Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 7120–7124, September 1990
Biochemistry

Structure of a protein superfiber: Spider dragline silk

(major ampullate/fibroin/DNA sequence/*Nephila*)

MING XU AND RANDOLPH V. LEWIS*

Molecular Biology Department, University of Wyoming, P.O. Box 3944 University Station, Laramie, WY 82071-3944

Communicated by Sidney Udenfriend, July 5, 1990 (received for review May 7, 1990)

ABSTRACT **Spider major ampullate (dragline) silk is an extracellular fibrous protein with unique characteristics of strength and elasticity. The silk fiber has been proposed to consist of pseudocrystalline regions of antiparallel β-sheet interspersed with elastic amorphous segments. The repetitive sequence of a fibroin protein from major ampullate silk of the spider *Nephila clavipes* was determined from a partial cDNA clone. The repeating unit is a maximum of 34 amino acids long and is not rigidly conserved. The repeat unit is composed of three different segments: (*i*) a 6 amino acid segment that is conserved in sequence but has deletions of 3 or 6 amino acids in many of the repeats; (*ii*) a 13 amino acid segment dominated by a polyalanine sequence of 5–7 residues; (*iii*) a 15 amino acid, highly conserved segment. The latter is predominantly a Gly-Gly-Xaa repeat with Xaa being alanine, tyrosine, leucine, or glutamine. The codon usage for this DNA is highly selective, avoiding the use of cytosine or guanine in the third position. A model for the physical properties of fiber formation, strength, and elasticity, based on this repetitive protein sequence, is presented.**

One distinctive feature of spiders is the ability to produce silk threads. The silk is synthesized in specialized silk glands in the abdomen. The orb-web-spinning spiders can produce silk from six different types of glands (1). Each of the six fibers has different mechanical properties. However, they all have several features in common. They are (*i*) composed completely of protein; (*ii*) undergo a transition from a soluble to an insoluble form that is virtually irreversible; (*iii*) composed of amino acids dominated by alanine, serine, and glycine (2) but, unlike silkworm silk, have substantial quantities of other amino acids, such as glutamine, tyrosine, leucine, and valine; and (*iv*) composed of protein(s) of unknown number and molecular weight.

Due to its size and accessibility the major ampullate gland has been the focus of most studies. Synthesis of the silk protein occurs in specialized cells at the tail of the gland, and the silk protein is secreted into the lumen of the gland where it is stored in a soluble form (1). The protein then progresses down a narrow duct where the polypeptide chains presumably begin to align. By the time the protein reaches the end of the duct, it is as insoluble as the extruded silk. The fiber diameter can be varied by virtue of a valve at the end of the duct (1).

Various physical and mechanical properties of the silks have been studied over the past 40 years. These have shown major ampullate (dragline) silk to have nearly unmatched characteristics of strength and elasticity. It has a tensile strength [200,000 psi (1 psi = 6.9 kPa), ref. 3 and our unpublished data] greater than steel and of the same order of magnitude as Kevlar. These silk fibers also have an elasticity of up to 35% (4), which is thought to be based on an entropic mechanism (5). Major ampullate silk undergoes a process termed supercontraction by shrinking to as little as 60% of its original length in water but not in organic solvents (6).

Physiochemical studies have shown the silk to contain a significant proportion of β-sheet via x-ray diffraction (7) or Fourier transform infrared spectroscopy (Z. Dong and R.V.L., unpublished work). It has also been shown that the elasticity is not from changes in the β-sheet region but only in the amorphous regions (ref. 8 and our Fourier transform infrared spectroscopy data). The current model of the fibers is a series of pseudocrystalline regions composed of stacked antiparallel β-sheets derived from multiple polypeptide chains. Interspersed with these are regions of relatively undefined structure.

The silk fibers are insoluble except in very harsh chaotropic agents such as LiSCN, $LiClO_4$, or 88% (vol/vol) formic acid. Once dissolved, the protein precipitates if dialyzed or if diluted with typical buffers. The silk protein solution can be forced through a narrow-gauge needle and fiber will reform (unpublished work).

Despite this breadth of data virtually nothing is known about the silk proteins themselves. The best molecular mass published for the major ampullate silk protein is 326 kDa (9), but due to possible problems with SDS/PAGE and this type of protein, the molecular mass value may be significantly distorted. Amino acid compositions have been published on a variety of species and silks (10, 11). The recent comprehensive data from Work and Young (2) are the best data available due to possible contamination of previous silk samples with multiple types of silk. No crosslinks of any type have been detected in spider silks, although the level of detection in these studies does not preclude them at a level of 0.1% of the total amino acids in most cases. No sequence information has been published yet, to our knowledge, on spider silk.[†] However, only recently the protein sequence from a portion of the silkworm cDNA was published (12).

MATERIALS AND METHODS

Purification and Identification of Silk Proteins. The spiders (*Nephila clavipes*) were purchased from Marine Specimens (Big Pine Key, FL). The first step was to get pure silk from a single type of silk gland. By using an idea from Work (13), an apparatus was designed to draw a single silk fiber from one spinnerette of the spider. A single silk fiber was wrapped on a spool, and a variable-speed electric drill was used to pull 0.5–1 mg of pure silk from the spinnerette.

Protein Cleavage, Peptide Purification, and Sequencing. Preliminary results showed a "ragged" amino terminus when the whole silk fiber was sequenced. Thus, the silk protein was cleaved to provide peptides for partial sequencing. Peptide fragments were generated by cleavage in 6 M HCl at 155°C for 3 min. The peptide fragments were purified by HPLC using a C_{18} reverse-phase column with a pyridine/acetate buffer (pH 4.0) and 1-propanol as the organic modifier (14). The

*To whom reprint requests should be addressed.

[†] The sequence reported in this paper has been deposited in the GenBank data base (accession no. M37137).

purified peptides were sequenced by using standard procedures on an Applied Biosystems gas-phase protein sequencer (model 470A) equipped with a 120A phenylthiohydantoin analysis system.

Synthetic DNA Probe. Synthesis of the DNA probes was done by using an automated Applied Biosystems DNA synthesizer (model 380A). The product DNA was used as a hybridization probe without further purification. Because glycine and alanine are the major components of the protein, using four different codons for these amino acids at the third position increased the possibility of matching the most frequent codon on these fragments.

cDNA Library Construction from Silk Gland mRNA. To get mRNA from silk glands, the spiders were forcibly silked to stimulate mRNA synthesis (9). The major ampullate silk glands were dissected from the abdomen of the spiders and immersed in liquid nitrogen immediately. The silk glands were ground to a powder in a small amount of liquid nitrogen with a pestle and mortar. RNA was extracted by the SDS/hot phenol method (15). An oligo-dT column (two passages) was used to isolate the mRNA from total RNA (16).

Reverse transcription of the mRNA to cDNA was done by using the RiboClone cDNA synthesis system from Promega. After making cDNA, the radiolabeled cDNA was passed over a Sepharose-4B gel filtration column (1 m × 1.6 cm). cDNAs larger than 500 base pairs (bp) were collected for ligation.

pBluescript plasmid (Stratagene) was cut with *Sma* I to create a blunt end. To reduce the self-ligation rate of the plasmids, alkaline phosphatase from calf intestine (Boehringer Mannheim) was used to remove the 5′ phosphate of the plasmid (17). The calf intestine phosphate was inactivated at 75°C for 30 min, and an Elutip-d column (Schleicher & Schuell) was used to purify the vectors. Ligation of the cDNA with pBluescript was done at 4°C overnight with T4 DNA ligase (17) and then transformed into competent XL-1 blue *Escherichia coli.* The transformed cells were spread on plates and incubated overnight at 37°C.

Screening the cDNA Library. The synthetic oligodeoxynucleotide probes were labeled with ^{32}P by T4 polynucleotide kinase (17). White colonies from the plates were transferred to 96-well assay plates containing YT (yeast extract/tryptone) medium with ampicillin and grown again at 37°C overnight. Each colony was transferred to Hybond-N hybridization transfer membranes (Amersham) by using a 32-pin stamp and allowed to grow overnight on ampicillin plates. NaOH was used to lyse the bacteria, and the DNA was fixed by baking the membranes in a vacuum oven at 80°C for 2 hr (17). The membranes were screened with kinased radiolabeled DNA probes by using the method of Wood *et al.* (18) because of the likelihood of complex DNA structure. Positive colonies were confirmed by Southern blotting of the insert DNA after agarose gel electrophoresis. Over 20 positive colonies were identified.

DNA Sequencing. Three colonies were sequenced by using the method of Sanger *et al.* (19) with deoxyadenosine [γ-[^{35}S]thio]triphosphate used for labeling. As sizes of the inserts differed, ranging from 800 bp to 2.4 kilobase pairs (kbp), each clone only provided clear reading of 300–350 bp. To read the whole sequence, the Erase-a-Base kit (Promega) was used to create a series of nested deletions for sequencing. Partial sequencing results from the 2.4-kilobase (kb) DNA showed a repetitive sequence and a high (G+C) content that can cause the inserts to delete and religate. To get complete sequence information, *Hae* III was used to cut the 2.4-kb insert into smaller fragments, ranging from 150 to 900 bp. These *Hae* III fragments were separated by 1% low-melting-point agarose (Bethesda Research Laboratories) and then purified by the hot phenol method (17) and Elutip-d. The purified fragments of different sizes were subcloned (as described above) into pBluescript KS (+/−) plasmids and M13 phages mp18 and mp19 (from Stratagene) for sequencing. A schematic of the sequencing strategy is shown in Fig. 1.

Northern (RNA) Blotting and Hybridization. mRNA was purified by passing whole RNA through the oligo-dT affinity column (as above) and through a denaturing formaldehyde-agarose gel (14). The mRNA was blotted onto Zeta-Probe (Bio-Rad) membranes (17). *Hae* III-digested fragments were separated by agarose gel electrophoresis (same procedure as above). A nick-translation kit (Bethesda Research Laboratories) was used to make radiolabeled probes with the *Hae* III-generated 900-bp internal coding region fragment as template. The membrane with mRNA was hybridized at 75°C with the probes to determine the mRNA size of the silk proteins (Bio-Rad instruction book). The membranes were then washed in distilled water at 96°C.

RESULTS

Although a variety of methods were attempted to cleave the spider silk protein, the only successful one was cleavage in 6 M HCl for 3 min at 155°C. After HPLC, several peptides were isolated and sequenced. None were larger than a hexapeptide, and the largest ones had the following sequences: GQGAG, GAGQG, GYGGLG, and AAAA. Additionally, several peptides had partial sequences of these. Based on the first two peptides, the following DNA probe was synthesized: 5′-CCNCGNCCNGTYCC-3′ (N = A, T, C, or G and Y = C or T).

Silk protein mRNA is thought to be the major mRNA present in the silk glands. Therefore the cDNA library was screened by testing transformed colonies. Of over 1800 clones tested 20 positives were found. After Southern blotting of the inserts, 12 were still positive, and the largest 6 of these positive clones were selected for further study.

The sequence of the largest, a 2.4-kbp insert, is shown in Fig. 2. The beginning of the poly(A) tail (base 2337) is shown, as is the polyadenylylation signal site (base 2319). The stop codon (base 2155) occurs at the end of a stretch of 69 amino acids that are distinctively different from the repeating segments. The total amino acid composition of this clone is very close to the composition of the silk fiber itself. The peptides isolated from the silk-fiber digests are present in this clone.

The repeating protein segments are shown with maximum alignment in Fig. 3. From this figure the repeating sequence is clearly seen. Within the repeat there appear to be three segments. The first is the initial 6 amino acids. This segment is the least conserved, but, interestingly, the deletions are all

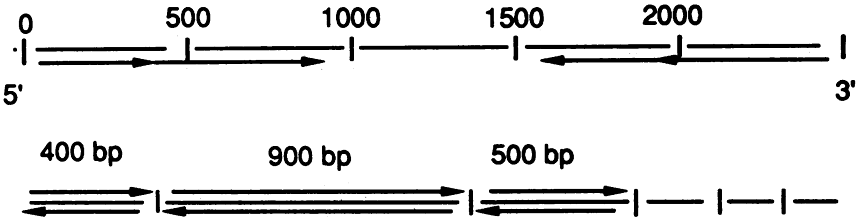

FIG. 1. Schematic of sequencing strategy. The 2.4-kbp insert is shown with arrows denoting the regions and directions sequenced; *Hae* III-digest fragments are shown below. The Erase-a-Base system was used to make nested deletions for all regions >250 bp.

```
CAA GGG GCA GGT GCA GCA GCA GCA GCA GCT GGA GGT GCC GGA CAA GGA GGA TAT GGA GGT CTT GGT GGA CAA GGA GCT GGT CAA GGT GGA   90
Gln Gly Ala Gly Ala Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Gly Gln Gly Ala Gly Gln Gly Gly   30

TAT GGA GGT CTT GGT GGA CAA GGT GCC GGA CAA GGA GCT GGT GCA GCC GCC GCA GCA GCA GCT GGT GGT GCC GGA CAA GGA GGA TAT GGA  180
Tyr Gly Gly Ley Gly Gly Gln Gly Ala Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly   60

GGT CTT GGA AGC CAA GGT GCT GGA CGA GGT GGA CAA GGA GCT GGA GCA GCC GCT GCA GCT GCG GGT GGT GCC GGA CAA GGA GGT TAT GGA  270
Gly Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly   90

GGT CTT GGA AGT CAA GGT GCA GGA CGA GGT GGA TTA GGT GGA CAA GGG GCA GGT GCA GCA GCC GCT GCA GCA GCT GGA GGT GCC GGA CAA  360
Gly Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Leu Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln  120

GGA GGA TAT GGA GGC CTT GGA AAC CAA GGT GCT GGA CGA GGT GGA CAA GGT GCA GCA GCA GCA GCA GCT GGA GGT GCT GGA CAA GGA GGA  450
Gly Gly Tyr Gly Gly Leu Gly Asn Gln Gly Ala Gly Arg Gly Gly Gln Gly Ala Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly  150

TAT GGA GGT CTT GGA AGC CAA GGT GCA GGA CGA GGT GGA TTA GGT GGA CAA GGT GCA GGT GCA GCA GCA GCA GCA GCC GGA GGT GCT GGA  540
Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Leu Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Ala Gly Gly Ala Gly  180

CAA GGC GGA TAC GGT GGT CTT GGT GGA CAA GGT GCC GGA CAA GGA GGC TAT GGA GGA CTT GGA AGC CAA GGT GCT GGA CGA GGA GGA TTA  630
Gln Gly Gly Tyr Gly Gly Leu Gly Gly Gln Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Leu  210

GGT GGA CAA GGT GCA GGT GCA GCA GCA GCA GCA GCA GCT GGA GGT GCC GGA CAA GGA GGA CTA GGT GGA CAA GGT GCT GGA CAA GGA GCT  720
Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Leu Gly Gly Gln Gly Ala Gly Gln Gly Ala  240

GGA GCA TCC GCT GCA GCA GCT GGT GGT GCC GGA CAA GGA GGA TAT GGA GGT CTT GGA AGC CAA GGT GCT GGA CGA GGT GGA GAA GGT GCA  810
Gly Ala Ser Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Glu Gly Ala  270

GGC GCA GCC GCA GCA GCA GCC GGA GGT GCT GGA CAA GGA GGA TAC GGT GGT CTT GGT GGA CAA GGT GCC GGA CAA GGA GGC TAT GGA GGA  900
Gly Ala Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Gly Gln Gln Ala Gly Gln Gly Gly Tyr Gly Gly  300

CTT GGA AGC CAA GGT GCT GGA CGA GGA GGA TTA GGT GGA CAA GGT GCA GGT GCA GCA GCA GCT GGA GGT GCC GGG CAA GGA GGA CTA GGT  990
Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Leu Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Leu Gly  330

GGA CAA GGT GCT GGA CAA GGA GCT GGA GCA GCC GCT GCA GCA GCT GGT GGT GCC GGA CAA GGA GGA TAT GGA GGT CTT GGA AGC CAA GGT 1080
Gly Gln Gly Ala Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Try Gly Gly Leu Gly Ser Gln Gly  360

GCA GGA CGA GGT GGA TTA GGT GGA CAA GGG GCA GGT GCA GTA GCC GCT GCA GCA GCT GGA GGT GCC GGA CAA GGA GGA TAT GGA GGT CTT 1170
Ala Gly Arg Gly Gly Leu Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu  390

GGA AGC CAA GGT GCT GGA CGA GGT GGA CAA GGA GCT GGA GCA GCC GCT GCA GCA GCT GGT GGT GCC GGA CAA AGA GGT TAT GGA GGT CTT 1260
Gly Ser Gln Gly Ala Gly Arg Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Arg Gly Tyr Gly Gly Leu  420

GGA AAT CAA GGT GCA GGA CGA GGT GGA TTA GGT GGA CAA GGG GCA GGT GCA GCA GCC GCT GCA GCA GCT GGA GGT GCC GGA CAA GGA GGA 1350
Gly Asn Gln Gly Ala Gly Arg Gly Gly Leu Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly  450

TAT GGA GGC CTT GGA AAC CAA GGT GCT GGA CGA GGT GGA CAA GGT GCA GCA GCA GCA GCT GGA GGT GCC GGA CAA GGA GGA TAT GGA GGT 1440
Tyr Gly Gly Leu Gly Asn Gln Gly Ala Gly Arg Gly Gly Gln Gly Ala Ala Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Tyr Gly Gly  480

CTT GGA AGC CAA GGT GCT GGA CGA GGT GGA CAA GGT GCA GGC GCA GCC GCA GCA GCA GCC GTA GGT GCT GGA CAA GAA GGA ATA CGT GGA 1530
Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Ala Val Gly Ala Gly Gln Glu Gly Ile Arg Gly  510

CAA GGT GCC GGA CAA GGA GGC TAT GGA GGA CTT GGA AGC CAA GGT TCT GGT CGA GGA GGA TTA GGT GGA CAA GGT GCA GGT GCA GCA GCA 1620
Gln Gly Ala Gly Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ser Gly Arg Gly Gly Leu Gly Gly Gln Gly Ala Gly Ala Ala Ala  540

GCA GCA GCT GGA GGT GCT GGA CAA GGA GGA TTA GGT GGA CAA GGT GCT GGA CAA GGA GCT GGA GCA GCC GCT GCA GCA GCT GGT GGT GTT 1710
Ala Ala Ala Gly Gly Ala Gly Gln Gly Gly Leu Gly Gly Gln Gly Ala Gly Gln Gly Ala Gly Ala Ala Ala Ala Ala Ala Gly Gly Val  570

AGA CAA GGA GGA TAT GGA GGT CTT GGA AGC CAA GGT GCT GGA CGA GGT GGA CAA GGT GCA GGC GCA GCC GCA GCA GCA GCC GGA GGT GCT 1800
Arg Gln Gly Gly Tyr Gly Gly Leu Gly Ser Gln Gly Ala Gly Arg Gly Gly Gln GLy Ala Gly Ala Ala Ala Ala Ala Ala Gly Gly Ala  600

GGA CAA GGA GGA TAT GGT GGT CTT GGT GGA CAA GGT GTT GGC CGA GGT GGA TTA GGT GGA CAG GGT GCA GGC GCA GCG GCA GCT GGT GGT 1890
Gly Gln Gly Gly Tyr Gly Gly Leu Gly Gly Gln Gly Val Gly Arg Gly Gly Leu Gly Gly Gln Gly Ala Gly Ala Ala Ala Ala Gly Gly  630

GCT GGA CAA GGA GGA TAT GGT GGT GTT GGT TCT GGG GCG TCT GCT GCC TCT GCA GCT GCA TCC CGG TTG TCT TCT CCT CAA GCT AGT TCA 1980
Ala Gly Gln Gly Gly Tyr Gly Gly Val Gly Ser Gly Ala Ser Ala Ala Ser Ala Ala Ala Ser Arg Leu Ser Ser Pro Gln Ala Ser Ser  660

AGA CTT TCA TCA GCT GTT TCC AAC TTG GTT GCA ACT GGT CCT ACT AAT TCT GCG GCC TTG TCA AGT ACA ATC AGT AAC GTG GTT TCA CAA 2070
Arg Val Ser Ser Ala Val Ser Asn Leu Val Ala Ser Gly Pro Thr Asn Ser Ala Ala Leu Ser Ser Thr Ile Ser Asn Val Val Ser Gln  690

ATT GGC GCC AGC ATC CTG GTC TTT CTG GAT GTG ATG TCC TCA TTC AAG CTC TTC TCG AGG TTG TTT CTG CTC TTA TCC AGA TCT TAG GTT 2160
Ile Gly Ala Ser Ile Leu Val Phe Leu Asp Val Met Ser Ser Phe Lys Leu Phe Ser Arg Leu Phe Leu Leu Leu Ser Arg Ser End

CTT CCA GCA TCG GCC AAG TTA ACT ATG GTT CCG CTG GAC AAG CCA CTC AGA TCG TTG GTC AAT CAG TTT ATC AAG CCC TAG GTT AAA TGT 2250

AAA ATC AAG AGT TGC TAA AAC TTA ATG AAC TCG GGC TGT TTA TTT GTG TTA GGT TTT AAA ATA TTT TCA ATA AAT ATT ATG CAT ATA AAA 2340

(polyA)
```

FIG. 2. Silk protein DNA and protein sequence. Sequence of a 2.4-kbp cDNA insert is shown. The sequence was read from both strands of both nested deletions and *Hae* III fragments. The stop codon is labeled End, and the polyadenylylation site is underlined.

in sets of 3; either 3 or 6 amino acids are deleted. The second segment is the middle 13 amino acids, which are highly conserved with an occasional alanine for glycine substitution and some variability in the length of the polyalanine stretch (5–7). The final segment is the last 15 amino acids. This segment is very highly conserved, and again, interestingly, any deletions are sets of 3 amino acids.

Another noteworthy feature is the codon usage (Fig. 4).

```
--------QGAGAAAAAA-GGAGQGGYGGLGGQG
---------------------AGQGGYGGLGGQG
------AGQGAGAAAAAAAGGAGQGGYGGLGSQG
AGR---GGQGAGAAAAAA-GGAGQGGYGGLGSQG
AGRGGLGGQGAGAAAAAAAGGAGQGGYGGLGNQG
AGR---GGQ--GAAAAAA-GGAGQGGYGGLGSQG
AGRGGLGGQ-AGAAAAAA-GGAGQGGYGGLGGQG
---------------------AGQGGYGGLGSQG
AGRGGLGGQGAGAAAAAAAGGAGQ---GGLGGQG
------AGQGAGASAAAA-GGAGQGGYGGLGSQG
AGR---GGEGAGAAAAAA-GGAGQGGYGGLGGQG
---------------------AGQGGYGGLGSQG
AGRGGLGGQGAGAAAA---GGAGQ---GGLGGQG
------AGQGAGAAAAAA-GGAGQGGYGGLGSQG
AGRGGLGGQGAGAVAAAAAGGAGQGGYGGLGSQG
AGR---GGQGAGAAAAAA-GGAGQRGYGGLGNQG
AGRGGLGGQGAGAAAAAAAGGAGQGGYGGLGNQG
AGR---GGQ--GAAAAA--GGAGQGGYGGLGSQG
AGR---GGQGAGAAAAAA-VGAGQEGIR---GQG
---------------------AGQGGYGGLGSQG
SGRGGLGGQGAGAAAAAA-GGAGQ---GGLGGQG
------AGQGAGAAAAAA-GGVRQGGYGGLGSQG
AGR---GGQGAGAAAAAA-GGAGQGGYGGLGGQG
VGRGGLGGQGAGAAAA---GGAGQGGYGGV-GSG
----------ASAASAAAASRLSS
```

FIG. 3. Spider silk protein repeats. The repeating protein units have been arranged from amino-terminal end to the end of the repeating segment, which is 69 amino acids from the carboxyl terminus. The units are arranged for maximum identity, and the dashes are deletions.

The bias is extremely high toward avoiding the use of cytosine and guanine in the third base position. Especially noticeable are glycine, 5.8% guanine or cytosine; glutamine, 1.5% guanine; and arginine, 4.5% guanine or cytosine.

The Northern (RNA) blot analysis showed bands at 7, 9, and 12 kb indicating possible multiple genes, alternative mRNA processing, or degradation of the mRNA. As the bands were distinct and sharp, degradation seems unlikely.

DISCUSSION

The repeating segments of spider major ampullate silk are not rigidly conserved. The repeats shown in Fig. 2, in fact, contain only one pair of identical repeats. This contrasts with the fibroin protein from silkworm, where nearly exact repeats are seen, although the published data from silkworm silk protein cDNA has many fewer repeats than presented here for the spider silk protein (12).

Based on the structure of these repeating units it is possible to propose a mechanism for fiber formation, for the pseudocrystalline β-sheet regions, and for the elasticity. The Gly-Gly-Xaa segments of the repeat form the β-sheet crystalline regions and thus provide the alignment of the different protein chains into stacked arrays to produce a fiber. Computer graphic simulations indicate a β-sheet structure is an energetically favorable conformation for these segments. Further evidence for the accuracy of this prediction are circular dichroism (CD) studies on synthetic peptides based on the complete repeating unit (P30) or the final 15 amino acid segment (P15). Both show the formation of β-sheet structure at high peptide concentrations (R.V.L., unpublished data).

The polyalanine region, although short, could easily adopt an α-helical structure. In the relaxed state there is no evidence for that in our Fourier transform infrared spectroscopy studies or the fiber x-ray diffraction studies. However, when the fiber is stretched we see clear evidence for helical regions in the Fourier transform infrared spectroscopy spectra (Z. Dong and R.V.L., unpublished work); these helical regions disappear when the fiber relaxes. Additional data from the CD studies show a clearly defined polyalanine helix in the presence of trifluoroethanol at concentrations >40% for P30, but no such helix is seen in the glycine-rich peptide P15. It is tempting to suggest that the fiber elasticity is from a reversible helix formation of the polyalanine region. Further evidence for this is that in spider silks showing little elasticity we have found no evidence for polyalanine regions, although glycine-rich-region peptides were found (unpublished data).

A repeating structure for major ampullate silk is certainly not unexpected from the previously determined amino acid composition. However, there are several features that are unexpected and possibly unique. The lack of uniformity in the repeats differs from that seen in silkworm silk fibroin protein and many other repetitive proteins. The sequence of the β-sheet regions is very dissimilar to the same regions in silkworm silk fibroin, where the sequence is based on the classical Gly-Ala-Gly-Ser type of repeat. Finally, the polyalanine region proposed to be involved in the elasticity must use a very different mechanism than elastin, where linked β-turns are the structures involved.

The elucidation of the repetitive protein structure of this spider silk protein will provide an opportunity to determine the basis for the unusual physical properties of these protein fibers.

TTT	Phe	0	TCT	Ser	4	TAT	Tyr	18	TGT	Cys	0
TTC	Phe	0	TCC	Ser	1	TAC	Tyr	2	TGC	Cys	0
TTA	Leu	9	TCA	Ser	0	TAA	End	0	TGA	End	0
TTG	Leu	0	TCG	Ser	0	TAG	End	0	TGG	Trp	0
CTT	Leu	19	CCT	Pro	0	CAT	His	0	CGT	Arg	1
CTC	Leu	0	CCC	Pro	0	CAC	His	0	CGC	Arg	0
CTA	Leu	2	CCA	Pro	0	CAA	Gln	65	CGA	Arg	15
CTG	Leu	0	CCG	Pro	0	CAG	Gln	1	CGG	Arg	0
ATT	Ile	0	ACT	Thr	0	AAT	Asn	1	AGT	Ser	1
ATC	Ile	0	ACC	Thr	0	AAC	Asn	2	AGC	Ser	10
ATA	Ile	1	ACA	Thr	0	AAA	Lys	0	AGA	Arg	2
ATG	Met	0	ACG	Thr	0	AAG	Lys	0	AGG	Arg	0
GTT	Val	3	GCT	Ala	52	GAT	Asp	0	GGT	Gly	121
GTC	Val	0	GCC	Ala	33	GAC	Asp	0	GGC	Gly	11
GTA	Val	2	GCA	Ala	94	GAA	Glu	2	GGA	Gly	169
GTG	Val	0	GCG	Ala	3	GAG	Glu	0	GGG	Gly	6

FIG. 4. Codon usage. Each codon is listed with its frequency of use. The codons are listed for the coding region only. The lack of codons with cytosine or guanine in the third position is especially notable.

We thank Purnima Ray for her assistance in the silk collection and hydrolysis study and William Kruggel for the peptide sequencing. This work was supported by funding from the Office of Naval Research.

1. Lucas, F. (1964) *Discovery* **2,** 20–26.
2. Work, R. W. & Young, C. T. (1987) *J. Arachnol.* **15,** 65–80.
3. Denny, M. W. (1980) in *Mechanical Properties of Biological Materials*, eds. Vincent, L. & Curry, M. (Cambridge Univ. Press, Cambridge), pp. 247–272.
4. Denny, M. W. (1976) *J. Exp. Biol.* **65,** 483–505.
5. Gosline, J. M. & Denny, M. W. (1984) *Nature (London)* **309,** 551–552.
6. Work, R. W. & Morosoff, N. (1982) *Text. Res. J.* **52,** 349–356.
7. Warwicker, J. O. (1960) *Faraday Soc. Trans.* **52,** 554–557.
8. Warwicker, J. O. (1960) *J. Mol. Biol.* **2,** 350–362.
9. Candela, G. C. & Cintron, J. (1981) *J. Exp. Zool.* **216,** 1–6.
10. Zemlin, J. C. (1968) *A Study of the Mechanical Behavior of Spider Silks* (U.S. Army Natick Lab., Natick, MA), Tech. Rep. 69-29-CM, AD 684333.
11. Tillinghast, E. K. (1984) *Insect Biochem.* **14,** 115–120.
12. Mita, K. (1988) *J. Mol. Biol.* **203,** 917–925.
13. Work, R. W. & Emerson, P. D. (1982) *J. Arachnol.* **10,** 1–10.
14. Ogden, R. C. & Adams, D. A. (1987) *Methods Enzymol.* **152,** 61–87.
15. Taylor, D. W., Cordingley, J. S. & Butterworth, A. E. (1984) *Mol. Biochem. Parasitol.* **10,** 305–318.
16. Aviv, H. & Leder, P. (1972) *Proc. Natl. Acad. Sci. USA* **69,** 1408–1412.
17. Maniatis, T., Fritsch, E. E. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab., Cold Spring Harbor, NY).
18. Wood, W. I., Gitschier, J., Lasky, L. A. & Lawn, R. M. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 1585–1588.
19. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.