Beckwitt R, Arcidiacono S, Stote R. 1998. Evolution of repetitive proteins: spider silks from *Nephila clavipes* (Tetragnathidae) and *Araneus bicentenarius* (Araneidae). *Insect Biochem Mol Biol.* 28:121–130.

Blackledge TA, Hayashi CY. 2006. Silken toolkits: biomechanics of silk fibers spun by the orb web spider *Argiope argentata* (Fabricius 1775). *J Exp Biol.* 209:2452–2461.

Blasingame E, Tuton-Blasingame T, Larkin L, et al. (12 co-authors). 2009. Pyriform spidroin 1, a novel member of the silk gene family that anchors dragline silk fibers in attachment discs of the black widow spider, *Latrodectus hesperus*. *J Biol Chem.* 284:29097–29108.

Bremer K. 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution* 42:795–803.

Chamary JV, Hurst LD. 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol.* 6:9.

Chinali A, Vater W, Rudakoff B, Sponner A, Unger E, Grosse F, Guehrs KH, Weisshart K. 2010. Containment of extended length polymorphisms in silk proteins. *J Mol Evol.* 70:325–338.

Coddington JA. 2005. Phylogeny and classification of spiders. In: Ubick D, Paquin P, Cushing PE, Roth V, editors. Spiders of North America: an identification manual. American Arachnological Society, p. 18–24, www.americanarachnology.org.

Denny M. 1976. The physical properties of spider's silk and their role in the design of orb-webs. *J Exp Biol.* 65:483–506.

Durand D, Halldórsson BV, Vernot B. 2006. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J Comput Biol.* 13: 320–335.

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

Eisoldt L, Thamm C, Scheibel T. 2012. Review: the role of terminal domains during storage and assembly of spider silk proteins. *Biopolymers* 97:355–361.

Elices M, Plaza GR, Arnedo MA, Pérez-Rigueiro J, Torres FG, Guinea GV. 2009. Mechanical behavior of silk during the evolution of orb-web spinning spiders. *Biomacromolecules* 10:1904–1910.

Ewing B, Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* 8:186–194.

Ewing B, Hillier L, Wendl MC, Green P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* 8:175–185.

Foelix RF. 2010. Biology of spiders. 3rd ed. New York: Oxford University Press.

Galtier N, Gouy M, Gautier C. 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci.* 12:543–548.

Garb JE, Ayoub NA, Hayashi CY. 2010. Untangling spider silk evolution with spidroin terminal domains. *BMC Evol Biol.* 10:243.

Garb JE, DiMauro T, Lewis RV, Hayashi CY. 2007. Expansion and intragenic homogenization of spider silk genes since the triassic: evidence from mygalomorphae (tarantulas and their kin) spidroins. *Mol Biol Evol.* 24:2454–2464.

Garb JE, DiMauro T, Vo V, Hayashi CY. 2006. Silk genes support the single origin of orb webs. *Science* 312:1762.

Garb JE, Hayashi CY. 2005. Modular evolution of egg case silk genes across orb-weaving spider superfamilies. *Proc Natl Acad Sci U S A.* 102:11379–11384.

Gatesy J, Hayashi C, Motriuk D, Woods J, Lewis R. 2001. Extreme diversity, conservation, and convergence of spider silk fibroin sequences. *Science* 291:2603–2605.

Gordon D. 2004. Viewing and editing assembled sequences using Consed. In: Baxevanis D, Davison DB, editors. Current protocols in bioinformatics. New York: John Wiley. p. 11.2.1–11.2.43.

Gordon D, Abajian C, Green P. 1998. Consed: a graphical tool for sequence finishing. *Genome Res.* 8:195–202.

Gordon D, Desmarais C, Green P. 2001. Automated finishing with autofinish. *Genome Res.* 11:614–625.

Gosline JM, Guerette PA, Ortlepp CS, Savage KN. 1999. The mechanical design of spider silks: from fibroin sequence to mechanical function. *J Exp Biol.* 202:3295–3303.

Griswold CE, Coddington JA, Platnick NI, Forster RR. 1999. Towards a phylogeny of entelegyne spiders (Araneae, Araneomorphae, Entelegynae). *J Arachnol.* 27:53–63.

Griswold CE, Ramirez MJ, Coddington JA, Platnick NI. 2005. Atlas of phylogenetic data for entelegyne spiders (Araneae: Araneomorphae: Entelegynae) with comments on their phylogeny. *Proc Calif Acad Sci.* 56(Suppl II):1–324.

Guerette PA, Ginzinger DG, Weber BHF, Gosline JM. 1996. Silk properties determined by gland-specific expression of a spider fibroin gene family. *Science* 272:112–115.

Hagn F, Eisoldt L, Hardy JG, Vendrely C, Coles M, Scheibel T, Kessler H. 2010. A conserved spider silk domain acts as a molecular switch that controls fibre assembly. *Nature* 465:239–242.

Hayashi CY, Blackledge TA, Lewis RV. 2004. Molecular and mechanical characterization of aciniform silk: uniformity of iterated sequence modules in a novel member of the spider silk fibroin gene family. *Mol Biol Evol.* 21:1950–1959.

Hayashi CY, Lewis RV. 1998. Evidence from flagelliform silk cDNA for the structural basis of elasticity and modular nature of spider silks. *J Mol Biol.* 275:773–784.

Hayashi CY, Lewis RV. 2000. Molecular architecture and evolution of a modular spider silk protein gene. *Science* 287:1477–1479.

Hayashi CY, Lewis RV. 2001. Spider flagelliform silk: lessons in protein design, gene structure, and molecular evolution. *Bioessays* 23: 750–756.

Hayashi CY, Shipley NH, Lewis RV. 1999. Hypotheses that correlate the sequence, structure, and mechanical properties of spider silk proteins. *Int J Biol Macromol.* 24:271–275.

Hinman MB, Lewis RV. 1992. Isolation of a clone encoding a second dragline silk fibroin. *Nephila clavipes* dragline silk is a two-protein fiber. *J Biol Chem.* 267:19320–19324.

Holland GP, Creager MS, Jenkins JE, Lewis RV, Yarger JL. 2008. Determining secondary structure in spider dragline silk by carbon-carbon correlation solid-state NMR spectroscopy. *J Am Chem Soc.* 130:9871–9877.

Holland GP, Jenkins JE, Creager MS, Lewis RV, Yarger JL. 2008. Quantifying the fraction of glycine and alanine in β-sheet and helical conformations in spider dragline silk using solid-state NMR. *Chem Commun.* 2008:5568–5570.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754–755.

Ittah S, Cohen S, Garty S, Cohn D, Gat U. 2006. An essential role for the C-terminal domain of a dragline spider silk protein in directing fiber formation. *Biomacromolecules* 7:1790–1795.

Jenkins JE, Creager MS, Butler EB, Lewis RV, Yarger JL, Holland GP. 2010. Solid-state NMR evidence for elastin-like β-turn structure in spider dragline silk. *Chem Commun (Camb).* 46:6714–6716.

Jenkins JE, Creager MS, Lewis RV, Holland GP, Yarger JL. 2010. Quantitative correlation between the protein primary sequences and secondary structures in spider dragline silks. *Biomacromolecules* 11:192–200.

Katz L, Burge CB. 2003. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Res.* 13: 2042–2051.

Kovoor JJ. 1987. Comparative structure and histochemistry of silk-producing organs in arachnids. In: Nentwig W, editor. Ecophysiology of spiders. Berlin (Germany) and New York: Springer-Verlag. p. 159–186.

Kovoor JJ, Muñoz-Cueva A. 1998. Structure and function of the silk-gland system in Oxyopidae (Araneae). In: Selden P, editor. Proceedings of the 17th European Colloquium of Arachnology. Edinburgh (UK); British Arachnological Society, Burnham Beeches (UK). p. 133–141.

Kuntner M, Coddington JA, Hormiga G. 2008. Phylogeny of extant nephilid orb-weaving spiders (Araneae, Nephilidae): testing morphological and ethological homologies. *Cladistics* 24:147–217.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25:1451–1452.

Meyer IM, Miklós I. 2005. Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs. *Nucleic Acids Res.* 33:6338–6348.

Motriuk-Smith D, Smith A, Hayashi CY, Lewis RV. 2005. Analysis of the conserved N-terminal domains in major ampullate spider silk proteins. *Biomacromolecules* 6:3152–3159.

Peng L, Sakharkar KR, Sakharkar MK. 2009. Genome architecture—number, size, and length distributions of exons and introns in six crown eukaryotic genomes. *Int J Integr Biol.* 5:87–91.

Platnick NI, Coddington JA, Forster RR, Griswold CE. 1991. Spinneret morphology and the phylogeny of haplogyne spiders (Araneae, Araneomorphae). *Am Museum Novitates* 3016:1–73.

Posada D. 2008. jModelTest: phylogenetic model averaging. *Mol Biol Evol.* 25:1253–1256.

Richter CJ. 1970. Morphology and function of the spinning apparatus of the wolf spider *Pardosa amentata* (Cl.) (Araneae, Lycosidae). *Zeitschrift fur Morphologie der Tiere* 68:37–68.

Rising A, Widhe M, Johansson J, Hedhammar M. 2011. Spider silk proteins: recent advances in recombinant production, structure-function relationships, and biomedical applications. *Cell Mol Life Sci.* 68:169–184.

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

Scharff N, Coddington JA. 1997. A phylogenetic analysis of the orb-weaving spider family Araneidae (Arachnida, Araneae). *Zool J Linn Soc.* 120:355–434.

Shultz JW. 1987. The origin of the spinning apparatus in spiders. *Biol Rev.* 62:89–113.

Sorenson MD, Franzosa EA, editors. 2007. TreeRot, version 3. Boston: Boston University.

Sponner A. 2007. Spider silk as a resource for future biotechnologies. *Entomol Res.* 37:238–250.

Sponner A, Schlott B, Vollrath F, Unger E, Grosse F, Weisshart K. 2005. Characterization of the protein components of *Nephila clavipes* dragline silk. *Biochemistry (NY)* 44:4727–4736.

Starrett J, Garb JE, Kuelbs A, Azubuike UO, Hayashi CY. 2012. Early events in the evolution of spider silk genes. *PLoS One* 7: e38084.

Sved J, Bird A. 1990. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc Natl Acad Sci U S A.* 87:4692–4696.

Swofford DL. 2002. PAUP*. Phylogenetic analysis using parsimony. Sunderland (MA): Sinauer Associates.

Tran JC, Zamdborg L, Ahlf DR, et al. (18 co-authors). 2011. Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* 480:254–258.

Vasanthavada K, Hu X, Falick AM, La Mattina C, Moore AMF, Jones PR, Yee R, Reza R, Tuton T, Vierra C. 2007. Aciniform spidroin, a constituent of egg case sacs and wrapping silk fibers from the black widow spider *Latrodectus hesperus*. *J Biol Chem.* 282:35088–35097.

Vernot B, Stolzer M, Goldman A, Durand D. 2008. Reconciliation with non-binary species trees. *J Comput Biol.* 15:981–1006.

Xu M, Lewis RV. 1990. Structure of a protein superfiber: spider dragline silk. *Proc Natl Acad Sci U S A.* 87:7120–7124.

Zhang J. 2003. Evolution by gene duplication: an update. *Trends Ecol Evol.* 18:292–298.

Zhao AC, Zhao TF, Nakagaki K, et al. (11 co-authors). 2006. Novel molecular and mechanical properties of egg case silk from wasp spider, *Argiope bruennichi*. *Biochemistry (NY)* 45: 3348–3356.