Extreme Diversity, Conservation, and Convergence of Spider Silk Fibroin Sequences

John Gatesy,*†‡ Cheryl Hayashi,*†‡ Dagmara Motriuk, Justin Woods, Randolph Lewis

Spiders (Araneae) spin high-performance silks from liquid fibroin proteins. Fibroin sequences from basal spider lineages reveal mosaics of amino acid motifs that differ radically from previously described spider silk sequences. The silk fibers of Araneae are constructed from many protein designs. Yet, the repetitive sequences of fibroins from orb-weaving spiders have been maintained, presumably by stabilizing selection, over 125 million years of evolutionary history. The retention of these conserved motifs since the Mesozoic and their convergent evolution in other structural superproteins imply that these sequences are central to understanding the exceptional mechanical properties of orb weaver silks.

There are over 34,000 described species of Araneae (*1*). Each species uses silk, and some ecribellate orb weavers (Araneoidea) have a varied tool kit of task-specific silks with divergent mechanical properties (*2*). Araneoid major ampullate silk, the primary dragline, is extremely tough. Minor ampullate silk, used in web construction, has high tensile strength. An orb web's capture spiral, in part composed of flagelliform silk, is stretchy and can triple in length before breaking (*3*). Each of these fibers is composed of one or more proteins encoded by the spider silk fibroin gene family (*4*). Previously sequenced araneoid fibroins are dominated by iterations of four simple amino acid motifs: polyalanine (A_n) (*n*, iterations of amino acid or motif), alternating glycine and alanine (GA), GGX (where X represents a small subset of amino acids), and $GPG(X)_n$ (P, proline) (*5*). However, orb weavers represent only a fraction of spider diversity (*1*). The silks of non-araneoid spiders have not been characterized at the DNA sequence level (Fig. 1).

Spiders draw fibers from dissolved fibroin proteins that are stored in specialized sets of abdominal glands. Recently, it has been suggested that this spinning mechanism may impact the generation of high-performance silk fibers more than the sequences of the fibroins themselves (*2, 6*). If this was so, amino acid sequences of different silk proteins should not have been conserved over long evolutionary intervals. To this point, comparative analyses of spider fibroins have been limited, so the evolutionary stability of these molecules is unclear.

DNA sequences have been described for only two genera of orb weavers (Fig. 1). To expand this database, we made seven cDNA libraries of silk glands from five spider genera. cDNA data were supplemented by information from two genomic libraries and polymerase chain reaction (PCR)–amplified sequences (*7*). We obtained partial cDNA or gene sequences for 23 fibroins from six families of Araneae. These data greatly extend the phylogenetic diversity of characterized fibroins (Fig. 1).

Like previously published fibroin sequences from spiders (*4, 5, 8, 9*) and lepidopterans (*10, 11*), our sequences encode repetitive alanine- and glycine-rich proteins. In each molecule, iterated amino acid motifs are organized into higher order ensemble repeats. Ensemble repeats within each fibroin were aligned, and a consensus ensemble repeat was generated for each molecule (*12*).

In part, silk DNA sequences from non-araneoid spiders (Fig. 2) reiterate the importance of amino acid motifs that compose orb weaver fibroins. GA, GGX, and A_n form the consensus ensemble repeat units of silk fibroins from the pisaurid fishing spider, *Dolomedes*. The association of these three motifs in *Dolomedes* silk proteins mirrors the pattern seen in major and minor ampullate fibroins of orb weavers (Fig. 3). GA, GGX, and A_n motifs also are distributed, sometimes sparsely, among ensemble repeat units from successively more basal lineages of spiders (Haplogynae and Mygalomorphae). A_n is represented in each of the fibroins from these taxa and from all lineages of Araneae studied thus far (Figs. 2 and 3). Mygalomorphae, tarantulas and close relatives, diverged from Araneomorphae, "true" spiders, at least 240 million years ago in the Middle Triassic (*13*) (Fig. 1); thus, A_n motifs may have been maintained in different spider silks since that time.

Although the fibroins of *Plectreurys* (Haplogynae) and *Euagrus* (Mygalomorphae) are internally repetitive, the ensemble repeats from these basal taxa (Fig. 2) are unlike analogous units from previously described silks (Fig. 3). Each of the fibroins from these primitive groups contains runs of serine. *Plectreurys* cDNA1 is highly internally repetitive with iterations of A_n, S_n, $(GX)_n$, and $(AQ)_n$ (S, serine; Q, glutamine). *Plectreurys* cDNA3 has a unique molecular architecture, with the 5′ end encoding a tandem array of long repeat units and the 3′ end encoding 15 repeats of a much shorter ensemble unit. The ~344–amino acid *Euagrus* repeat unit is a complex mixture of serine and alanine–rich sequence that includes a string of threonine, an amino acid that is rare in araneoid fibroins (Fig. 2).

Aside from an overall modular structure, scattered GA, GGX, and A_n motifs, and amino acid matches in the nonrepetitive COOH-terminus, there is only limited sequence similarity between araneoid fibroins and those from *Plectreurys* and *Euagrus* (Figs. 2 and 3). Perhaps these differences are not surprising given the contrasting ecologies of these species (*1, 14*), but our data clearly indicate that spiders use a diversity of proteins to spin silk threads. The fibroin repeats of basal Araneae suggest that spider silk design may not be especially dependent on specific sequences, but comparisons of fibroins among orb weavers contradict this notion (Fig. 3).

In combination with published data (*4, 5, 8, 9*), our sequences allow comparisons between the two basal-most clades of ecribellate orb weavers, Araneidae and "derived araneoids" (Fig. 1), for four groups of fibroins (*15*): major ampullate spidroin 1–like (MaSp1), major ampullate spidroin 2–like (MaSp2), minor ampullate spidroins (MiSp), and flagelliform silk protein (Flag). Differences among fibroins within each of these four groups are primarily variations in the arrangement and frequency of A_n, GA, GGX, and $GPG(X)_n$ motifs (Fig. 3). A_n, GA, and GGX are present in consensus repeats for both araneid and derived araneoid MiSp orthologs. Major ampullate fibroins are similarly conserved among araneoids. Stable repeats for MaSp1 are A_n, GA, and GGX, and stable repeats for MaSp2 are $GPG(X)_n$ and A_n. These motifs are retained even in major ampullate fibroins of the widow *Latrodectus*, a cobweb-weaving araneoid that does not spin a conventional orb web. The long Flag repeats are divergent within Araneoidea, but

Department of Molecular Biology, University of Wyoming, Laramie, WY 82071, USA.

*Present address: Department of Biology, University of California, Riverside, CA 92521, USA.
†These authors contributed equally to this work.
‡To whom correspondence should be addressed.

both araneid and derived araneoid repeat units are composed primarily of clustered GPG(X)$_n$ and GGX motifs (Fig. 3).

Fossil evidence suggests that the divergence of Araneidae from derived araneoids occurred no later than the Early Cretaceous (Fig. 1). Therefore, the motifs conserved within MaSp1, MaSp2, MiSp, and Flag have been maintained, presumably by stabilizing selection, for over 125 million years (*16*). Motifs that have been retained over such long evolutionary periods are likely to be critical to the divergent mechanical properties of the specialized orb weaver silks.

All of the motifs that are conserved in araneoid fibroins are found in other structural proteins that also have exceptional mechanical properties. There is strong evidence that A$_n$ and GA regions create β sheets in silk fibers, and these crystalline structures are hypothesized to be critical for the high tensile strength of spider silk (*17*). A$_n$ and GA are found in several proteins that also are characterized by high strength: lepidopteran fibroins from *Bombyx*, *Antheraea*, and *Galleria* (*10, 11*); oyster shell matrix protein (*18*); and mussel byssus proteins (*19*). Furthermore, GPG(X)$_n$ has been implicated as the elastic module of MaSp2 and Flag (*5*), and analogous PG motifs confer stretchiness to wheat gluten and elastin (*20*). GPG(X)$_n$ is duplicated in proteins of the elastic byssal threads of mussels (*19*). The structure of GGX motifs in spider silk currently is not clear (*21*). Regardless, GGX motifs are maintained in MaSp1, MiSp, and Flag and are distributed widely among *Antheraea* and *Galleria* fibroins, mussel byssus proteins, oyster shell matrix protein, and abductin (a component of the hinge ligament that connects molluskan shells) (*22*). The convergent use of A$_n$, GA, GGX, and GPG(X)$_n$ in other structural superproteins is further evidence for the importance of these motifs to the mechanical properties of different araneoid silks.

Many arthropods produce silk, but silk use has a widely scattered phylogenetic distribution (*14*). The profoundly different modes of silk production in lepidopteran insects (labial glands) versus spiders (abdominal glands) suggest that silks from these taxa evolved convergently (*23*). It also would be difficult to argue that the structural proteins used by bivalve mollusks are closely related to spider fibroins. These "aquatic silks" likely evolved common motifs in parallel with terrestrial arthropod fibroins. Such evolutionary convergences are documented regularly at the gross anatomical level, but comparable molecular examples are rare [e.g., (*24*)].

Recent attempts at artificially spinning orb weaver fibroins have not been successful in replicating the strength and toughness of native silk fibers (*25*). Furthermore, the mechanical properties of spider silks have been shown to vary with the spinning rate (*26*). However, to suggest that the drawing process, rather than the primary sequences of fibroins, exerts the dominant influence on fiber formation and properties (*2, 6*) is an overstatement. Evidently, both factors are important. Four simple amino acid motifs account for most repetitive sequences in characterized araneoid fibroins. The retention of these motifs in spider silks since the Mesozoic implies that fibroin sequences are critical to any detailed understanding of the elasticity, toughness, and strength of different araneoid silk fibers. The convergent evolution of these motifs in other structural superproteins reiterates the importance of these sequences. However, clusters of the amino acid motifs conserved in orb weaver silks make up only a

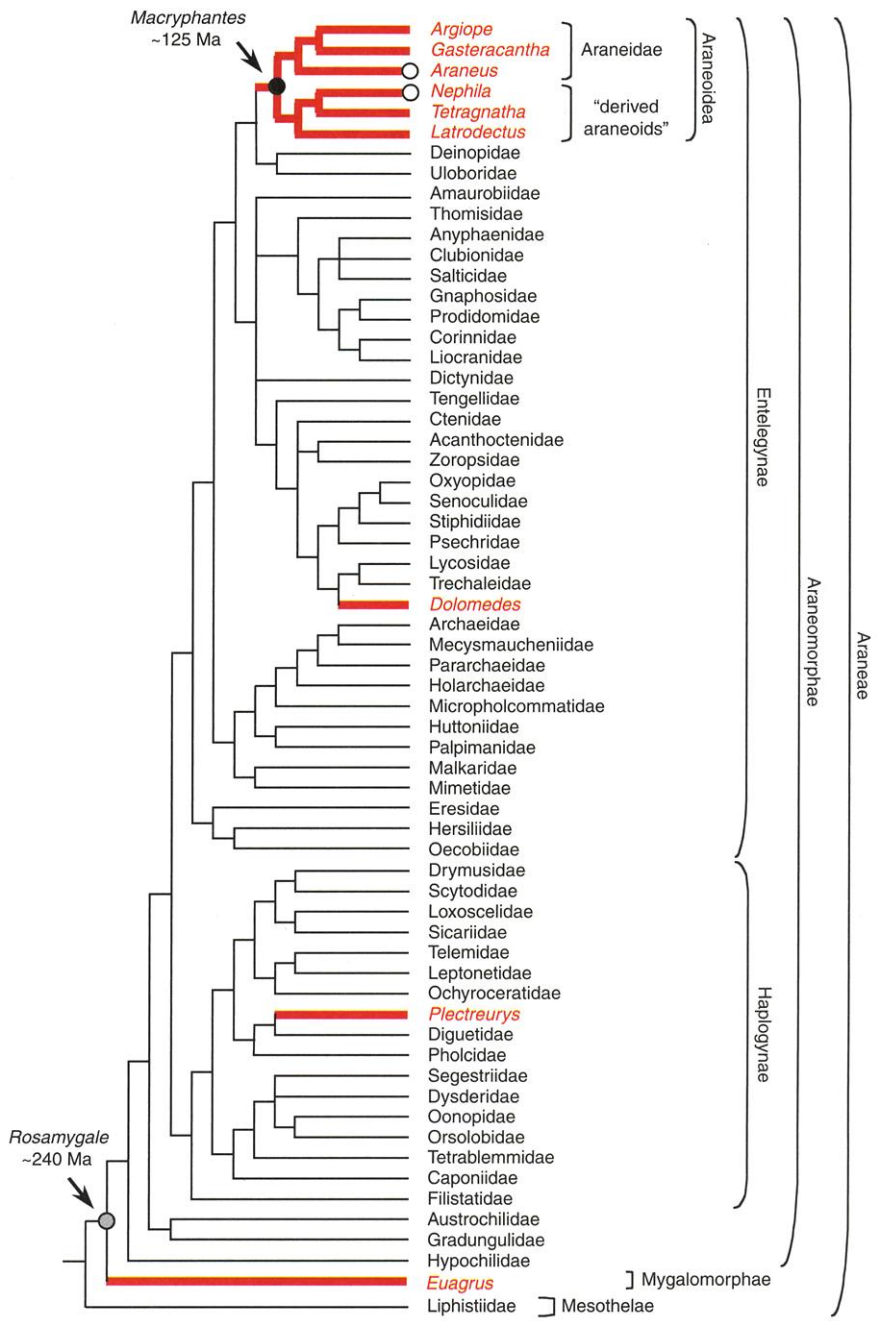

Fig. 1. Hypothesized phylogenetic relationships of Araneae based on morphological evidence (*1, 28*). Previously published spider fibroin sequences are from the two genera marked by white circles. Including data presented here, fibroin sequences have been characterized for the taxa in red. Circles at internal nodes mark fossil calibration points. Extinct taxa that calibrate these nodes, *Macryphantes* (black circle) (*16*) and *Rosamygale* (gray circle) (*13*), are indicated, and higher level taxa are to the right of the brackets (Ma, millions of years ago). *Dolomedes*, *Plectreurys*, and *Euagrus* are from the families Pisauridae, Plectreuridae, and Dipluridae, respectively.

Fig. 2. Consensus ensemble repeat units for non-araneoid spider fibroins. Single-letter symbols for amino acids are used (*29*), and GGX, GA, and A_n motifs are indicated in green, yellow, and red, respectively. *Plectreurys* cDNA1 and *Plectreurys* cDNA2 were derived from the larger ampule-shaped glands of *Plectreurys*, and *Plectreurys* cDNA3 and *Plectreurys* cDNA4 were from the smaller ampullate glands of this spider.

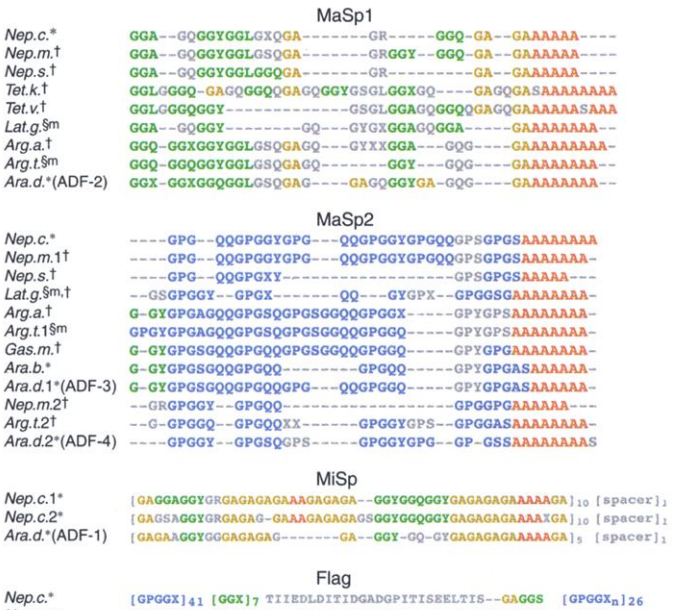

Fig. 3. Consensus ensemble repeat units for four araneoid fibroin ortholog groups. Single-letter symbols for amino acids are used (*29*), and GGX, GA, A_n, and GPG(X)$_n$ motifs are indicated in green, yellow, red, and blue, respectively. Dashes indicate gaps inserted into the alignments of ensemble repeat units for MaSp1, MaSp2, and MiSp. The "[spacer]" region of the MiSp fibroins is a serine-rich sequence that is 137 amino acids long in *Nephila clavipes* (GenBank accession number AF027735). Abbreviations are as follows: *Nep.c.*, *Nephila clavipes*; *Nep.m.*, *N. madagascariensis*; *Nep.s.*, *N. senegalensis*; *Tet.k.*, *Tetragnatha kauaiensis*; *Tet.v.*, *T. versicolor*; *Lat.g.*, *Latrodectus geometricus*; *Arg.a.*, *Argiope aurantia*; *Arg.t.*, *Argiope trifasciata*; *Ara.d.*, *Araneus diadematus*; *Gas.m.*, *Gasteracantha mammosa*; and *Ara.b.*, *Araneus bicentenarius*. Symbols indicate the following: *, previously published sequence; †, PCR/genomic clone; §m, cDNA from major ampullate glands; and §f, cDNA from flagelliform glands. The previous designations for *Araneus diadematus* fibroins (*4*) are shown in parentheses (ADF1 through ADF4).

fraction of the repetitive sequences in silk proteins from *Plectreurys* and *Euagrus*. These molecules from basal Araneae hint at the potential wealth of protein designs yet to be discovered in spider silks.

References and Notes

1. J. Coddington, H. Levi, *Annu. Rev. Ecol. Syst.* **22**, 565 (1991).
2. F. Vollrath, *Int. J. Biol. Macromol.* **24**, 81 (1999).
3. J. Gosline, P. Guerette, C. Ortlepp, K. Savage, *J. Exp. Biol.* **202**, 3295 (1999).
4. P. Guerette, D. Ginzinger, B. Weber, J. Gosline, *Science* **272**, 112 (1996).
5. C. Hayashi, R. Lewis, *J. Mol. Biol.* **275**, 773 (1998).
6. P. Calvert, *Nature* **393**, 309 (1998).
7. cDNA libraries were made from major ampullate glands of *Argiope trifasciata* (Araneidae) and *Latrodectus geometricus* (Theridiidae), flagelliform glands of *Argiope trifasciata*, ampullate glands of *Dolomedes tenebrosus* (Pisauridae), two sets of silk glands from *Plectreurys tristis* (Plectreuridae), and silk glands of the mygalomorph *Euagrus chisoseus* (Dipluridae). Glands from *Dolomedes* were the four pairs of spindly ampullate glands with long tails that are connected to the spinnerets by extensive looped ducts. Glands from *Plectreurys* were the two largest pairs of ampule-shaped glands. The relatively uniform silk glands of *Euagrus* were combined in the RNA extraction for this species. Genomic libraries were constructed for *Nephila madagascariensis* (Tetragnathidae) and for *Argiope trifasciata*. Data from the nine libraries were augmented by PCR-amplified genomic sequences from eight araneoids (Fig. 3). Sequences were submitted to GenBank (accession numbers AF350262 through AF350286). Specific methods are given as supplementary information (*27*)
8. R. Beckwitt, S. Arcidiacono, R. Stote, *Insect Biochem. Mol. Biol.* **28**, 121 (1998).
9. GenBank accession numbers for the sequences are as follows: M37137, M92913, AF027735 through AF027737, and AF218621 through AF218624.
10. K. Mita, S. Ichimura, T. James, *J. Mol. Evol.* **38**, 583 (1994).
11. GenBank accession numbers for the sequences are as follows: AF083334, AF095239, and AF095240.
12. DNA sequences for the various fibroins were translated into amino acid sequences. For each fibroin, ensemble repeat units (higher order aggregations of iterated sequence motifs) were aligned algorithmically then manually with MacVector (Oxford Molecular, Hunt Valley, MD), and a consensus ensemble repeat unit was generated. The consensus ensemble repeat was defined as the most common amino acid (or gap) at each position in the alignment of repeats for each molecule. When there was a tie at an alignment position, X denoted this ambiguity in the consensus sequence.
13. P. Selden, J. Gall, *Palaeontology* **35**, 211 (1992).
14. C. Craig, *Annu. Rev. Entomol.* **42**, 231 (1997).
15. Araneoid silk sequences initially were allocated to different fibroin ortholog groups in *Nephila clavipes* (*5*, *9*), according to sequence characteristics of ensemble repeat units. These assignments were assessed through phylogenetic analyses of nonrepetitive COOH-terminal sequences, expression patterns, fit to independent phylogenetic hypotheses for Araneoidea, and minimization of gene duplication events. Specific methods and interpretations of phylogenetic patterns are available as supplementary information (*27*).
16. P. Selden, *Palaeontology* **33**, 257 (1990).
17. A. Simmons, E. Ray, L. Jelinski, *Macromolecules* **27**, 5235 (1994).
18. S. Sudo *et al.*, *Nature* **387**, 563 (1997).
19. X. Qin, K. Coyne, J. Waite, *J. Biol. Chem.* **272**, 32623 (1997).
20. A. Tatham, P. Shewry, B. Miflin, *FEBS Lett.* **177**, 205 (1984).
21. B. Thiel, K. Guess, C. Viney, *Biopolymers* **41**, 703 (1996).
22. Q. Cao, Y. Wang, H. Bayley, *Curr. Biol.* **7**, 677 (1997).
23. J. Schultz, *Biol. Rev.* **62**, 89 (1987).
24. L. Chen, A. DeVries, C. Cheng, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 3817 (1997).
25. A. Seidel, O. Liivak, L. Jelinski, *Macromolecules* **31**, 6733 (1998).
26. B. Madsen, Z. Shao, F. Vollrath, *Int. J. Biol. Macromol.* **24**, 301 (1999).
27. Supplementary information is available at www.sciencemag.org/cgi/content/full/291/5513/2603/DC1.
28. G. Hormiga, N. Scharff, J. Coddington, *Syst. Biol.* **49**, 435 (2000).
29. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; D, Asp; E, Glu; F, Phe; G, Gly; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; X, a small subset of amino acids, and Y, Tyr.

30. We thank G. Binford, T. Blackledge, R. Gillespie, S. Hayashi, M. Hedin, J. Jones, and A. Leroy for help in collecting spiders; M. Hinman for amino acid analyses; and A. de Queiroz, S. Gatesy, M. O'Leary, P. Langer, P. Selden, and two anonymous reviewers for improving the manuscript. This work was supported by grants from NSF (MCB-9806999) and the U.S. Army Research Office (DAAG55-98-1-0262).

17 November 2000; accepted 26 February 2001