Molecular Architecture and Evolution of a Modular Spider Silk Protein Gene

Cheryl Y. Hayashi* and Randolph V. Lewis

Spider flagelliform silk is one of the most elastic natural materials known. Extensive sequencing of spider silk genes has shown that the exons and introns of the flagelliform gene underwent intragenic concerted evolution. The intron sequences are more homogenized within a species than are the exons. This pattern can be explained by extreme mutation and recombination pressures on the internally repetitive exons. The iterated sequences within exons encode protein structures that are critical to the function of silks. Therefore, attributes that make silks exceptional biomaterials may also hinder the fixation of optimally adapted protein sequences.

Araneoid spiders are capable of spinning up to seven unique silks. Some of these silks are renowned high-performance fibers. For example, dragline silk has exceptional tensile strength whereas flagelliform silk, the elastic filament that forms the capture spiral of an orb-web, may have >200% extensibility (*1*). In *Nephila clavipes* (Araneae: Tetragnathidae), the gene that encodes flagelliform silk (*Flag*) is transcribed into an mRNA of about 15.5 kb (*2*). Most of Flag is composed of numerous iterations of three different amino acid motifs: $GPGG(X)_n$, GGX (G, Gly; P, Pro; X, other), and a 28-residue "spacer" (Fig. 1A). These motifs are organized into complex ensembles of about 440 amino acids that have similar tandem arrays of the glycine-rich motifs combined with a single spacer motif (Fig. 1B).

Here we report genomic sequences of *Flag* genes that encompass both 5′ and 3′ ends and substantial portions of the intervening repetitive region. We cloned partial *Flag* genes from genomic DNAs of *N. clavipes* (*N.c.*) and *Nephila madagascariensis* (*N.m.*) (*3*). The resulting 36 kb of sequence (GenBank accession nos. AF218621–AF218624) represents the most extensive DNAs known for any spider silk. In contrast to other spider silk genes, *Flag* is not encoded by a single enormous exon (*4, 5*). Instead, the *Flag* gene is evenly divided into exonic and intronic regions (*6*). In all, the *Flag* locus is estimated to span 30 kb and to contain 13 exons (Fig. 1D).

The first two exons encode the nonrepetitive amino-terminal region. The final exon contains both repetitive sequence and the nonrepetitive carboxy terminus (Fig. 1D). Exons 3 to 13 each encodes an individual ensemble repeat (Fig. 1, B to D). These repeated exons are of similar length [about 1320 base pairs (bp)] and identical organization. When these exons are aligned, the spacer motifs emerge as the most conserved sequences within and between species (Fig. 2A). The main differences among the repeated exons are the variable numbers of tandem $GPGG(X)_n$ and GGX motifs. In comparisons of species, only exons 3 and 13 of *N.c.* are more similar to the corresponding exons of *N.m.* than to the other repeated exons of *N.c.* (Fig. 2B). Even given these exceptions, on average the repeated exons are more alike within (73%) than between (68%) species at the DNA level. Given that both species have identical motifs and ensemble repeat structures, differential selection cannot easily account for this divergence pattern. Instead, such homogenization of sequence repeats within species is indicative of concerted evolution (*7*).

The introns separating the repeated exons also share high similarity (Fig. 1, C and D). Introns 3 to 12 are each about 1420 bp long and, in general, are easily aligned to one another (Fig. 3A) (*8*). Within species, the introns are on average 87% similar. The degree of homogenization in *N.c.* is extreme, with introns 5, 6, and 7 sharing 99.9% identity. Between species, the introns are much less similar (75%). Thus, the repeated introns are less divergent within a species than between species (Fig. 3B).

The corresponding exons and introns from *N.c.* and *N.m.* were compared to further examine between-species divergence of repetitive and nonrepetitive regions. The nonrepetitive exons 1 (591 bp) and 2 (405 bp) and the nonrepetitive 3′ portion of exon 13 (372 bp) are the most highly conserved coding sequences (Fig. 4). These regions differ by less than 5% between the two species. In contrast, the repeated exon sequences are five times more divergent between species. Thus, the repeated exons are not only homogenized within species but are also more divergent between species than the nonrepetitive coding sequences (Fig. 4).

Even with the high divergence among repeated exons, there is evidence of purifying selection. The most obvious indication is that the organization of motifs within the ensemble repeats is strictly maintained in both species. Each ensemble repeat has just one centrally located spacer that is flanked by $GPGG(X)_n$ and GGX repeats (Fig. 2A). More detailed comparison of the exons shows that the observed types of synonymous substitution are biased toward adenine and thymine, which is consistent with the substantial codon usage preferences present in spider silk genes (*2*, *4*, *9*). Furthermore, there are additional, apparent constraints on the observed types of nonsynonymous substitution. Most amino

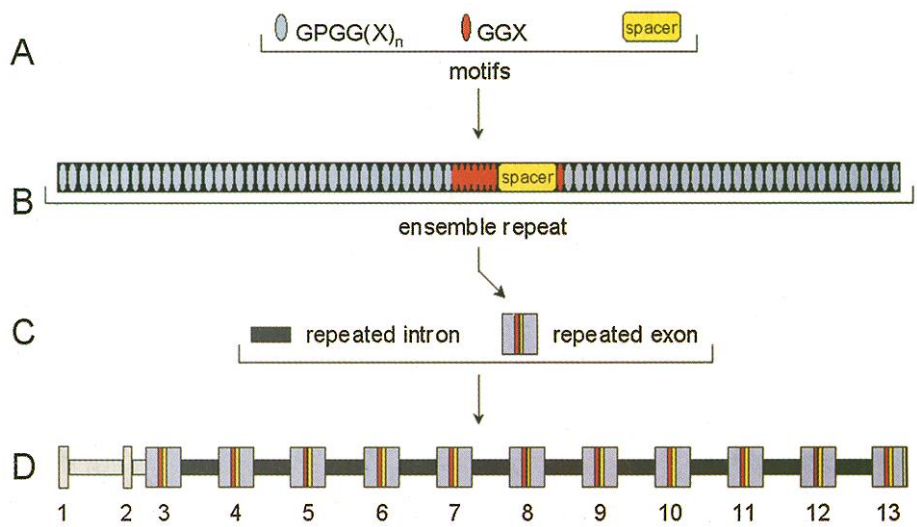

Fig. 1. The *Flag* gene contains hierarchical sets of components. (**A**) The repetitive coding region is composed of codons for three different amino acid sequence motifs. (**B**) Iterations of the three motifs are organized into complex ensemble repeats of about 440 amino acids. (**C**) Each ensemble repeat is encoded by a single exon. These repeated exons are separated by repeated introns. (**D**) The *Flag* gene spans about 30 kb. Exons and introns are numbered, and regions of nonrepetitive sequence are shaded gray.

Department of Molecular Biology, University of Wyoming, Laramie, WY 82071–3944, USA.

*To whom correspondence should be addressed. E-mail: hayashi@uwyo.edu

acid differences occur in the X positions of the GPGG(X)$_n$ and GGX motifs. Although these positions are the most variable among the repeats, they tend to be replaced from a very small subset of amino acids—namely, alanine, serine, tyrosine, and valine. The final apparent constraint is that insertions and deletions are often entire GPGG(X)$_n$ or GGX motifs. In fact, most of the divergence between the aligned exons is due to short insertions and deletions of these motifs and not point mutations (Fig. 2A).

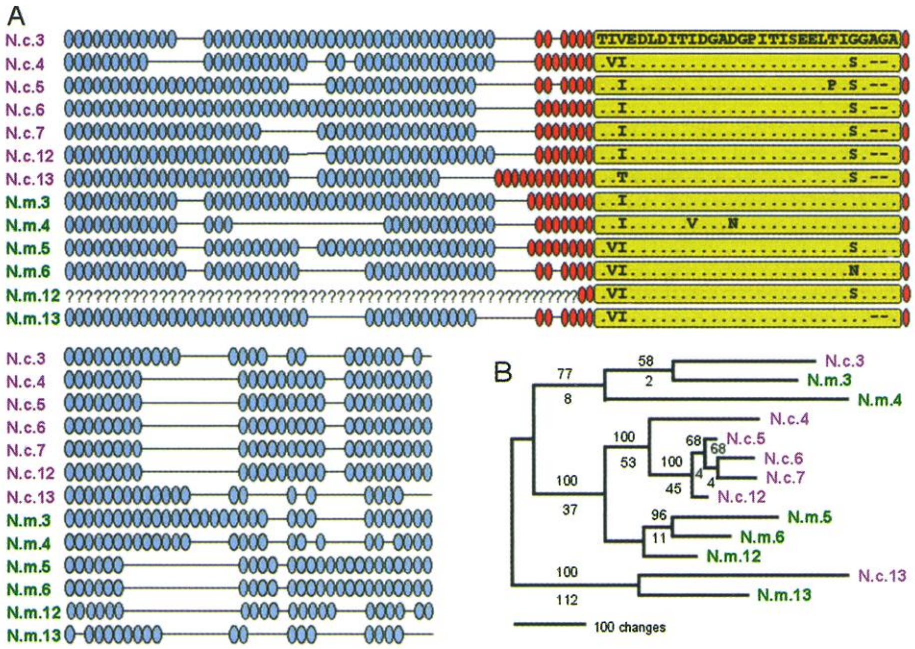

Fig. 2. Each ensemble repeat is encoded by an individual repeated exon. (**A**) The translated alignment of nucleotides (*14*) from *N.c.* and *N.m.* is depicted with the motif symbols from Fig. 1A and the numbering of exons in Fig. 1D. (**B**) Parsimony analysis (*15*) was done to show the similarity among exons by the possession of shared characters. The tree shown is one of the two shortest trees (2609 steps) that was chosen after one round of successive approximations (*16*). Bootstrap values (*17*) are shown above the internodes; Bremer support scores (*18*) are below. Branch lengths are proportional to the number of character changes. An additional 605 steps are required to make all the exons pair by corresponding number (e.g., N.c.5 with N.m.5).

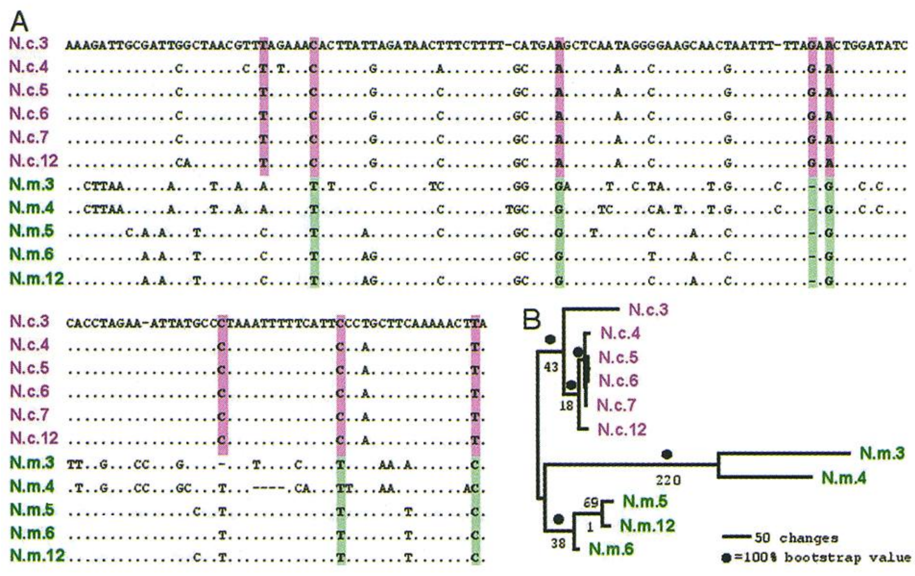

Fig. 3. Iterated introns are homogenized within species. (**A**) A portion of the alignment (*14*) is shown, and variable sites that are fixed within a species are highlighted. (**B**) Parsimony analysis (*15*) of the aligned intron sequences resulted in three trees (1158 steps), which differed only in the resolution of N.c.5, N.c.6, and N.c.7, that share 99.9% identity. Support indices and branch lengths are as in Fig. 2B. To force the pairing of introns by position in the gene (e.g., N.c.3 with N.m.3) required an additional 801 steps.

Because *Nephila* rely on their orb-webs for prey capture, there are strong selective forces that operate on Flag to produce a functional capture spiral. Thus, the high divergence of the repeated exons might be due to purely molecular mechanisms. The internally repetitive nature of the exons is likely to promote errors from slippage during replication. Furthermore, because codons for glycine, proline, and alanine compose almost 75% of the coding sequence, there is an abundance of guanine and cytosine bases. Strings of these base pairs are thought to create recombination hot spots (*5*, *10*). The resulting unequal crossover events could, in part, account for the length differences observed in the numbers of tandem GPGG(X)$_n$ and GGX motifs (Fig. 2A).

The large variation in length that has been observed in alleles of spider silk genes is further evidence for the prevalence of replication slippage and unequal crossing-over events (*5*). Many of the differences between alleles are insertions and deletions of ensemble repeats or shorter motifs that can be explained by replication slippage. Similarly, allelic length differences are common in the lepidopteran silkworm fibroin gene, which is also rich in cytosine and guanine and highly internally repetitive (*10*, *11*).

The repeated introns are more similar (87%) (*8*) within a species than are the repeated exons (73%) (Figs. 2B and 3B). The intron sequences could have been homogenized by recombination events, selection, or some combination of the two processes. A possible function of the intron sequences may be stabilization of the Flag precursor mRNA. However, the extreme 99.9% similarity of *N.c.* introns 5 to 7 is strong evidence for homogenization by recombination. To argue for selection alone would require the convergent evolution of hundreds of sequence changes among the introns of *N.c.* and *N.m.* (Fig. 3A). Thus, as with the exons, recombination and/or conversion events are likely to have limited the within-gene divergence of the repeated introns. However, in contrast to the exons, the introns are not internally repetitive and lack a sequence rich in cytosine and gua-

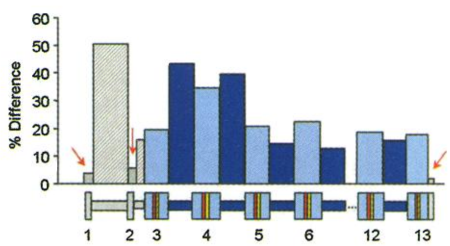

Fig. 4. Corresponding gene regions between *N.c.* and *N.m.* were aligned pairwise (*19*) and percent divergence is plotted. Colors of exons and introns are as in Fig. 1. Arrows point to the low divergence in exons 1 and 2 and to the carboxy terminal portion of exon 13.

nine. Without these factors that promote replication slippage and unequal crossovers, the introns exhibit less divergence than the repeated exons and tend to require fewer alignment gaps (Figs. 2B and 3).

The greater divergence of the *Flag* repetitive exons relative to the introns could be due to different directional selection regimes operating on the Flag fibers of *N.c.* versus *N.m.* If so, then a unique sequence element that is advantageous to one species is expected to be present in all of its repeated exons and absent from the repeated exons of another species. Given that there are very few fixed differences among the ensemble repeats within *N.c.* or *N.m.*, directional selection does not seem to account for rapid sequence divergence relative to the repeated introns. Also, purifying selection on the silk protein structure is reflected in the strict maintenance of the ensemble organization (Fig. 2A). Instead of speeding up evolution, functional constraints should decrease sequence divergence of the exons. Thus, the greater exonic divergence still can be best explained by the molecular architecture of the *Flag* gene.

Flag was known to be a modular protein with three basic motifs composing a large ensemble repeat. The genomic organization of the *Flag* gene suggests a new hierarchical level of modularity. Not only are the ensemble repeats encoded by repeated exons, but the intervening introns are also iterated copies. This molecular architecture results in efficient within-gene concerted evolution. Probably through some combination of gene conversion and unequal crossing-over at repetitive exons, *Flag* remains fairly homogenized over its entire 15,500-bp coding sequence. However, this same highly repetitive architecture apparently prevents the coding sequences of *Flag* from completely homogenizing. The evolution of the *Flag* gene represents a case in which homogenization of repeats through purifying selection and recombination is offset by mutational mechanisms inherent in the basic structure of the DNA sequences (*12*). Thus, the repetitive genetic architecture of spider silk encourages sequence homogenization as well as rapid sequence divergence. This conflict has implications for the interpretation of high-performance silks as optimally adapted supermolecules (*1*, *13*).

References and Notes

1. J. Gosline, M. DeMont, M. Denny, *Endeavour* **10**, 37 (1986).
2. C. Hayashi and R. Lewis, *J. Mol. Biol.* **275**, 773 (1998).
3. A λFixII (Stratagene) library of *N.c.* genomic DNA was a gift from M. Hinman. A λGem-12 (Promega) library was constructed from *N.m.* genomic DNA. Libraries were screened with the radiolabeled oligonucleotide CCWCCWGGWCCNNNWCCWCCWGGWCC (W = A or T; N = A, G, C, or T). *Flag* inserts were subcloned into pGEM (Promega) vectors and sequenced in both directions with universal or gene-specific primers. To sequence through long, highly repetitive regions, sets of nested deletions were created with the Erase-A-Base kit (Promega), or transposons were inserted using the Genome Priming System (New England Biolabs). An additional 2.8 kb of the *Flag* gene from *N.c.* was amplified by polymerase chain reaction with the primers CGCTTCTGAAACGAAAAAGG and GCGAACATTCTTCCTACAGA, ligated into pGEM3z-f(+) (Promega) and duplicate clones were sequenced as described above.
4. M. Xu and R. Lewis, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 7120 (1990); M. Hinman and R. Lewis, *J. Biol. Chem.* **267**, 19320 (1992); M. Colgin and R. Lewis, *Protein Sci.* **7**, 667 (1998).
5. R. Beckwitt, S. Arcidiacono, R. Stote, *Insect Biochem. Mol. Biol.* **28**, 121 (1998).
6. *Flag* gene exons and introns can be distinguished by several criteria. First, five exons directly correspond to the partial cDNAs Flag5′ (GenBank accession no. AF027972) and Flag3′ (GenBank accession no. AF027973). Second, all introns are flanked by the typical GT and AG boundary sequences. Third, exons could be continuously translated in only one reading frame, whereas introns could not be translated for any appreciable length in any reading frame. cDNA data (*2*) were used in analyses for *N.c.* exons 1 and 12.
7. E. Zimmer, S. Martin, S. Beverly, Y. Kan, A. Wilson, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 2158 (1980). Conservation of 5′ and 3′ terminal members of the repeated exons is consistent with some genetic models [M. Lassner and J. Dvorak, *Nucleic Acids Res.* **14**, 5499 (1986)].
8. The only exceptions are introns 3 and 4 of *N.m.* These introns have areas of high similarity to the other introns but also contain divergent regions that are difficult to align. Phylogenetic analysis and pairwise distances show that these introns are more similar to each other than to any of the other repeated introns (Fig. 3B). Despite their divergent sequences, these two introns are more similar to the other *N.m.* introns than to the *N.c.* introns. Introns 3 and 4 in *N.m.* are less homogenized than in *N.c.*, so the degree of homogenization differs between species.
9. P. Guerette, D. Ginzinger, B. Weber, J. Gosline, *Science* **272**, 112 (1996).
10. K. Mita, S. Ichimura, T. James, *J. Mol. Evol.* **38**, 583 (1994).
11. R. Manning and L. Gage, *J. Biol. Chem.* **255**, 9451 (1980).
12. G. Dover, *Nature* **299**, 111 (1982); G. Dover, A. Linares, T. Bowen, J. Hancock, *Methods Enzymol.* **224**, 525 (1993).
13. F. Vollrath, *Sci. Am.* **266** (3), 70 (1992); S. Osaki, *Nature* **384**, 419 (1996); F. Vollrath, *Int. J. Biol. Macromol.* **24**, 81 (1999).
14. Multiple alignments were constructed with MALIGN, v.2.1 [W. Wheeler and D. Gladstein (American Museum of Natural History, New York, 1994)] and adjusted with SeqApp, v.1.9a [D. Gilbert (University of Indiana, Bloomington, 1992)] to consolidate gaps and maintain reading frames of the exons. Gaps are shown as dashes, missing data are indicated by question marks, and periods show identity to the initial sequence.
15. D. Swofford, *PAUP: Phylogenetic Analysis using Parsimony*, v. 3.1.1. (Illinois Natural History Survey, Champaign, IL, 1993). Gaps were treated as a fifth character state. We do not consider individual exons or introns to be evolving independently and thus do not interpret parsimony trees as representing the phylogeny of the exons or introns.
16. J. Farris, *Syst. Zool.* **18**, 374 (1969).
17. J. Felsenstein, *Evolution* **39**, 783 (1985).
18. K. Bremer, *Cladistics* **10**, 295 (1994).
19. Pairwise alignments were constructed with the Clustal W option with a gap weight of five in *MacVector*, v.6.5. (Oxford Molecular Group, Oxford, 1998).

20. Supported by grants from NSF (BIR-9510799, MCB-9806999) and the Army Research Office (DAAH04-95-1-0531, DAAG55-98-1-0262). We thank M. Hinman for the *N.c.* λ library and A. de Queiroz, J. Gatesy, S. Gatesy, and anonymous reviewers for improving the manuscript.

19 October 1999; accepted 10 January 2000

Effects of Environment on Compensatory Mutations to Ameliorate Costs of Antibiotic Resistance

J. Björkman,[1,2]* I. Nagaev,[2]* O. G. Berg,[3] D. Hughes,[2] D. I. Andersson[1]†

Most types of antibiotic resistance impose a biological cost on bacterial fitness. These costs can be compensated, usually without loss of resistance, by second-site mutations during the evolution of the resistant bacteria in an experimental host or in a laboratory medium. Different fitness-compensating mutations were selected depending on whether the bacteria evolved through serial passage in mice or in a laboratory medium. This difference in mutation spectra was caused by either a growth condition–specific formation or selection of the compensated mutants. These results suggest that bacterial evolution to reduce the costs of antibiotic resistance can take different trajectories within and outside a host.

Among the major factors determining the frequency of resistance in a bacterial population are (i) the volume of antibiotic use, (ii) the costs of resistance to bacterial fitness, and (iii) the ability of bacteria to genetically compensate for such costs (*1*, *2*). Generally, both plasmid- and chromosomally conferred resistances cause fitness losses, even though exceptions are known. When resistance has a cost, compensatory mutations can ameliorate these costs, commonly without loss of resistance (*3*, *4*).

To determine whether the costs of resistance are compensated by different mutations under different growth conditions, we examined two