Published in final edited form as:

Insect Mol Biol. 2008 September ; 17(5): 465–474. doi:10.1111/j.1365-2583.2008.00828.x.

Identification and Characterization of Multiple *Spidroin 1* Genes Encoding Major Ampullate Silk Proteins in *Nephila clavipes*

William A. Gaines IV* and William R. Marcotte Jr.*,[†]

*Department of Genetics and Biochemistry, Clemson University, Clemson, SC 29634

[†]Department of Biological Sciences, Clemson University, Clemson, SC 29634

Abstract

Spider dragline silk is primarily composed of proteins called major ampullate spidroins (MaSp) that consist of a large repeat array flanked by non-repetitive N- and C-terminal domains. Until recently, there has been little evidence for more than one gene encoding each of the two major spidroin silk proteins, MaSp1 and MaSp2. Here, we report the deduced N-terminal domain sequences for two distinct *MaSp1* genes from *Nephila clavipes* (*MaSp1A* and *MaSp1B*) and for *MaSp2*. All three *MaSp* genes are co-expressed in the major ampullate gland. A search of the GenBank database also revealed two distinct *MaSp1* C-terminal domain sequences. Sequencing confirmed that both *MaSp1* genes are present in all seven *Nephila clavipes* spiders examined. The presence of nucleotide polymorphisms in these genes confirmed that *MaSp1A* and *MaSp1B* are distinct genetic loci and not merely alleles of the same gene. We have experimentally determined the transcription start sites for all three *MaSp* genes and established preliminary pairing between the two *MaSp1* N- and C-terminal domains. Phylogenetic analysis of these new sequences and other published *MaSp* N- and C-terminal domain sequences illustrated that duplications of *MaSp* genes may be widespread among spider species.

Keywords

major ampullate spidroin; spider silk; *Nephila clavipes*; N-terminal domain; gene duplication

Introduction

Silks are highly-insoluble proteinaceous fibers produced by numerous species of insects and spiders. Orb-weaving spiders (*Araneidae*) like *Nephila clavipes* can produce up to seven different types of silk that originate from different abdominal glands (Lewis, 2006). These fibers vary widely in mechanical properties and have evolved to accommodate specific needs of the spider. The frame and radii of the web are made from dragline silk that exhibits appreciable elasticity and strength resulting in an incredibly tough fiber (Gosline *et al.*, 1999). Flagelliform silk is the most elastic of the silks and is used for the capture spiral of the web (Hayashi & Lewis 1998). Aciniform silk is used for wrapping prey and while intermediate with respect to strength when compared to the weaker flagelliform and stronger dragline silks, it is superior to both in toughness (Blackledge & Hayashi, 2006).

Among the different types of silks, dragline silk is the most well-studied. Dragline silk is used not only to construct the outer frame and radii of the orb-shaped web but also as a hanging lifeline that allows the spider to evade and/or escape from predators. The core

constituents of dragline silk are two fibrous proteins produced in the major ampullate gland that are called major ampullate spidroins 1 and 2 (MaSp1 and MaSp2; Xu & Lewis, 1990; Hinman & Lewis, 1992; Sponner *et al.*, 2005a). These proteins are stored at high concentration in the lumen of the gland as a spinning dope (Hijirida *et al.*, 1996). When there is demand for dragline fiber, this spinning dope is drawn down a long, tapered duct in which there are documented changes in both pH and the ionic environment (Knight & Vollrath, 2001). As the proteins traverse the length of the duct and emerge from the spinneret, these soluble spidroin proteins condense and assemble/transition into a highly organized, insoluble fiber (Knight *et al.*, 2000).

MaSp1 and MaSp2 are large proteins of about 250 to 350 kDa that share a general domain architecture (Sponner *et al.*, 2005a; Ayoub *et al.*, 2007). Both proteins contain a large, central, repetitive domain that consists of approximately 100 tandem copies of a 30 to 40 amino acid repeat sequence. The consensus repeat sequences for both MaSp1 and MaSp2 are glycine-rich and end in poly-alanine motifs (usually four to seven residues long). For MaSp1, the consensus repeat includes (GGX)_n motifs (where X = A, L, Q, or Y) and very low proline content. In contrast, the MaSp2 consensus repeat has significant proline content and characteristic motifs such as GPG and QQ (Gatesy *et al.*, 2001). The repetitive domains of different spidroins display a relatively high level of amino acid sequence variation that has been implicated in providing the elasticity and toughness that is characteristic of the different fibers (Hayashi & Lewis, 1998; Hayashi *et al.*, 1999; Rising *et al.*, 2005).

The repeat arrays of both MaSp1 and MaSp2 (and other spidroins) are flanked by non-repetitive N-terminal and C-terminal domains of approximately 150 and 100 amino acids, respectively. The deduced C-terminal domain sequences of MaSp proteins have been known for almost two decades (Xu & Lewis, 1990; Hinman & Lewis, 1992). Functional studies using partial dragline silk spidroins have implicated the C-terminal domain in the organized transition from a soluble spidroin solution to an insoluble fiber during spinning (Huemmerich *et al.*, 2004; Sponner *et al.*, 2005b; Ittah *et al.*, 2006; Stark *et al.*, 2007). There is also a conserved cysteine residue in the C-terminal domain of most MaSp proteins that has been suggested to be important in assembly of the final fiber as a result of intermolecular cross-linking through disulfide bond formation (Sponner *et al.*, 2005a).

In contrast, N-terminal domain *MaSp* sequences were not known until more recently due, at least in part, to difficulties in cloning and/or sequencing the large and highly-repetitive spidroin genes and cDNAs (Motriuk-Smith *et al.*, 2005). Analysis of these sequences demonstrates that they are the most highly conserved domains of spidroins. The only biochemical characterization of this domain is a study on recombinant protein where circular dichroism (CD) spectroscopy confirmed the largely α -helical composition and size exclusion chromatography indicated homodimerization (Hedhammar *et al.*, 2008). While a number of these reported sequences have been deduced from genomic DNA, it has allowed evaluation of the N-terminal domains for specific motifs that would support their authenticity. One of these motifs, a signal peptide that would be consistent with secretion of spidroins into the duct, now appears to be a clearly recognized component of essentially all known N-terminal domains (Rising *et al.*, 2006).

Interestingly, the N-terminal and C-terminal domains found on mature spidroins are not only conserved between MaSp1 and MaSp2, but also among many silk types and spider species (Beckwitt & Arcidiacono, 1994; Hayashi *et al.*, 2004; Rising *et al.*, 2006). This suggests that despite their relatively small size, they play an important role for the function of the silk. Individual N- and C-terminal domains themselves bear little resemblance to each other except that they are both rich in serine (~13% for N-terminal and 23% for C-terminal) and both are predicted to exist as largely amphipathic α -helical secondary structures (Sponner *et*

al., 2005b; Rising *et al.*, 2006). In both cases, this could be consistent with terminal globular domains resulting from the interaction of the amphipathic helices.

To further our knowledge of N-terminal sequences of the major ampullate spidroins, we have isolated and characterized both genomic and cDNA sequences from *N. clavipes*. We report sequences that are consistent with those reported for other species, including the presence of a signal peptide for secretion. Further, we present data that indicates not only the presence of multiple *MaSp1* genes in the *N. clavipes* genome but also that both of these genes are expressed in individual major ampullate glands.

Results

Using a combination of degenerate primer PCR, inverse PCR and asymmetric PCR, we have determined genomic sequences present at the 5' ends of several genes from *Nephila clavipes* that encode MaSp proteins. The areas sequenced included approximately the first four copies of the spidroin repeat array and similarity of the deduced amino acid sequences to previously reported spidroin proteins has allowed us to define them as spidroin 1 (MaSp1) or spidroin 2 (MaSp2). As can be seen in Fig. 1C (underlined), the first two sequences are particularly rich in (GGX)_n motifs that are characteristic of MaSp1 but less common, and not tandemly repeated, in MaSp2 (Gatesy *et al.*, 2001). The third sequence contains GPG and QQ motifs and a predominance of proline all of which are characteristic of MaSp2 but not MaSp1. Interestingly, while the general repeat structure of the first two sequences resemble MaSp1, these sequences cannot be directly aligned suggesting that they may represent allelic variants or separate genes. All three sequences contain poly-alanine motifs in the repeat structure, a characteristic of both MaSp1 and MaSp2. Therefore, these three sequences will be referred to as *Nephila clavipes* MaSp1 variant A (*MaSp1A*), *MaSp1B*, and *MaSp2*.

To determine if all three sequences are expressed in major ampullate glands and to establish the 5' ends of potential transcripts, we conducted 5' RLM-RACE on total RNA isolated from a single major ampullate gland (see Experimental Procedures for details; Maruyama & Sugano, 1994). Complimentary DNA was reverse transcribed using a universal primer designed to anneal to all three *MaSp* sequences upstream of the repeat arrays. Subsequent PCR was conducted with a primer that binds the RLM-RACE ribo-oligonucleotide adaptor and a nested *MaSp* universal primer. A single band at approximately 300 bp was produced from the experimental RNA sample that was treated with tobacco acid phosphatase (+TAP) but not from the -TAP control RNA sample (Fig. 2) indicating that this product(s) was derived from a capped RNA. RLM-RACE products were isolated from the gel and cloned. Sequence analysis confirmed that all three sequences were present in the single RLM-RACE band seen in Fig. 2 indicating that all were expressed in the single major ampullate gland from which the RNA template was isolated. In addition, this analysis defined the 5' ends of the transcripts (asterisks in Fig. 1A).

We also determined approximately 350 bp of 5' flanking sequence for two of the three gene sequences (*MaSp1A* and *MaSp2*). Evaluation of these regions with promoter prediction software identified a putative promoter in the 5' regions of both *MaSp1A* and *MaSp2* (Reese, 2001). This promoter contains a canonical TATA box 32 nt and 31 nt upstream from a predicted transcription start site in *MaSp1A* and *MaSp2*, respectively (Fig. 1A). The predicted start sites correspond exactly with those determined experimentally. The 5' untranslated regions (UTRs) are short (27-28 nt) and the first ATG in the sequence has a local nucleotide context congruent with the *Drosophila* translation initiation consensus (i.e. Kozak site; Miyasaka, 2002). This predicted translation start site is consistent with the

predicted translations of *MaSp* genes from other species (Motriuk-Smith *et al.*, 2005; Rising *et al.*, 2006).

Comparison of the non-repetitive N-terminal domains encoded by these three sequences showed that *MaSp1A* and *MaSp1B* share 91% amino acid identity with each other while *MaSp2* has ~72% identity with either *MaSp1A* or *MaSp1B* (Fig. 1B). The differences among the three N-terminal domains are largely conservative amino acid substitutions. Consistent with secretion of these proteins into the lumen of the major ampullate duct, analysis of the three *MaSp* protein sequences using SignalP predicts that the N-terminal 23-24 amino acids encode a signal peptide (Bendtsen *et al.*, 2004). Additionally, secondary structure prediction suggests that the N-terminal regions between the signal peptide and repeat arrays are predominantly alpha-helix with ~60% of the amino acids participating in five helical structures (Fig. 1B). The identification of the signal peptide and the prediction of five helices are consistent with characterization of previously identified *MaSp* genes in other species (Motriuk-Smith *et al.*, 2005; Rising *et al.*, 2006).

Having established that all three sequences are expressed in an individual spider, it was of interest to try to clarify if *MaSp1A* and *MaSp1B* sequences represent either allelic or splice variants of a single gene or are indeed distinct genes. To do this, we isolated *MaSp1A* and *MaSp1B* genomic N-terminal sequences from seven adult *N. clavipes* females. For each individual, genomic DNA was amplified using variant-specific primers, cloned into a plasmid vector and independent plasmids were sequenced. The results are summarized in Table 1 and show that all seven individuals have copies of both *MaSp1A* and *MaSp1B* sequences. However, while all sequences for *MaSp1B* were identical, the *MaSp1A* sequences displayed a single nucleotide polymorphism (SNP) at nucleotide position 458 (relative to the transcription start site). The presence of both *MaSp1A* SNPs in five out of the seven individuals tested suggests they are heterozygous for the SNP alleles of *MaSp1A* but also have a *MaSp1B* gene. Close examination of *N. clavipes* *MaSp1* C-terminal sequences available in public datasets also identified two variants (Sponner *et al.*, 2004; previously called Variants 4 and 5, GenBank accession nos. AY654291 and AY654292, respectively). Interestingly, in a manner similar to the N-terminal sequences characterized above, the two *MaSp1* C-terminal domains share ~90% identity with each other and ~72% identity to the single *MaSp2* C-terminal domain found in the public datasets.

This observation suggested to us the presence of two distinct *MaSp1* genes in which the two N-terminal domain variants of *MaSp1* (A and B) are paired individually with the two C-terminal variants (4 and 5). To explore this possibility, we first needed to determine if both C-terminal variants were present within our sample of seven individuals. Genomic DNA was subjected to PCR using variant-specific *MaSp1* C-terminal primers and the products were distinguished by digestion at a polymorphic *XhoI* site. As seen in Table 1, the results are similar to our findings for the *MaSp1A* and *MaSp1B* variants. Specifically, each individual possesses both *MaSp1* variant 4 and *MaSp1* variant 5 C-terminal sequences. We also cloned the PCR products from individuals 1 and 3 for both variants 4 and 5 and sequenced 3 independent isolates of each. Individual 3 had a single sequence representing C-terminal variant 4 and a single sequence representing variant 5. Individual 1 also had a single sequence representing C-terminal variant 4 but this sequence had a SNP when compared to individual 3. On the other hand, individual 1 possessed two sequences representing C-terminal variant 5 that differed by a single polymorphic nucleotide. One of these was identical to that found in individual 3. Therefore, all individuals in this small sample contain *MaSp1* C-terminal variants 4 and 5 and there exist at least two SNP alleles of each. Taken together, the data in Table 1 strongly suggests that *MaSp1A* and *MaSp1B* are not simply allelic variants but most likely represent two distinct genes.

To further explore this possibility, we attempted to establish direct linkage between the N- and C-terminal variants of *MaSp1*. Due to the long and repetitive nature of the *MaSp1* genes, however, direct PCR of entire *MaSp1* genes was not a viable option. Instead, we used variant-specific *MaSp1* C-terminal reverse primers in separate reactions to prime a long primer extension reaction on genomic DNA template from a single individual. The extension reactions were repeated for 15 rounds and the conditions of the reaction were such that the products should have extended beyond the N-terminal start codon. By priming at each C-terminus independently, the relative abundance of the associated N-terminus should increase within each genomic DNA sample. Subsequently, both *MaSp1A* and *MaSp1B* were simultaneously amplified for each reaction using a universal primer set and the ratio of *MaSp1A* to *MaSp1B* was determined by direct sequencing of the PCR products and quantitation of peak areas at polymorphic nucleotides. It was found that the abundance of *MaSp1A* increased significantly in the sample primed with the *MaSp1* C-terminal variant 5-specific primer and that the abundance of *MaSp1B* increased significantly in the sample primed with the *MaSp1* C-terminal variant 4-specific primer (Fig. 3). A Student's t-test revealed that each of the three priming groups (C-terminal variant 4, variant 5, and the no priming control) are significantly different from each other (p-value < 0.01). We propose that the C-terminal domains previously referred to as *MaSp1* C-terminal variant 4 and 5 be renamed as *MaSp1B* and *MaSp1A* C-terminal domains, respectively, to indicate the pairing of N- and C-terminal domains.

Nucleotide based distance trees (Neighbor Joining, complete deletion) were estimated for our *N. clavipes* sequences along with the published *MaSp* sequences of other species (Fig. 4). The overall tree topology changed little when synonymous sites or amino acid sequences were used. Additionally, estimated maximum parsimony and UPGMA tree topologies were very similar (results not shown). As observed previously, clades generally group first by spider genus then by spidroin type (*MaSp1* or *MaSp2*). In both trees, *MaSp* sequences from *N. madagascariensis* are most similar to *N. clavipes* and those from *Euprosthonops australis* are the most distantly related. It is interesting to note that *Nc.MaSp1* C-terminal variant 4 is more closely related to the *Nm.MaSp1* C- and *Ns.MaSp1* C-terminal domains than it is to *Nc.MaSp1* C-terminal variant 5 (Fig. 4B). It is entirely possible that these other *Nephila* species share the same *MaSp1* gene duplication and only one of their *MaSp1* C-terminal sequences are represented in the database. The *Nc.MaSp1A* and *Nc.MaSp1B* N-terminal sequences form a clade in the N-terminal tree (Fig. 4A). However, this is likely due to the low representation of *Nephila* species in this data set. It's also noteworthy that several species appear to have multiple *MaSp1* or *MaSp2* gene sequences in the database. These potential *MaSp1* and *MaSp2* paralogs are marked by an asterisk in Fig. 4. The groups of sequences marked by double asterisk are *MaSp1* genes which have been experimentally confirmed as paralogs in *N. clavipes* by the present paper, in *Euprosthonops australis* (Rising *et al.*, 2007), and in *Latrodectus hesperus* and *L. geometricus* (Ayoub & Hayashi, 2008).

Discussion

Historically, the dragline silk of *N. clavipes* is among the best-characterized spider silk fibers with respect to its mechanical properties. While full sequences for the *Nc.MaSp1* and *Nc.MaSp2* genes are not currently available, partial cDNAs for these two genes were the first dragline silk sequences isolated (Xu & Lewis, 1990; Hinman & Lewis, 1992). Since then, the isolation of additional C-terminal sequences from other spider species has revealed considerable similarity across species and silk types (Beckwitt & Arcidiacono, 1994; Hayashi *et al.*, 2004). Other studies have characterized the 5' gene sequences for dragline, flagelliform, and cylindriform silk genes and utilized promoter prediction software to look for potential regulatory elements (Motriuk-Smith *et al.*, 2005; Rising *et al.*, 2006; Ayoub *et*

et al., 2007). Such elements have been noted and the predicted transcription start sites have allowed a comparison of the deduced primary and secondary structures of silk protein N-terminal domains.

MaSp protein primary structure has been characterized in some detail as have some elements of secondary structure and the information reported here contributes directly to the breadth of our knowledge of these sequences. Our 5' RLM-RACE results are the first direct empirical determination of the transcription start site for any major ampullate silk gene. Interestingly, not only are the predicted sequences of N-terminal coding regions conserved among different genera of spiders and silk types, but the presence of a signal peptide and five predicted helical structures seem to be conserved among different silk types as well (Motriuk-Smith *et al.*, 2005; Rising *et al.*, 2006). From this we can infer that the silk N-terminal domain plays an important role in silk function since the different elements of this domain have been maintained from a time that predates the divergence of the spider species presented here (Selden & Gall, 1992; Rising *et al.*, 2006).

The transcription start sites identified are consistent with promoter prediction results for these same regions and are also consistent with the predicted start sites identified in *MaSp* 5' gene sequences from other species (Motriuk-Smith *et al.*, 2005; Rising *et al.*, 2006; Ayoub *et al.*, 2007). The first AUG codon downstream of the transcription start site is, in all cases, immediately followed by a 23-24 amino acid sequence that returns a maximum signal peptide probability score (1.000) by SignalP analysis. This analysis also shows that there are not secondary transcription start sites. Therefore, in the absence of an internal ribosome entry site at downstream AUG codons, it would not be possible to generate short isoforms of MaSp proteins. The absence of sequences associated with downstream AUG codons that elicit favorable signal peptide probability scores further argues against the production of shorter isoforms. Others have noted a distinct codon bias in repetitive regions adjacent to the C-terminal domains of *N. clavipes* (Xu & Lewis, 1990) and in the full length repetitive regions of *L. hesperus* (Ayoub *et al.*, 2007). As is frequently seen in highly expressed proteins, the most common amino acids in spidroins (in this case G, A, Q, and P) have the most skewed codon usages and the sequences reported here display an essentially identical codon bias.

While the sequences reported here are consistent with those described previously, we demonstrate that the *N. clavipes* genome contains at least three major ampullate silk protein genes, *MaSp1A*, *MaSp1B* and *MaSp2*. Based on the N-terminal sequence data, the *MaSp1A* gene is present as at least two SNP alleles. It is highly unlikely that this polymorphism is the result of errors in the PCR reaction for several reasons. First, the polymerase used (Phusion™) is a very high fidelity polymerase, especially for the short PCR reactions used to isolate 5' gene sequences (from genomic DNA or cDNA). Second, the same polymorphism was identified in multiple adult female individuals. Third, this observation is consistent with the recent report of multiple *MaSp1* loci in *Latrodectus* (Ayoub & Hayashi, 2008). SNP alleles may also exist for the *MaSp1B* gene. More importantly, we have shown that all three gene transcripts are present in single major ampullate glands of individual adult female spiders. This demonstrates that both *MaSp1* genes are functional and are expressed simultaneously in time and space. The presence of two *MaSp1* genes in *N. clavipes* is not limited to spiders harvested in coastal South Carolina as the same result was found in spiders harvested in southern Georgia/northern Florida (data not shown).

Our characterization of C-terminal sequences along with those available in public datasets also argues for at least two MaSp1 silk protein genes in *N. clavipes*. We have identified two SNP alleles for the C-terminal domains of both *MaSp1* genes (previously called variants 4 and 5; Sponner *et al.*, 2004). In fact, all seven individuals that were sequenced had both

MaSp1 N-terminal variants (A and B) and both C-terminal variants (4 and 5). If an allelic relationship for these variants was assumed, then the 100% heterozygosity of this sample population would be very unlikely. More significantly, five of the seven sequenced individuals had both SNP alleles of *MaSp1A* and a copy of *MaSp1B*. Since *N. clavipes* is diploid (Araujo *et al.*, 2005), these results argue very strongly against an allelic relationship between *MaSp1A* and *MaSp1B*.

Using a variant-specific priming approach, we have attempted to establish linkage between the two different C-terminal sequences and *MaSp1A* /*MaSp1B* N-terminal sequences. Our results show that variant 5-specific priming in the C-terminus increases the frequency with which *MaSp1A* N-terminal sequences are isolated by subsequent PCR. Conversely, variant 4-specific priming increases the frequency with which *MaSp1B* N-terminal sequences can be isolated. This observation strongly suggests that the variant 5 and variant 4 C-terminal domains are part of the *MaSp1A* and *MaSp1B* genes, respectively, and argues against a mechanism that involves alternative splicing of a single gene. Consistent with this conclusion, the complete gene sequence for *L. hesperus MaSp1* is known to encode a single exon (Ayoub *et al.*, 2007). While the relationship between the *MaSp1* N- and C-termini will only be unambiguously established through isolation of complete gene sequences for *MaSp1A* and *MaSp1B*, all the results presented support the existence of two *MaSp1* loci in *N. clavipes*.

Having found that there are two *MaSp1* genes expressed in the major ampullate gland, it is important to ask what, if any, benefit this gene duplication event may provide to the spider. Because *MaSp1* represents the most abundant protein in dragline silk, one possibility is that two *MaSp1* genes may simply be more efficient in producing the large amounts of *MaSp1* protein required. If that is the case, a single *MaSp2* gene may suffice for the less abundant dragline silk protein. It has been observed that endopolyploidy may occur in silkworm (*B. mori*) silk glands (Perdrix-Gillot, 1979) and possibly also spider silk and venom glands (Rasch & Connelly, 2005). Multiple gene copies, in addition to endopolyploidy, could increase the number of *MaSp1* and *MaSp2* templates available for transcription to meet the silk demand.

Alternately, the two *MaSp1* proteins may differ compositionally and functionally in their repetitive domains. By altering the ratio of *MaSp1A* and *MaSp1B* expression, a spider may be able to tailor physical properties of the silk fiber to adjust to changing needs. This would be consistent with observations that individual spiders of *Nephila edulis* can produce major ampullate silks with differing physical qualities; however, it is more likely that these fiber differences are caused by alterations in spinning conditions (Madsen *et al.*, 1999).

Despite the fact that *N. clavipes* has been historically well characterized with respect to spider silk sequencing, this second *MaSp1* gene has been overlooked until now. Our work in *Nephila* spiders and the recent identification of multiple *MaSp1* loci in *Latrodectus* (Ayoub & Hayashi, 2008) and in *Euprosthenops* (Rising *et al.*, 2007) raise the possibility that other spiders also have novel *MaSp* genes yet to be discovered. A search of the GenBank database showed that *Nephila pilipes* (a close relative to *N. clavipes*) also has two *MaSp1* C-terminal variants that are 96% identical to each other (GenBank accession Nos. AY666051, AY666053; see Fig. 4B). The report characterizing these sequences states that three of three *N. pilipes* individuals for which genomic DNA was sequenced had both *MaSp1* C-terminal variants present (Tai *et al.*, 2004). Additionally, two *MaSp2* N-terminal variants are present in *L. hesperus* and three *MaSp2* C-terminal variants are present in *Argiope amoena* (see Fig. 4; Ayoub *et al.*, 2007; Pan *et al.*, 2007). However, since these sequences are not well characterized in the literature, it is difficult to say whether the *MaSp2* variants are alleles or

gene paralogs. Overall, it appears that *MaSp* gene duplications are widespread among spider species.

Experimental Procedures

Spider Collection and Nucleic Acid Isolation

Seven live, adult female *Nephila clavipes* spiders were collected at Georgetown, South Carolina and gland material was harvested within 1 week of collection. To encourage elevated spidroin gene expression, spiders were silked from the major ampullate gland for 30 minutes followed by a ten minute recovery period. Within five minutes of sacrificing a spider, major ampullate glands were dissected from the abdomen and placed in RNAlater solution (Ambion, Austin, TX). RNA was isolated from major ampullate glands of individuals 1 and 3 using an illustra RNeasy mini kit (GE Healthcare, Piscataway, NJ). DNA was isolated for all 7 individuals from non-major ampullate gland material by grinding in a tissue lysis buffer (150mM Tris pH 8.0, 300mM NaCl, 10mM EDTA, 0.5% SDS, 0.5 mg/ml proteinase K), incubation at 45°C for 4 hrs, phenol:chloroform:isoamyl alcohol (25:24:1) extraction, and ethanol precipitation.

Genomic DNA PCR

5' *MaSp* Gene Characterization—Five nucleotide sequences for the N-terminal regions of major ampullate spidroin genes from public datasets (GenBank accession nos. DQ059135, DQ059136, DQ059133, DQ379381, DQ379382) were aligned. These sequences, from *MaSp1* and *MaSp2*, show a high level of sequence similarity and allowed the development of nested sets of degenerate primers in regions of high conservation (Primers Sp1 degenerate F1/R1 and F2/R2, Supplementary Table 1) for PCR using *N. clavipes* genomic DNA as template. The sequence information obtained from degenerate primer PCR allowed the development of primers for inverse PCR (Primers Sp1 and Sp2 inverse F1/R1, Supplementary Table 1). Genomic DNA was digested with *Hind* III and subjected to dilute ligation followed by ethanol precipitation. This DNA served as the template for inverse PCR using the outward facing primer set. Asymmetric PCR was performed using a non-degenerate forward primer (Primers Sp1 and Sp2 inverse F1, Supplementary Table 1) and degenerate reverse primers designed to bind either the *MaSp1* or *MaSp2* repeat consensus (Primers Sp1 and Sp2 asymmetric R1, Supplementary Table 1). All PCR utilized Phusion™ polymerase (New England Biolabs, Ipswich, MA) and all PCR products were cloned into pGEM®-T Easy vector (Promega Corp., Madison, WI) prior to sequencing using ABI BigDye® terminator sequencing with a T7 promoter primer.

MaSp1 Variant-specific N- and C-Terminal Coding Region Characterization

Variant-specific primers were used to amplify the *MaSp* N- (variants A and B) and C-terminal (variants 4 and 5) coding regions from genomic DNA template of seven individual spiders (Primers Sp1- C Var4 and Var5 F1/R1, Supplementary Table 1) and the products were cloned separately. For all seven individuals, five independent clones were sequenced for *MaSp1* N-terminal variant A and three independent clones were sequenced for *MaSp1* N-terminal variant B. PCR products for *MaSp1* C-terminal variants 4 and 5 were cloned for only individuals 1 and 3 and three independent clones were sequenced for each variant. The remaining *MaSp1* C-terminal PCR products were identified as variant 4 or 5 by the presence/absence of a polymorphic *Xho* I restriction site.

***MaSp1* N-terminal to C-terminal Linkage Determination**—N to C linkage was determined by long single primer extensions from the C- to N-terminus followed by PCR of both *MaSp1* N-termini to determine relative abundance. Genomic DNA (400 ng) from individual 5 was treated with ligase to repair nicks. Triplicate extension reactions containing

40 ng genomic DNA each were performed using Phusion™ polymerase and a *MaSp1* variant 4 primer, a variant 5 primer and no primer for a total of nine reactions (Primers *MaSp1*-C Var4/Var5 R2, Supplementary Table 1). The reaction conditions included the supplied GC buffer and 6% DMSO to relax secondary structures. The ten minute extension cycles were repeated 15 times. Aliquots of the primer extension products were then added directly to PCR reactions using a *MaSp1* N-terminal universal primer pair to amplify both variant A and variant B sequences (Primers Sp1-N universal F1/R1). The PCR products were gel purified then sequenced directly using the forward universal *MaSp1* N-terminal primer. Sequence chromatograms were compared at each polymorphic nucleotide site between *MaSp1A* and *MaSp1B*. For each polymorphic double peak, the area under each peak was determined by enlarging the chromatogram (~400%) and determining the area of a triangle created under the peak. The average difference in peak area was taken as the difference in relative abundance of *MaSp1A* to *MaSp1B* for that reaction.

RNA Ligase Mediated-RACE (RLM-RACE)

RNA isolated from individual 1 was subjected to RLM-RACE using a protocol similar to previously described methods (Maruyama & Sugano, 1994). Briefly, 2.5 µg of total RNA was treated with 5 units of shrimp alkaline phosphatase (SAP, Promega, Madison, WI) then split into two 1.25 µg fractions. One of these fractions was treated with 5 units of tobacco acid pyrophosphatase (TAP, Epicentre, Madison, WI) and the other served as a negative control. After both the SAP and TAP treatments, DNA was purified with a 5:1 acid phenol/chloroform mix and precipitated with isopropanol. Subsequently, 400 pmol of a ribo-oligonucleotide adaptor (Primer RLM-RACE ribo, Supplementary Table 1) was ligated to the 5' end of the each RNA fraction using 2.5 units of T4 RNA ligase (Epicentre, Madison, WI). cDNA synthesis was then performed using M-MuLV reverse transcriptase (New England Biolabs, Ipswich, MA) and a low degeneracy universal primer (Primer Sp-N 5' RACE R1, Supplementary Table 1) within the N-terminal domain that will anneal to *MaSp1A*, *MaSp1B*, and *MaSp2*. The SAP, TAP, ligase, and RT reactions each included 10 units of ribonuclease inhibitor protein (Ambion, Austin, TX). RNA was then degraded with RNase A and PCR was performed using a forward primer that binds to the cDNA derived from the ribo-oligonucleotide adaptor (Primers RLM-RACE deoxyribo, Supplementary Table 1) and a low degeneracy reverse primer within the N-terminal domain that was universal for all three genes and also nested relative to the RT primer (Primer Sp-N 5' RACE R2, Supplementary Table 1). The PCR products were then gel purified, cloned, and sequenced.

Sequence Analyses

Sequences alignments and general sequence manipulations were performed in Vector NTI 10 (Invitrogen). Prediction of the promoters in *MaSp1A* and *MaSp2* were done by Neural Network Promoter Prediction v2.2 (Reese, 2001). SignalP 3.0 was used to predict signal peptides (Bendtsen *et al.*, 2004). Secondary structure prediction was done using Hierarchical Neural Network (Combet *et al.*, 2000). Codon usage was determined with CountCodon (Nakamura *et al.*, 2000). BLASTP using BLOSUM62 matrix was used to determine protein similarity scores and ClustalW was used to evaluate amino acid substitutions as conservative (same size and polarity), semi-conservative (same size or polarity) or non-conservative (not same size or polarity). Trees were constructed by first aligning the nucleotide sequences for *MaSp* N-terminal and C-terminal domains in ClustalW. Next, MEGA4 was used to generate trees using the following parameters: Neighbor Joining, complete deletion of gaps, 1000 bootstraps, p-distance model (Tamura *et al.*, 2007). Accession numbers for all sequences presented in the trees are compiled in Supplementary Table 2.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

W. R. Marcotte Jr acknowledges financial support through DHHS-NIH-NIBIB. W. A. Gaines IV acknowledges financial support through a SC Agricultural Experiment Station New Generation Graduate Fellowship, a Clemson University College of Agriculture, Forestry, and Life Sciences Wade Stackhouse Graduate Fellowship and a Sigma Xi Grant-in-Aid of Research. We would like to thank Jackie Palmer and Jeff Vernon for assistance with spider collection. Technical Contribution No. 5500 of the Clemson University Experiment Station.

Literature Cited

- Araujo D, Cella MD, Brescovit DA. Cytogenetic analysis of the neotropical spider *Nephilengys cruentata* (Araneomorphae, Tetragnathidae): standard staining norms, C-bands and base-specific fluorochromes. *Braz J Biol.* 2005; 65:193–202. [PubMed: 16097721]
- Ayoub NA, Garb JE, Tinghitella RM, Collin MA, Hayashi CY. Blueprint for a high-performance biomaterial: full-length spider dragline silk genes. *PLoS ONE.* 2007; 2:e514. [PubMed: 17565367]
- Ayoub NA, Hayashi CY. Multiple recombining loci encode MaSp1, the primary constituent of dragline silk, in widow spiders (*Latrodectus*: Theridiidae). *Mol Biol Evol.* 2008; 25:277–286. [PubMed: 18048404]
- Beckwitt R, Arcidiacono S. Sequence conservation in the C-terminal region of spider silk proteins (Spidroin) from *Nephila clavipes* (Tetragnathidae) and *Araneus bicentenarius* (Araneidae). *J Biol Chem.* 1994; 269:6661–6663. [PubMed: 8120021]
- Bendtsen JD, Nielsen H, von Heijne G, Brunak S. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol.* 2004; 340:783–795. [PubMed: 15223320]
- Blackledge TA, Hayashi CY. Silken toolkits: biomechanics of silk fibers spun by the orb web spider *Argiope argentata*. *J Exp Biol.* 2006; 209:2452–2461. [PubMed: 16788028]
- Combet C, Blanchet C, Geourjon C, Deleage G. NPS@: network protein sequence analysis. *Trends Biochem Sci.* 2000; 25:147–150. [PubMed: 10694887]
- Gatesy J, Hayashi C, Motriuk D, Woods J, Lewis R. Extreme diversity, conservation, and convergence of spider silk fibroin sequences. *Science.* 2001; 291:2603–2605. [PubMed: 11283372]
- Gosline JM, Guerette PA, Ortlepp CS, Savage KN. The mechanical design of spider silks: from fibroin sequence to mechanical function. *J Exp Biol.* 1999; 202:3295–3303. [PubMed: 10562512]
- Hayashi CY, Lewis RV. Evidence from flagelliform silk cDNA for the structural basis of elasticity and modular nature of spider silks. *J Mol Biol.* 1998; 275:773–784. [PubMed: 9480768]
- Hayashi CY, Shipley NH, Lewis RV. Hypotheses that correlate the sequence, structure, and mechanical properties of spider silk proteins. *Int J Biol Macromol.* 1999; 24:271–275. [PubMed: 10342774]
- Hedhammar M, Rising A, Grip S, Martinez AS, Nordling K, Casals C, Stark M, Johansson J. Structural properties of recombinant nonrepetitive and repetitive parts of major ampullate spidroin 1 from *Euprosthenops australis*: implications for fiber formation. *Biochemistry.* 2008; 47:3407–3417. [PubMed: 18293938]
- Hijirida DH, Do KG, Michal C, Wong S, Zax D, Jelinski LW. ¹³C NMR of *Nephila clavipes* major ampullate silk gland. *Biophys J.* 1996; 71:3442–3447. [PubMed: 8968613]
- Hinman MB, Lewis RV. Isolation of a clone encoding a second dragline silk fibroin. *Nephila clavipes* dragline silk is a two-protein fiber. *J Biol Chem.* 1992; 267:19320–19324. [PubMed: 1527052]
- Huemmerich D, Helsen CW, Quedzuweit S, Oschmann J, Rudolph R, Scheibel T. Primary structure elements of spider dragline silks and their contribution to protein solubility. *Biochemistry.* 2004; 43:13604–13612. [PubMed: 15491167]
- Ittah S, Cohen S, Garty S, Cohn D, Gat U. An essential role for the C-terminal domain of a dragline spider silk protein in directing fiber formation. *Biomacromolecules.* 2006; 7:1790–1795. [PubMed: 16768399]

- Johnson WC. Analyzing protein circular dichroism spectra for accurate secondary structures. *Proteins*. 1999; 35:307–312. [PubMed: 10328265]
- Knight DP, Knight MM, Vollrath F. Beta transition and stress-induced phase separation in the spinning of spider dragline silk. *Int J Biol Macromol*. 2000; 27:205–210. [PubMed: 10828366]
- Knight DP, Vollrath F. Changes in element composition along the spinning duct in a *Nephila* spider. *Naturwissenschaften*. 2001; 88:179–182. [PubMed: 11480706]
- Lewis RV. Spider silk: ancient ideas for new biomaterials. *Chem Rev*. 2006; 106:3762–3774. [PubMed: 16967919]
- Madsen B, Shao ZZ, Vollrath F. Variability in the mechanical properties of spider silks on three levels: interspecific, intraspecific and intraindividual. *Int J Biol Macromol*. 1999; 24:301–306. [PubMed: 10342779]
- Maruyama K, Sugano S. Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene*. 1994; 138:171–174. [PubMed: 8125298]
- Miyasaka H. Translation initiation AUG context varies with codon usage bias and gene length in *Drosophila melanogaster*. *J Mol Evol*. 2002; 55:52–64. [PubMed: 12165842]
- Motriuk-Smith D, Smith A, Hayashi CY, Lewis RV. Analysis of the conserved N-terminal domains in major ampullate spider silk proteins. *Biomacromolecules*. 2005; 6:3152–3159. [PubMed: 16283740]
- Nakamura Y, Gojobori T, Ikemura T. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res*. 2000; 28:292. [PubMed: 10592250]
- Pan HC, Song DX, Zhou KY, Zhu GP. Cloning and prokaryotic expression of major ampullate spidroin gene of spider. *Sheng Wu Gong Cheng Xue Bao*. 2007; 23:446–451. [PubMed: 17577991]
- Perdrix-Gillot S. DNA synthesis and endomitoses in the giant nuclei of the silk gland of *Bombyx mori*. *Biochimie*. 1979; 61:171–204. [PubMed: 465570]
- Rasch EM, Connelly BA. Genome size and endonuclear DNA replication in spiders. *J Morphol*. 2005; 265:209–214. [PubMed: 15971267]
- Reese MG. Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. *Comput Chem*. 2001; 26:51–56. [PubMed: 11765852]
- Rising A, Johansson J, Larson G, Bongcam-Rudloff E, Engström W, Hjälm G. Major ampullate spidroins from *Euprosthenops australis*: multiplicity at protein, mRNA and gene levels. *Insect Mol Biol*. 2007; 16:551–561. [PubMed: 17680798]
- Rising A, Hjälm G, Engstrom W, Johansson J. N-terminal nonrepetitive domain common to dragline, flagelliform, and cylindriiform spider silk proteins. *Biomacromolecules*. 2006; 7:3120–3124. [PubMed: 17096540]
- Rising A, Nimmervoll H, Grip S, Fernandez-Arias A, Storckenfeldt E, Knight DP, Vollrath F, Engstrom W. Spider silk proteins-mechanical property and gene sequence. *Zoolog Sci*. 2005; 22:273–281. [PubMed: 15795489]
- Selden PA, Gall JC. A Triassic Mygalomorph spider from the Northern Vosges, France. *Palaeontology*. 1992; 35:211–235.
- Sponner A, Schlott B, Vollrath F, Unger E, Grosse F, Weisshart K. Characterization of the protein components of *Nephila clavipes* dragline silk. *Biochemistry*. 2005a; 44:4727–4736. [PubMed: 15779899]
- Sponner A, Unger E, Grosse F, Weisshart K. Conserved C-termini of Spidroins are secreted by the major ampullate glands and retained in the silk thread. *Biomacromolecules*. 2004; 5:840–845. [PubMed: 15132670]
- Sponner A, Vater W, Rommerskirch W, Vollrath F, Unger E, Grosse F, Weisshart K. The conserved C-termini contribute to the properties of spider silk fibroins. *Biochem Biophys Res Commun*. 2005b; 338:897–902. [PubMed: 16253207]
- Sreerama N, Woody RW. Estimation of protein secondary structure from circular dichroism spectra: comparison of CONTIN, SELCON, and CDSSTR methods with an expanded reference set. *Anal Biochem*. 2000; 287:252–260. [PubMed: 11112271]

- Stark M, Grip S, Rising A, Hedhammar M, Engstrom W, Hjalm G, Johansson J. Macroscopic fibers self-assembled from recombinant miniature spider silk proteins. *Biomacromolecules*. 2007; 8:1695–1701. [PubMed: 17402782]
- Tai PL, Hwang GY, Tso IM. Inter-specific sequence conservation and intra-individual sequence variation in a spider silk gene. *Int J Biol Macromol*. 2004; 34:295–301. [PubMed: 15556231]
- Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol*. 2007; 24:1596–1599. [PubMed: 17488738]
- Xu M, Lewis RV. Structure of a protein superfiber: spider dragline silk. *Proc Natl Acad Sci U S A*. 1990; 87:7120–7124. [PubMed: 2402494]

Figure 1. Sequences of *N. clavipes* *MaSp1A*, *MaSp1B*, *MaSp2*. (A) 5' nucleotide sequences from the TATA box (underlined) to the start codon (bold). Predicted transcription start site (TSS) is marked by asterisk. (B) Aligned translations of the MaSp N-terminal non-repetitive domains. Predicted signal peptide is boxed and predicted helices are highlighted. (C) Translations of the 5' end of the repeat array. GGX motifs (where X = Q, Y, L, A) are underlined, prolines are bold, and characteristic Sp2 motifs (GPG, GPS, QQ) are highlighted. Note that the start codon in part A corresponds to the 1st amino acid of part B, and that the last amino acid shown in part B immediately precedes the 1st amino acid of part C. Sequences have been submitted to GenBank under Accession Nos. EU599238, EU599239, EU599240, EU599241, EU599242 and EU599243.

Figure 2.

5' RLM-RACE of *MaSp-N* using *N. clavipes* major ampullate gland RNA. +TAP was treated with tobacco acid pyrophosphatase and -TAP was not. The expected product size for all 3 *MaSp-N* domains is ~300 bp. Marker sizes are indicated on the left.

Figure 3.

Changes in (A) *MaSp1A-N* and (B) *MaSp1B-N* abundance caused by variant-specific priming at *MaSp1-C* domains. V4 is template primed from the *MaSp1-C* Var4, V5 from the *MaSp1-C* Var5, and NP is a no priming control. Numbers represent the mean change in chromatogram peak areas relative to the no priming control. (C) The proposed linkage between the two *MaSp1* N-terminal and C-terminal domains as suggested by the graphs above.

Figure 4. Nucleotide distance trees of *MaSp*-N (A) and *MaSp*-C (B) sequences. Trees were generated by MEGA 4 using neighbor joining method and complete deletion of gaps. Numbers at nodes indicate bootstrap and scale bar represents sequence distance. Sequence groups for which there is strong evidence of duplicated *MaSp* genes are marked with double asterisk (**) and sequences which may represent duplicated genes or allele variants are marked with single asterisk (*). The novel sequences presented in this paper are boxed. Species key: Nc: *Nephila clavipes*, Nm: *N. inaurata madagascariensis*, Ns: *N. senegalensis*, Np: *N. pilipes*, Na: *N. antipodiana*, Aa: *Argiope amoena*, At: *Argiope trifasciata*, Lh: *Latrodectus hesperus*, Lg: *Latrodectus geometricus*, Ea: *Euprosthenops australis*.

Table 1*MaSp1* N- and C-terminal domains present in seven *N. clavipes* spiders^a

Spider	<i>Sp1A</i> -N	<i>Sp1B</i> -N	<i>Sp1-C</i> Var4	<i>Sp1-C</i> Var5
1	+(C)	+	+(C)	+(G,T)
2	+(C,A)	+	+ ns	+ ns
3	+(C,A)	+	+(G)	+(G)
4	+(C)	+	+ ns	+ ns
5	+(C,A)	+	+ ns	+ ns
6	+(C,A)	+	+ ns	+ ns
7	+(C,A)	+	+ ns	+ ns

^a += domain is present; C,A,G,T = SNP allele is present; ns = not sequenced