Pergamon

Insect Biochemistry and Molecular Biology 28 (1998) 121–130

Insect Biochemistry and Molecular Biology

Evolution of repetitive proteins: spider silks from *Nephila clavipes* (Tetragnathidae) and *Araneus bicentenarius* (Araneidae)[1]

Richard Beckwitt [a,b,*], Steven Arcidiacono [b], Robert Stote [b]

[a] *Department of Biology, Framingham State College, Framingham, MA 01702, USA*
[b] *Biotechnology Division, U.S. Army, Natick Research, Development and Engineering Center, Natick, MA 01760, USA*

Received 21 May 1997; accepted 29 September 1997

Abstract

Spider silks are highly repetitive proteins, characterized by regions of polyalanine and glycine-rich repeating units. We have obtained two variants of the Spidroin 1 (NCF-1) silk gene sequence from *Nephila clavipes*. One sequence (1726 bp) was from a cloned cDNA, and the other (1951 bp) was from PCR of genomic DNA. When these sequences are compared with each other and the previously published Spidroin 1 sequence, there are differences due to sequence rearrangements, as well as single base substitutions. These variations are similar to those that have been reported from other highly repetitive genes, and probably represent the results of unequal cross-overs. We have also obtained 708 bp of sequence from PCR of genomic DNA from *Araneus bicentenarius*. This sequence shows considerable similarity to a dragline sequence (ADF-3) from *A. diadematus*, as well as Spidroin 2 (NCF-2) from *N. clavipes*. Minor but consistent differences in the repeating unit sequence between *A. bicentenarius* and *A. diadematus* suggest that concerted evolution or gene conversion processes are acting to maintain similarity among repeat units within a single gene. © 1998 Elsevier Science Ltd. All rights reserved.

Keywords: Spider silk; Spidroin; Fibroin; Repetitive protein; Concerted evolution; *Araneus bicentenarius*; *Nephila clavipes*

1. Introduction

Spider silks have been the subject of considerable recent interest, much of it directed toward their potential for bio-engineering as high performance fibers (Kaplan et al., 1994; Mello et al., 1994). In order to understand how the physical characteristics of the silk fiber are produced, it is useful to examine the range of variation that is possible among native silks. Silks are also interesting as members of a class of unusual proteins: highly repetitive in sequence, and composed of a limited range of amino acids.

Lewis and co-workers have published partial cDNA sequences of two silk genes (Spidroin 1 and 2, also called NCF-1 and NCF-2) from *Nephila clavipes* (Xu and Lewis, 1990; Hinman and Lewis, 1992). Guerette et al. (1996) have published four different partial cDNA sequences (ADF-1, ADF-2, ADF-3, ADF-4) from *Araneus diadematus*, and used Northern blots to demonstrate gland-specific synthesis. Although there are differences among these sequences, the repetitive regions of them all are characterized by regions of 4–10 alanines and glycine-rich segments. In addition, there is considerable sequence similarity in the non-repetitive C-terminal regions (Beckwitt and Arcidiacono, 1994).

The nomenclature for spider silks has not yet stabilized. The situation is likely to become more complicated, as more species are investigated. It also appears that there is not a simple correspondence between the different silk glands and the different silk genes that are expressed. On the other hand, there is also evidence that there are similarities among the silk genes present in different species, and characteristics that set this gene family apart from other silks. In this paper, we will follow the conventions of Guerette et al. (1996), although it has the limitation that related proteins from different species do not have similar names.

In this paper, we report on apparent rearrangements in allelic variants of the NCF-1 (Spidroin 1) gene of *N.*

* Corresponding author. Tel: (508)626-4791; Fax: (508)626-4794.

[1] GenBank assigned accession number(s):*Araneus*: U20328; *Nephila*: cDNA = U20329, PCR = U37520

0965-1748/98/$19.00 © 1998 Elsevier Science Ltd. All rights reserved.
PII: S0965-1748(97)00083-0

clavipes. The *N. clavipes* sequences were obtained from a cDNA clone made from major ampullate gland mRNA as well as PCR amplification of genomic DNA. The cDNA sequence has also been transferred to the pET21 expression system, and the protein product characterized (Arcidiacono et al., 1998). We also report partial sequence data from *Araneus bicentenarius* from PCR amplification of genomic DNA.

2. Materials and methods

2.1. cDNA preparation

Major ampullate glands were dissected from live spiders, and immediately frozen in liquid nitrogen. Glands from twelve spiders were ground under liquid nitrogen, with a mortar and pestle. Total RNA was extracted using the CsCl protocol (Ausubel et al., 1987). Total RNA was used to synthesize cDNA with the Librarian kit (Invitrogen).

2.2. Preparation of genomic DNA

Approximately 1 g of frozen abdominal tissue (one spider) was crushed with a mortar and pestle under liquid nitrogen, and extracted using Proteinase K and phenol/chloroform (Ausubel et al., 1987). DNA from these preparations was of high molecular weight (> 20 kbp) when examined by gel electrophoresis.

2.3. Polymerase chain reaction

The PCR primers used and their sources are:

1. (1) P-ALA, 5′-GCGGGATCCATGGCAGCAGCAGCAGCAGCT-3′ (Xu and Lewis, 1990, bases 649 > 669);
2. (2) ABR, 5′-GGGAAGCTTGTGCGGCTGGAGTAGTAGGTCCACTA-3′ (Beckwitt and Arcidiacono, 1994, bases 59 > 34);
3. (3) S1R, 5′-GGCGAATTCACCTAGGGCTTGATAAACTGATTGAC-3′ (Xu and Lewis, 1990, bases 2242 > 2218);
4. (4) S1L, 5′-CCCGGATCCGGAGGTGCCGGACAAGGAGGATATGGAGGT-3′ (Xu and Lewis, 1990, bases 31 > 60).

Slightly different PCR protocols were used with *N. clavipes* and *A. bicentenarius* genomic DNA's. For *A. bicentenarius*, PCR was carried out following the methods outlined in Saiki et al. (1988). Genomic DNA templates for each reaction were diluted to 1 μg/ml in H_2O. A negative control, consisting of all reaction components except template DNA was included in each set of reactions. After addition of the DNA template (1 μg of genomic DNA), each reaction was overlain with 1 drop of mineral oil, and denatured at 94°C for 5 minutes. *Taq* polymerase (2.5 units, Perkin-Elmer/Cetus) was then added to each reaction ("hot start"). Since the primers based on spider silk repeats have the potential to bind to many different sites within a gene, and we were unsure of the sequences to be found, we adapted the "touchdown" PCR procedure (Don et al., 1991; Roux, 1994). In this procedure, the annealing temperature is initially set quite high, and then lowered by 1–2°C after each 3 cycles. The initial annealing temperature was set at 70°C and lowered in 1°C decrements to 60°C. Once the lowest annealing temperature had been reached, the reactions were subjected to an additional 15 cycles with both annealing and extension at 72°C. The cycle times consisted of: denaturing at 94°C for 1 minute, annealing for 1 minute, and elongation at 72°C for 2 minutes. The final extension step was increased by 5 minutes to insure full-length double-stranded products, after which the reactions were held at 4°C until analyzed. Temperature cycling was performed on a Thermal Cycler (Perkin-Elmer/Cetus).

For *N. clavipes*, PCR was modified to use the Taq Extender protocol (Stratagene Cloning Systems). The reactions were run for 30 cycles of denaturing at 94°C for 1 minute and annealing and extension at 72°C for 12 minutes, followed by a final extension at 72°C for 10 minutes.

2.4. Cloning

PCR products were purified from unincorporated primer using Microcon-100 tubes (Amicon) or agarose gel electrophoresis and the Sephaglas Band-Prep kit (Pharmacia). The PCR products were cloned into pUC18 and used to transform *E. coli* strain XL-1 Blue (Stratagene) or NM522.

The cDNA was ligated into pUC18 as above. *E. coli* strain DHaF' was transformed, and the resulting library screened with synthetic oligonucleotides based on the NCF-1 (Spidroin 1) sequence (5′-TAWCCWCCYTGWCCWGCWCCWCCWGCWGC-3′) as well as a 288 bp PCR product from the non-repetitive portion of NCF-1 (Beckwitt and Arcidiacono, 1994).

2.5. DNA sequencing

The NCF-1 cDNA clone (with an insert of about 1700 base pairs) was digested with *Pst*I restriction endonuclease. The fragments were subcloned into pUC18 for sequencing. In addition, a series of nested deletions was prepared with exonuclease. A single clone, containing the *N. clavipes* PCR product of about 2000 bp, was also subcloned using *Pst*I and nested deletions. Clones containing various *A. bicentenarius* PCR products of less than 600 bp were sequenced without subcloning. Plasmid templates were prepared for sequencing using the standard mini-prep protocol (Sambrook et al., 1989), including a PEG precipitation. Sequencing was done using the A.L.F. automated DNA sequencer (Pharmacia), using the Auto-Cycle kit. The sequences were confirmed by Lofstrand Laboratories (Bethesda, MD).

2.6. Sequence analysis

Computer analysis of DNA and amino acid sequences was done using the DNASTAR package of computer programs (PC Version, DNASTAR, Inc.) and the GCG package (Genetics Computer Group, Inc.) as implemented on the VAX cluster at the NCI Frederick Biomedical Supercomputing Center.

3. Results

The cDNA clone of NCF-1 (*N. clavipes* Spidroin 1) that we obtained included 1726 bp of sequence (1637 bp of coding sequence, as well as 89 bp of the 3′ non-coding sequence, up to the poly-A tail). PCR of *N. clavipes* genomic DNA from an individual spider, using the P-ALA and S1R primers, produced a single PCR product of 1951 bp (1882 bp excluding the PCR primers). The predicted size of the PCR product, based on the sequence of Xu and Lewis (1990) was 1593 bp. The sequence included in-frame start and stop codons present in the PCR primers to allow subsequent expression of the protein product. There are no introns within this region of the NCF-1 (Spidroin 1) gene of *N. clavipes*.

When our cDNA and genomic PCR sequences are compared with each other or the NCF-1 cDNA sequence of Xu and Lewis (1990), there are some regions of exact identity, while other regions are less similar. Among the three sequences, no two are identical. In addition to silent, single-nucleotide substitutions, there are insertion/deletions of multiple codons between each pair. One possibility is that one or more sequences are in error. Since the sequences are so repetitive, and since our sequences were reconstructed from subcloned fragments, it is possible that some regions were inadvertently placed in the incorrect order. Our sequences were reconstructed from *Pst*I fragments. When each *Pst*I fragment is aligned along the Xu and Lewis sequence, they do not align in the same order and there are still regions with gaps. In addition, Xu and Lewis included a *Hae*III restriction map of their cDNA clone. When we repeated the *Hae*III digest on our clones, we obtained distinctly different fragment sizes, in each case consistent with the sequences.

The 3′ non-coding regions of the two cDNA sequences are very similar, differing only in the presence of 1-base insertion/deletions, usually in runs of repeated bases (Fig. 1). Similarities and differences in the coding regions are more apparent when the sequences are translated. Fig. 2 shows the multiple alignment of the amino acid sequences predicted from the three NCF-1 sequences. As can be seen, there are regions of strong similarity. About 140 amino acids (corresponding to the first part of our cDNA and PCR sequences) are nearly identical among all three; they differ only in the number of alanines in a stretch, or the presence of 1–3 amino acids in the glycine-rich regions. This region of similarity continues for another 180 amino acids in our cDNA and PCR sequences (up to amino acid 320 of our PCR sequence), although the Xu and Lewis sequence is more divergent. In the next 130 amino acids (up to amino acid 440 of our PCR sequence) there are also several gaps in the alignment, and the Xu and Lewis sequence and our PCR sequence appear the most similar.

It appears that entire repeats or parts of repeats have been added or lost. In particular, amino acids 362–464 of our PCR sequence are a near-perfect copy of amino acids 130–234 in the same sequence. In our cDNA sequence, this section is replaced by a single repeat (amino acids 318–337) that is unusually short, and is a copy of one that occurred earlier (222–241) The final 220 amino acids are nearly identical in all three sequences.

Individual repeats can vary within and between sequences to a considerable degree. This can be seen when all the repeat units (starting with the first G after a polyalanine region) from all three sequences are placed in a multiple sequence alignment (Fig. 3). The GCG PILEUP program begins with a pairwise alignment of the two most similar sequences, then aligns other sequences or clusters in a similar fashion. In the output, the most similar sequences are adjacent.

Several repeat units with identical amino acid sequence can be found in the three sequences, but not always in the same order (e.g. XL3, XL13, XL16, cDNA2, cDNA9, PCR1 and PCR8: each 30 amino acids long). Among these seven repeat units, only two are identical in DNA sequence (PCR8 and cDNA9), and there are up to nine silent substitutions at the nine variable sites (PCR1 and XL3, see Fig. 4).

PCR amplification of *A. bicentenarius* genomic DNA using the P-ALA and ABR primers produced a 403 bp product. It included 59 bases that were identical to our previous *A. bicentenarius* sequence (Beckwitt and Arcidiacono, 1994), including the ABR PCR primer. When the sequence was translated, it appeared much more similar to Spidroin 2 (NCF-2) than NCF-1. It is also very similar to the ADF-3 sequence found in the major ampullate gland silk from *A. diadematus* (Guerette et al., 1996). PCR amplification of *A. bicentenarius* genomic DNA using the S1L and ABR primers produced a product of 568 bp, which included the shorter PCR product within it. This was unexpected, since the S1L primer

```
                   .         .         .         .         .
cDNA 1638 ATGTAAAATCAAGAGTTGCTAAAACTTAATGAA.TCGGGCTGTTAAATTT 1686
          ||||||||||||||||||||||||||||||||| ||||||||||  ||||
XL   2248 ATGTAAAATCAAGAGTTGCTAAAACTTAATGAACTCGGGCTGTTT.ATTT 2296
                   .         .         .         .
cDNA 1687 GTGTTA.GTTTTAAAATATTTTCAATAAATATTATGCATAT 1726
          |||||| ||||||||||||||||||||||||||||||||||
XL   2297 GTGTTAGGTTTTAAAATATTTTCAATAAATATTATGCATAT 2337
```

Fig. 1. Alignment of the 3′ non-coding region of NCF-1 from Xu and Lewis (XL) with our cDNA clone (cDNA). Alignment performed with the GAP program of GCG, Gap Weight = 3.00, Length Weight = 0.100. The poly-A recognition sequence (AATAAA) is underlined, ‘.’ = gap.

was based on Spidroin 1 (NCF-1), and the original *A. bicentenarius* sequence upon which the ABR primer was based had been amplified using primers based on the NCF-1 sequence. We have combined the new sequence data with our previous *A. bicentenarius* sequence, to obtain a total of 708 bp of sequence (ABF-1), excluding the PCR primer sequences.

There are some differences in the pattern of repeats among the similar proteins from *N. clavipes*, *A. bicentenarius* and *A. diadematus*. The *N. clavipes* Spidroin 2 (NCF-2) repeats seem to alternate fairly regularly between GQQGP and GGYGP (either may appear first), and end with SGPGS. Repeat units of the Spidroin 2-like sequences in both the *A. bicentenarius* (ABF-1) and the *A. diadematus* (ADF-3) begin with GGYGPGS then have several GQQGP’s in succession, and end with YGP. There are only minor differences in the repeat units of the two *Araneus* species: in *A. diadematus*, the final GQQGP is often GGQGP instead. The result is like a theme and variations. In all three species, the GPG motif re-occurs every 5 residues and the QQ motif re-occurs every 5 or 10 residues, which may have some structural significance. The final repeat before the conserved C-terminal region is unique in each sequence, differing from other repeats as well as among species. This can be seen in the alignment of repeat units (Fig. 5). Note that all of the *Araneus* sequences are similar, and distinct from the *N. clavipes* sequences.

It is also possible to examine a longer portion of the conserved C-terminal region of ABF-1 in more detail. As can be seen in Fig. 6, the region just after the end of the repeating units is as strongly conserved in *A. bicentenarius* as it is in all of the other spider sequences.

4. Discussion

There is still some question about the formation of the dragline fiber. Hinman et al. (1994) present their current model for dragline silk, in which the proline-rich segments of the Spidroin 2 (NCF-2) repeat (GQQGP or GGYGP) are found in a series of linked β-turns, while the polyalanine regions of Spidroin 1 (NCF-1) and NCF-2 form a β-sheet. In their model, the GGX repeats of NCF-1 also adopt the linked β-turn structure. Simmons et al. (1996) present a different model for dragline silk, based on solid-state ^{2}H nuclear magnetic resonance data. In their model, there are two types of alanine-rich regions: one highly oriented and one poorly oriented. They propose that the molecule is folded into a series of alanine-rich, anti-parallel β-sheets and glycine-rich, amorphous regions. They also suggest that the chain reverses direction at GX sequences (where X is S or N). The allelic variation seen in our NCF-1 sequences, in particular the variation in the length of both the glycine-rich and alanine-rich regions, causes some alteration in the regular staggered pattern indicated by Simmons et al. (1996). Guerette et al. (1996) indicate that ADF-3 and ADF-4 are components of the *A. diadematus* dragline. Both are proline-rich, with repeats containing GPGGY (ADF-3 and ADF-4) and GQQGP (ADF-3).

There are several other repeating proteins that have amino acid sequences similar to spider silks. Elastins typically have polyalanine regions, as well as repeat units rich in glycine, proline and valine (Fazio et al., 1988). Similar sequences are also found in lamprin, an insoluble, fibrous protein found in the cartilage of lampreys (Robson et al., 1993). Robson et al. (1993) suggest that the evolutionary conservation of these repeat structures implies a conserved structural role: all are contained in proteins that are fibrillar, self-aggregating and composed largely of β-structure.

There are two evolutionary questions that arise from comparisons of allelic variants of NCF-1 and the pattern of similarities and differences among ADF-3, ABF-1 and NCF-2: (1) Within a single species, how is it possible for such large differences to arise within allelic variants of a single protein? (2) Among species, how is it possible for the pattern of small repeats to be altered uniformly throughout a larger structure of repeating units?

When examining sequences that are known to belong to a family of similar genes, one question that must be answered is whether variants represent alleles of a single locus or the products of related loci. In the case of the NCF-1 (Spidroin 1) sequences we have obtained from *N. clavipes* by cDNA cloning and genomic PCR, we are confident that they are allelic variants rather than different members of a gene family. All three of the sequences

```
 PCR  .......... .......... .......... .......... ....AAAAAA AGGAGQGGYG  16
cDNA  .......... .......... .........G YGGLGGQGAG QGAGAAAAAA AGGAGQGGYG  31
  XL  QGAGAAAAAA GGAGQGGYGG LGGQGAGQGG YGGLGGQGAG QGAGAAAAAA AGGAGQGGYG  60

 PCR  GLGSQGAGRG GQGAGAAAAA AGGAGQGGYG GLGSQGAGRG GLGGQGAGAA AAAAAGGVGQ  76
cDNA  GLGSQGAGRG GQGAGAAAAA AGGAGQGGYG GLGSQGAGRG GLGGQGAG.. AAAAAGGVGQ  89
  XL  GLGSQGAGRG GQGAGAAAAA AGGAGQGGYG GLGSQGAGRG GLGGQGAGAA AAAAAGGAGQ 120

 PCR  ...GGLGGQG AGQGA.GAAA AAAGGAGQGG YGGLGSQGAG RGGSGGQGAG AAAAAAG... 130
cDNA  ...GGLGGQG AGQGA.GAAA AAAGGAGQGG YGGLGSQGAG RGGSGGQGAG AAAAAAG... 143
  XL  GGYGGLGNQG AGRGGQGAAA AAAGGAGQGG YGGLGSQGAG RGGLGGQGAG AAAAAAGGAG 180

 PCR  .......... GAGQGGYGGL GSQGAGRGGL GGQGAGA... .......... .......... 157
cDNA  .......... GAGQGGYGGL GSQGAGRGGL GGQGAGA... .......... .......... 170
  XL  QGGYGGLGGQ GAGQGGYGGL GSQGAGRGGL GGQGAGAAAA AAAGGAGQGG LGGQGAGQGA 240

 PCR  .......... .......... .......... .AAAAAAGGA GQGGYGGLGG QGAGQGGYGG 186
cDNA  .......... .......... .......... .AAAAAAGGA GQGGYGGLGG QGAGQGGYGG 199
  XL  GASAAAAGGA GQGGYGGLGS QGAGRGGEGA GAAAAAAGGA GQGGYGGLGG QGAGQGGYGG 300

 PCR  LGSQGAGRGG LGGQGAGAAA AAAAGGAGQG GLGGQGAGQG AGAAAAAAGG AGQGGYGGLG 246
cDNA  LGSQGAGRGG LGGQGAG..A AAAAGGAGQG GLG....GQG AGAAAAAAGG AGQGGYGGLG 253
  XL  LGSQGAGRGG LGGQGAG... AAAAGGAGQG GLGGQGAGQG AGAAAAAAGG AGQGGYGGLG 357

 PCR  SQGAGR...G GQGAGA.AAA AAVGAGQGGY GG........ .......... ..QGAGQGGY 282
cDNA  SQGAGR...G GQGAGA.AAA AAGGAGQGGY GG........ .......... ..QGAGQGGY 290
  XL  SQGAGRGGLG GQGAGAVAAA AAGGAGQGGY GGLGSQGAGR GGQGAGAAAA AAGGAGQRGY 417

 PCR  GGLGSQGAGR GGLGGQGAGA AAAAAAGGAG QGGLGGQGAG QGAGAAAAAA GGAGQGGYGG 342
cDNA  GGLGSQGAGR GGLGGQGAGA AAAAAAAG.. .......... .......... .......... 318
  XL  GGLGNQGAGR GGLGGQGAGA AAAAAAGG.. .......... .......... ..AGQGGYGG 453

 PCR  LGNQGAGRGG QGAAAAAAGG AGQGGYGGLG SQGAGRGGLG GQGAGAAAAA AGGAGQGGYG 402
cDNA  .......... .......... .......... ..GAGQGGLG GQGAGAAAAA A......... 337
  XL  LGNQGAGRGG QG.AAAAAGG AGQGGYGGLG SQGAGR...G GQGAGAAAAA AVGAGQ...E 507

 PCR  GLGGQGAGQG GYGGLGSQGS GRGGLGGQGA GAAAAAAGGA GQGGLGGQGA GQGAGAAAAA 462
cDNA  .......... .......... .......... .......GGA GQGGLGGQGA GQGAGAAAAA 360
  XL  GIRGQGAGQG GYGGLGSQGS GRGGLGGQGA GAAAAAAGGA GQGGLGGQGA GQGAGAAAAA 567

 PCR  ....AGGVRQ GGYGGLGSQG AGRGGQGAGA AAAAAGGAGQ GGYGGLGGQG VGRGGLGGQG 518
cDNA  AAAAAGGVRQ GGYGGLGSQG AGRGGQGAGA AAAAAGGAGQ GGYGGLGGQG VGRGGLGGQG 420
  XL  ....AGGVRQ GGYGGLGSQG AGRGGQGAGA AAAAAGGAGQ GGYGGLGGQG VGRGGLGGQG 623

 PCR  AGAAAAGGAG QGGYGGVGSG ASAASAAASR LSSPQASSRV SSAVSNLVAS GPTNSAALSS 578
cDNA  AGAAAAVGAG QGGYGGVGSG ASAASAAASR LSSPQASSRV SSAVSNLVAS GPTNSAALSS 480
  XL  AGAAAAGGAG QGGYGGVGSG ASAASAAASR LSSPQASSRL SSAVSNLVAT GPTNSAALSS 683

 PCR  TISNVVSQIG ASNPGLSGCD VLIQALLEVV SALIQILGSS SIGQVNYGSA GQATQIVGQS VYQAL. 643
cDNA  TISNVVSQIG ASNPGLSGCD VLIQALLEVV SALIQILGSS SIGQVNYGSA GQATQIVGQS VYQALG 546
  XL  TISNVVSQIG ASNPGLSGCD VLIQALLEVV SALIQILGSS SIGQVNYGSA GQATQIVGQS VYQALG 749
```

Fig. 2. Multiple alignment of the Spidroin 1 (NCF-1) amino acid sequence from Xu and Lewis (XL) with that derived from our cDNA clone (cDNA) and genomic PCR (PCR). Alignment performed with the PILEUP program of GCG, Gap Weight = 3.00, Gap Length Weight = 0.100, standard symbol comparison table.

do have considerable regions of identity at the DNA level, including most of the final 650 bases of each. This is in contrast to the pattern shown by the four members of the spider silk gene family in *Araneus diadematus* (Guerette et al., 1996), where the differences in the repeating and non-repeating portions of the sequences are considerable.

In the case of the three variants of NCF-1, the repeat-

```
          1                                                 50
   PCR17  .......... ...GGAGQGG YGGVGS.... .......GAS AASAAA....
    XL21  .......... ...GGAGQGG YGGVGS.... .......GAS AASAAA....
  cDNA15  .......... ...VGAGQGG YGGVGS.... .......GAS AASAAA....
    PCR3  .......... ......GGVG QGGLGGQGA. .....GQGAG AAAAAA....
   cDNA4  .......... ......GGVG QGGLGGQGA. .....GQGAG AAAAAA....
   PCR10  .......... ......GGAG QGGLGGQGA. .....GQGAG AAAAAA....
   PCR14  .......... ......GGAG QGGLGGQGA. .....GQGAG AAAAAA....
    PCR7  .......... ......GGAG QGGLGGQGA. .....GQGAG AAAAAA....
  cDNA12  .......... ......GGAG QGGLGGQGA. .....GQGAG AAAAAAAAAA
    XL11  .......... ......GGAG QGGLGGQGA. .....GQGAG AAAAAA....
    XL18  .......... ......GGAG QGGLGGQGA. .....GQGAG AAAAAA....
     XL8  .......... ......GGAG QGGLGGQGA. .....GQGAG ASAAAA....
     XL2  GGAGQGGYGG LGGQGAGQGG YGGLGGQGA. .....GQGAG AAAAAAA...
   PCR15  .......... ...GGVRQGG YGGLGSQGAG R...GGQGAG AAAAAA....
  cDNA13  .......... ...GGVRQGG YGGLGSQGAG R...GGQGAG AAAAAA....
    XL19  .......... ...GGVRQGG YGGLGSQGAG R...GGQGAG AAAAAA....
   PCR11  .......... ...GGAGQGG YGGLGNQGAG R...GGQGAA AAAA......
    XL15  .......... ...GGAGQGG YGGLGNQGAG R...GGQGAA AAA.......
     XL5  .......... ...GGAGQGG YGGLGNQGAG R...GGQGAA AAAA......
    PCR1  .......... ...GGAGQGG YGGLGSQGAG R...GGQGAG AAAAAA....
    PCR8  .......... ...GGAGQGG YGGLGSQGAG R...GGQGAG AAAAAA....
   cDNA2  .......... ...GGAGQGG YGGLGSQGAG R...GGQGAG AAAAAA....
   cDNA9  .......... ...GGAGQGG YGGLGSQGAG R...GGQGAG AAAAAA....
    XL13  .......... ...GGAGQGG YGGLGSQGAG R...GGQGAG AAAAAA....
    XL16  .......... ...GGAGQGG YGGLGSQGAG R...GGQGAG AAAAAA....
     XL3  .......... ...GGAGQGG YGGLGSQGAG R...GGQGAG AAAAAA....
     XL9  .......... ...GGAGQGG YGGLGSQGAG R...GGEGAG AAAAAA....
  cDNA11  .......... .......... ......GGAG QGGLGGQGAG AAAAAA....
   cDNA8  .......... .......... ......GGAG QGGLGGQGAG AAAAAA....
    XL14  .......... ...GGAGQRG YGGLGNQGAG RGGLGGQGAG AAAAAAA...
    PCR6  GGAGQGGYGG LGGQGAGQGG YGGLGSQGAG RGGLGGQGAG AAAAAAA...
   cDNA7  GGAGQGGYGG LGGQGAGQGG YGGLGSQGAG RGGLGGQGAG AAAAA.....
    XL10  GGAGQGGYGG LGGQGAGQGG YGGLGSQGAG RGGLGGQGAG AAAA......
     XL7  GGAGQGGYGG LGGQGAGQGG YGGLGSQGAG RGGLGGQGAG AAAAAAA...
   PCR13  GGAGQGGYGG LGGQGAGQGG YGGLGSQGSG RGGLGGQGAG AAAAAA....
   PCR16  .......... ...GGAGQGG YGGLGGQGVG RGGLGGQGAG AAAA......
  cDNA14  .......... ...GGAGQGG YGGLGGQGVG RGGLGGQGAG AAAA......
    XL20  .......... ...GGAGQGG YGGLGGQGVG RGGLGGQGAG AAAA......
    PCR4  .......... ...GGAGQGG YGGLGSQGAG RGGSGGQGAG AAAAAA....
   cDNA5  .......... ...GGAGQGG YGGLGSQGAG RGGSGGQGAG AAAAAA....
   PCR12  .......... ...GGAGQGG YGGLGSQGAG RGGLGGQGAG AAAAAA....
    PCR2  .......... ...GGAGQGG YGGLGSQGAG RGGLGGQGAG AAAAAAA...
    PCR5  .......... ...GGAGQGG YGGLGSQGAG RGGLGGQGAG AAAAAAA...
   cDNA3  .......... ...GGAGQGG YGGLGSQGAG RGGLGGQGAG AAAAA.....
   cDNA6  .......... ...GGAGQGG YGGLGSQGAG RGGLGGQGAG AAAAAAA...
     XL4  .......... ...GGAGQGG YGGLGSQGAG RGGLGGQGAG AAAAAAA...
     XL6  .......... ...GGAGQGG YGGLGSQGAG RGGLGGQGAG AAAAAA....
    PCR9  ...VGAGQGG YGGQGAGQGG YGGLGSQGAG RGGLGGQGAG AAAAAAA...
  cDNA10  ...GGAGQGG YGGQGAGQGG YGGLGSQGAG RGGLGGQGAG AAAAAAAA..
    XL12  .......... ...GGAGQGG YGGLGSQGAG RGGLGGQGAG AVAAAAA...
    XL17  ...VGAGQEG IRGQGAGQGG YGGLGSQGSG RGGLGGQGAG AAAAAA....
```

Fig. 3. Multiple alignment of Spidroin 1 (NCF-1) repeats: Repeats are numbered in order from the 5′ end of the sequence (starting after a polyalanine region). Alignment performed with the PILEUP program of GCG, parameters as above. Sequences that are most similar are adjacent. The repeats in bold are identical in amino acid sequence (see text).

```
            1                                                50
Consensus   GG-GGTGCCG GACAAGGAGG ATATGGAGGT CTTGGAAGCC A-GGTGCTGG
    cDNA2   ..t....... .......... .......... .......... .a........
    cDNA9   ..t....... .......... .......... .......... .a........
    PCR1    ..a....... .......... .......... .......... .g........
    PCR8    ..t....... .......... .......... .......... .a........
    XL3     ..t....... .......... .......... .......... .a........
    XL13    ..a....... .......... .......... .......... .a........
    XL16    ..a....... .......... .......... .......... .a........

            51                                     90
Consensus   ACGAGGTGGA CAAGG-GC-G G-GCAGCCGC -GC-GC-GC-
    cDNA2   .......... .....t..a. .c........ a..c..a..t
    cDNA9   .......... .....t..a. .c........ a..a..a..c
    PCR1    .......... .....t..a. .c........ a..c..a..t
    PCR8    .......... .....t..a. .c........ a..a..a..c
    XL3     .......... .....a..t. .a........ t..a..t..g
    XL13    .......... .....a..t. .a........ t..a..a..t
    XL16    .......... .....t..a. .c........ a..a..a..c
```

Fig. 4. Multiple alignment of DNA sequences for 7 selected Spidroin 1 (NCF-1) repeats highlighted in Fig. 3. Strict consensus, '.' = as in consensus sequence, '-' = no consensus.

```
          1                                                50                     72
 10NCF-2  .......... .......... .......... GPGQQGPGGY GPGQQGP... .....S..GP GSASAAAAAA AA
 11NCF-2  .......... .......... .....GPGGY GPGQQGPGGY APGQQGP... .....S..GP GSASAAAAAA AA
  1NCF-2  .......... .......... ......PGGY GPGQQGPGGY GPGQQGP... .....S..GP GSAAAAAAAA AA
  7NCF-2  .......... .......... .......... GPGQQGPGGY GPGQQGP... .....S..GP GSAAAAAAAA A.
 12NCF-2  .......... .......... .....GPGGY GPGQQGPGGY APGQQGP... .....S..GP GSAAAAAAAA A.
  4NCF-2  ........SA ESGQQGPGGY GPGQQGPGGY GPGQQGPGGY GPGQQGP... .....S..GP GSAAAAAAAA ..
  8NCF-2  .......... .......... .....GPGGY GPGQQGPGGY GPGQQGP... .....S..GA GSAAAAAAA. ..
  5NCF-2  .......... .........S GPGQQGPGGY GPGQQGPGGY GPGQQGP... .....S..GP GSAAAAAAAA ..
  6NCF-2  .......... .........S GPGQQGPGGY GPGQQGPGGY GPGQQGL... .....S..GP GSAAAAAAA. ..
  2NCF-2  .......... .....GPGGY GPGQQGPGGY GPGQQGPGRY GPGQQGP... .....S..GP GSAAAAAA.. ..
  3NCF-2  .......... .......... GSGQQGPGGY GPRQQGPGGY GQGQQGP... .....S..GP GSAAAASAAA ..
  9NCF-2  .......... .......... GPGQQGLGGY GPGQQGPGGY GPGQQGP... .....GGYGP GSASAAAAAA ..
 13NCF-2  .......... .......... .......... .....GPGGY GPAQQGP... .....S..GP GIAASAASAG ..
 11ADF-3  GGYGPGSGQQ GPGQQGPGQQ GPGQQGPGQQ GPGQQGPGQQ GPGQQGPGQQ GPGGQGAYGP GASAAAGAA. ..
 12ADF-3  GGYGPGSGQQ GPGQQGPGQQ GPGQQGPGQQ GPGQQGPGQQ GPGQQGP... .......YGP GASAAAAAA. ..
  1ADF-3  .......... .......... ..ARAGSGQQ GPGQQGPGQQ GPGQQGP... .......YGP GASAAAAAA. ..
  2ADF-3  .......... ......GYGP GS.....GQQ GPSQQGPGQQ GPGGQGP... .......YGP GASAAAAAA. ..
  7ADF-3  .......... .....GGYGP GS.....GQQ GPGQQGPGQQ GPGGQGP... .......YGP GASAAAAAA. ..
  4ABF-1  .......... .....GGYGP GS.....GQQ GPGQQGPGQQ GPGQQGP... .......YGA GASAAAAAA. ..
 10ADF-3  .......... .....GGYGP GSGQQGPGQQ GPGQQGPGQQ GPGGQGP... .......YGP GASAAAAAA. ..
  5ADF-3  .......... .....GGYGP GS........ ..GQQGPGQQ GPGGQGP... .......YGP GASAAAAAA. ..
  9ADF-3  .......... .....GGYGP GS........ ..GQQGPGQQ GPGGQGP... .......YGP GASAAAAAA. ..
  1ABF-1  .......... .......... .......... .......... ....QGP... .......YGP GAAAAAAAA. ..
  2ABF-1  .......... .....GGYGP GS........ ..GQQGPGQQ GPGQQGP... .......YGP GAAAAAAAA. ..
  3ABF-1  .......... .....GGYGP GS........ ..GQQGPVQQ GPGQQGP... .......YGP GASAAAAAA. ..
  6ADF-3  .......... .....GGYGP GS........ ..G.QGPGQQ GPGGQGP... .......YGP GASAAAAAA. ..
  3ADF-3  .......... .....GGYGP GS........ .......GQQ GPGGQGP... .......YGP GSSAAAAAA. ..
  8ADF-3  .......... .....GGYGP G......... .YGQQGPGQQ GPGGQGP... .......YGP GASAASAAS. ..
 13ADF-3  .......... .....GGYGP GS.....GQQ GPGQQGPGQQ GPGGQGP... .......YGP GAASAAVSV. ..
  4ADF-3  .......... .....GGNGP GS........ ..GQQGAGQQ GPGQQGP... .......YGP GASAAAAAA. ..
  5ABF-1  .......... .......... .......... GGYGPGSGQQ GPGVRVA... ........AP VASAAA.... ..
 14NCF-2  .......... .......... ......PGGY GPAQQGPAGY GPGSAVAASA GAGSAG.YGP GSQASAAA.. ..
 14ADF-3  .......... .......... .......... .......... .......... .....GGYGP QSSSVPVASA VA
```

Fig. 5. Multiple alignment of amino acid repeats from Spidroin 2-like sequences (NCF-2, ABF-1 and ADF-3). PILEUP parameters as above.

```
           1                                                  50
Consensus  SRLSSP-A-S RVSSAVS-LV SSGPT--AAL S--ISN-VSQ ISASNPGLSG
    ADF-3  SRLSSPAASS RVSSAVSSLV SSGPTKHAAL SNTISSVVSQ VSASNPGLSG
    ABF-1  SRLSSSAASS RVSSAVSSLV SSGPTTPAAL SNTISSAVSQ ISASNPGLSG
    NCF-1  SRLSSPQASS RVSSAVSNLV ASGPTNSAAL SSTISNVVSQ IGASNPGLSG
    NCF-2  SRLASPDSGA RVASAVSNLV SSGPTSSAAL SSVISNAVSQ IGASNPGLSG
    ADF-2  SRLSSPSAAA RVSSAVSLVS NGGPTSPAAL SSSISNVVSQ ISASNPGLSG
    ADF-4  SVYLRLQPRL EVSSAVSSLV SSGPTNGAAV SGALNSLVSQ ISASNPGLSG
    ADF-1  NRLSSAGAAS RVSSNVAAIA SAGA...AAL PNVISNIYSG VLSS..GVSS

           51                                               98
Consensus  CDVLVQALLE VVSALV-ILG SSSIGQVNY- ---Q--Q-VG QSV-----
    ADF-3  CDVLVQALLE VVSALVSILG SSSIGQINYG ASAQYTQMVG QSVAQALA
    ABF-1  CDVLVQALLE VVSALVHILG SSSVGQINYG ASAQYAQMV. ........
    NCF-1  CDVLIQALLE VVSALIQILG SSSIGQVNYG SAGQATQIVG QSVYQALG
    NCF-2  CDVLIQALLE IVSACVTILS SSSIGQVNYG AASQFAQVVG QSVLSAF.
    ADF-2  CDILVQALLE IISALVHILG SANIGPVNSS SAGQSASIVG QSVYRALS
    ADF-4  CDALVQALLE LVSALVAILS SASIGQVNVS SVSQSTQMIS QALS....
    ADF-1  SEALIQALLE VISALIHVLG SASIGNVSSV GVNSALNAVQ NAVGAYAG
```

Fig. 6. Multiple alignment of conserved C-terminal portion of spider silks. Majority-rule consensus, '.' = gap, '-' = no consensus.

ing and non-repeating, C-terminal portions of the molecule seem to be under different constraints. In the C-terminal portion, there are a few, single-base substitutions (most of which are silent). In the repeating part, it is difficult to decide which portions of the molecule to align, and even regions that code for identical amino acids have frequent silent substitutions. It appears that there is weak control of the size of each block of repeats (as set off by stretches of polyalanine), and individual blocks can be shuffled and/or duplicated throughout the gene with little effect on the final protein. There is some indication that the last few repeats before the conserved C-terminal region are more tightly controlled. Within blocks, certain regions seem more highly conserved: blocks nearly always start with GGAGQGGY, and end with GQGAG before the polyalanine region. Given the highly repetitive nature of the NCF-1 sequence, it is likely that mis-matched recombination within the gene is a common occurrence. Mita et al. (1994) discuss the repetitive structure of *Bombyx mori* silk fibroin heavy chain. The sequence appears to be made up of three components, and is organized in a hierarchical fashion. They suggest that mis-matched recombination is responsible for the substantial size heterogeneity seen in different allelic variants. A similar suggestion was made earlier by Manning and Gage (1980).

Galli and Wieslander (1993) have described a set of allelic variants in a salivary gland protein from *Chironomus tentans* that exhibits from 12 to 22 repeats of a 477 bp unit. Paulsson et al. (1992) describe allelic variants in the silk proteins of *Chironomus tentans* in which the numbers of two distinct repeat types vary in inverse fashion, so that the total protein varies in length by only about 10%. *B. mori* fibroin alleles can vary by up to 15% (Manning and Gage, 1980). It appears that length variation may be a common feature in proteins with a repeating structure. In each of these examples, the repeats are nearly identical. The situation in *N. clavipes* may be slightly different. Because there is no complete sequence for the NCF-1 (Spidroin 1) gene, we do not know if there is any size variation in the completed protein. The repeat units are not very highly conserved, although there is some evidence that neighboring repeats (or blocks of repeats) in a tandem array are more similar than other repeat units, as would be expected in a simple model of unequal crossing-over.

The polyalanine regions that appear to punctuate the repeats may play an important role in the facility with which repeat units are rearranged. These regions appear to be built up from GCA trinucleotide repeats, although they usually end with GCT and occasionally contain GCG or GCC. Newfeld et al. (1994) have discussed the evolution of homopolymer repeats in the *mastermind* gene of *Drosophila*. Their model of drive-selection equilibrium may help to explain the variation in the number of alanines, within a limited range, although in the *mam* protein it appears that it is the DNA triplet that may be selected for, rather than amino acid. On the other hand, although elastins have many regions of polyalanine, and there may be tandem penta-peptide repeats, the larger units between regions of polyalanine are not so clearly repeated, and there is no evidence of allelic rearrangements (Raybould et al., 1994; Schlotterer and Tautz, 1994).

Mita et al. (1994) suggest that recombination in *B. mori* silk genes may be enhanced by either Chi-like sequences, or an 18 base pair boundary sequence (**CI**) that sets off the crystalline domains. Chi-like sequences appear several times in NCF-1 (not surprising since they could code for AGG), but not in NCF-2 or ABF-1. Their

CI sequence appears only once, with 70% similarity, in both NCF-1 and NCF-2, and not at all in ABF-1.

In the three Spidroin 2-like proteins (NCF-2, ADF-3 and ABF-1) the differences appear in each block of repeat. There must have been some mechanism to regularize the sequence in many blocks; they can not each be changing independently. It may be some mechanism such as mis-matched recombination or gene conversion, which insures that an advantageous pattern rapidly spreads throughout the gene. Schlotterer and Tautz (1994) discuss the model of concerted evolution, in which a variant sequence spreads among genes within a family of repeated genes (as for ribosomal DNA). It seems likely that a similar mechanism may operate among repeat units within a single repetitious gene. The data of Schlotterer and Tautz suggests that the spread is first along a single chromosome, and only more slowly to the homologous pair (or to other copies on non-homologous chromosomes). The process may have to operate at two levels, to explain the spread of penta-peptide repeats within one block, and the spread of similar blocks, punctuated by regions of polyalanine. The minor, but consistent variation in the repeat units between the two *Araneus* species (in particular the variant GGQGP) suggests that a similar mechanism may still be acting.

One question that remains unanswered by this work is the presence of a Spidroin 1 (NCF-1-type) sequence in *A. bicentenarius*. Our ABF-1 sequence was amplified with a primer based on NCF-1, but did not produce a sequence at all similar. Either such sequences are lacking from *A. bicentenarius*, the sequences differ in the region of the ABR primer, or the *A. bicentenarius* sequence is different from NCF-1 in the repeating region. Since Guerette et al. (1996) also did not find any Spidroin 1-like sequences in *A. diadematus*, it suggests that such genes may not be present within the genus *Araneus*.

Acknowledgements

We acknowledge the National Cancer Institute for allocation of computing time and staff support at the Frederick Biomedical Supercomputing Center of the Frederick Cancer Research and Development Center. Database searches were performed at the NCBI using the BLAST network service.

References

Arcidiacono, S., Mello, C., Kaplan, D., Cheley, S., Bayley, H., 1998. Purification and characterization of recombinant spidersilk expressed in *E. coli*. Applied Microbiology and Biotechnology 49, 31–38.

Ausubel, F.M., Brent, R., Kingston, R.F., Moore, D.D., Seidman, J.G., Smith, J.A., Struhl, K. (eds.), 1987. Current protocols in molecular biology. Greene Publishing Associates and Wiley-Interscience, New York.

Beckwitt, R., Arcidiacono, S., 1994. Sequence conservation in the C-terminal region of spider silk proteins (Spidroin) from *Nephila clavipes* (Tetragnathidae) and *Araneus bicentenarius* (Araneidae). J. Biol. Chem 269, 6661–6663.

Don, R.H., Cox, P.T., Wainwright, B.J., Baker, K., Mattick, J.S., 1991. "Touchdown" PCR to circumvent spurious priming during gene amplification. Nucleic Acids Res 19, 4008.

Fazio, M.J., Olsen, D.R., Kauh, E.A., Baldwin, C.T., Indik, Z., Ornstein-Goldstein, N., Yeh, H., Rosenbloom, J., Uitto, J., 1988. Cloning of full-length elastin cDNAs from a human skin fibroblast recombinant cDNA library: further elucidation of alternative splicing utilizing exon-specific oligonucleotides. J. Invest. Dermatol 91, 458–464.

Galli, J., Wieslander, L., 1993. A repetitive secretory protein gene of a novel type in *Chironomus tentans* is specifically expressed in the salivary glands and exhibits extensive length polymorphism. J. Biol. Chem 268, 11888–11893.

Guerette, P.A., Ginzinger, D.G., Weber, B.H.F., Gosline, J.M., 1996. Silk properties determined by gland-specific expression of a spider fibrion gene family. Science 272, 112–115.

Hinman, M.B., Lewis, R.V., 1992. Isolation of a clone encoding a second dragline silk fibroin. J. Biol. Chem 267, 19320–19324.

Hinman, M.B., Stauffer, S.L., Lewis, R.V., 1994. Mechanical and chemical properties of certain spider silks. In: Kaplan, D., Adams, W.W., Farmer, B., Viney, C. (eds.), Silk polymers: materials science and biotechnology. American Chemical Society Symposium Series, v. 544. American Chemical Society, Washington, DC.

Kaplan, D., Adams, W.W., Farmer, B., Viney, C., 1994. Silk: biology, structure, properties and genetics. In: Kaplan, D., Adams, W.W., Farmer, B., Viney, C. (eds.), Silk polymers: materials science and biotechnology. American Chemical Society Symposium Series, v. 544. American Chemical Society, Washington, DC.

Manning, R.F., Gage, L.P., 1980. Internal structure of the silk fibroin gene of *Bombyx mori*. II. Remarkable polymorphism of the organization of crystalline and amorphous coding sequences. J. Biol. Chem 255, 9451–9457.

Mello, C., Yeung, B., Senecal, S., Vouros, P., Kaplan, D., 1994. Analysis of *Nephila clavipes* dragline protein. In: Kaplan, D., Adams, W.W., Farmer, B., Viney, C. (eds.), Silk polymers: materials science and biotechnology. American Chemical Society Symposium Series, v. 544. American Chemical Society, Washington, DC.

Mita, K., Ichimura, S., James, T.C., 1994. Highly repetitive structure and its organization of the silk fibroin gene. J. Mol. Evol 38, 583–592.

Newfeld, S.J., Tachida, H., Yedvobnick, B., 1994. Drive-selection equilibrium: homopolymer evolution in the *Drosophila* gene *mastermind*. J. Mol. Evol 38, 637–641.

Paulsson, G., Hoog, C., Bernholm, K., Weislander, L., 1992. Balbiani ring 1 gene in *Chironomus tentans*: Sequence organization and dynamics of a coding minisatellite. J. Mol. Biol 225, 349–361.

Raybould, M.C., Birley, A.J., Hulten, M., 1994. Two new polymorphisms in the human elastin gene. Hum. Genet 93, 475–476.

Robson, P., Wright, G.M., Sitarz, E., Maiti, A., Rawar, M., Youson, J.H., Keeley, F.W., 1993. Characterization of lamprin, an unusual matrix protein from lamprey cartilage. J. Biol. Chem 268, 1440–1447.

Roux, K.H., 1994. Using mismatched primer-template pairs in touchdown PCR. BioTechniques 16, 812–814.

Saiki, R., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B., Erlich, H.A., 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239, 487–494.

Sambrook, J., Fritsch, E.F., Maniatis, T., 1989. Molecular cloning: a laboratory manual. Second edition. Cold Spring Harbor Laboratory Press.

Schlotterer, C., Tautz, D., 1994. Chromosomal homogeneity of *Drosophila* ribosomal DNA arrays suggests intrachromosomal exchanges drive concerted evolution. Curr. Biol 4, 777–783.

Simmons, A.C., Michal, C.A., Jelinski, L.W., 1996. Molecular orientation and two-component nature of the crystalline fraction of spider dragline silk. Science 271, 84–87.

Xu, M., Lewis, R.V., 1990. Structure of a protein superfiber: spider dragline silk. Proc. Natl. Acad. Sci. USA 87, 7120–7124.