Supporting Information

Title:

Analysis of the conserved N-terminal domains in major ampullate spider silk proteins

Authors and affiliations:

Dagmara Motriuk-Smith[1*], Alyson Smith[1], Cheryl Y. Hayashi[2], Randolph V. Lewis[1]

[1]University of Wyoming, Department of Molecular Biology, Laramie WY 82071

[2]University of California, Department of Biology, Riverside CA 92521

Corresponding author:

Dagmara Motriuk-Smith

University of Wyoming

College of Agriculture

Department of Molecular Biology

Dept. 3944

1000 E. University Ave

Laramie, WY 82071

e-mail: motriuk@uwyo.edu

phone: 307-766-6380

fax: 307-766-5098

Supporting Information; Figure 1A

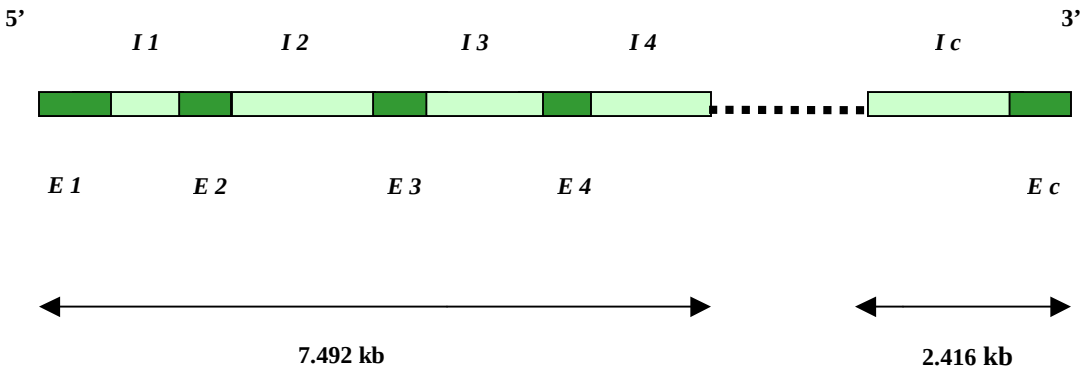
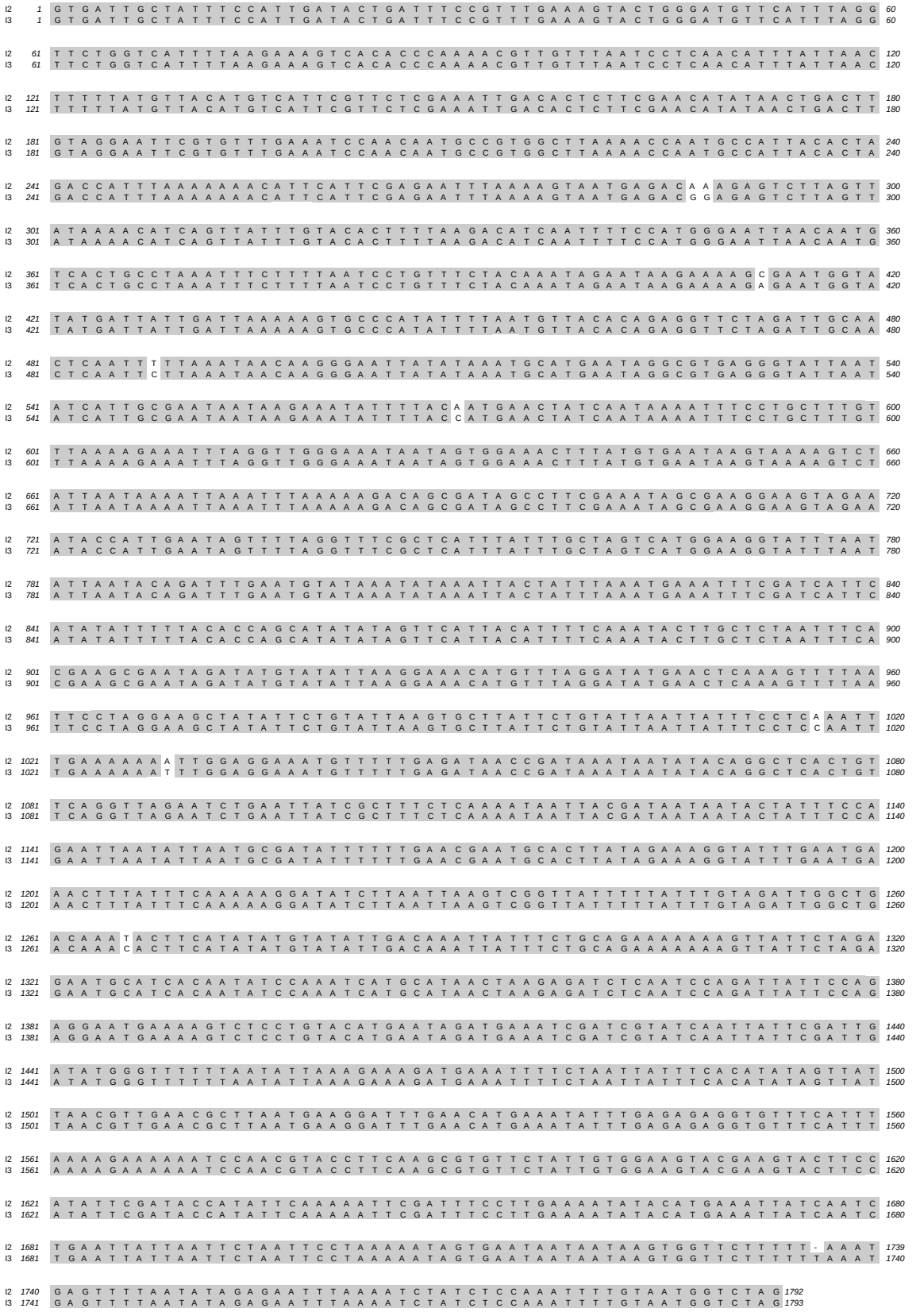


Figure1. Molecular architecture of *At.MaSp2* major ampullate silk fibroin. (A) Arrangement of exons (dark green) interrupted by introns (light green) in the *At.MaSp2* fibroin gene. (B) Alignment of intron sequences showing a high level of conservation between introns *I2* and *I3*. Abbreviations used: E, exon; I, intron. Exons 1 through 4 (*E1-E4*) represent the 5'end of silk fibroin coding region. The 3' end of the fibroin coding region is represented by exon c (*Ec*). Dots represent missing DNA sequence.

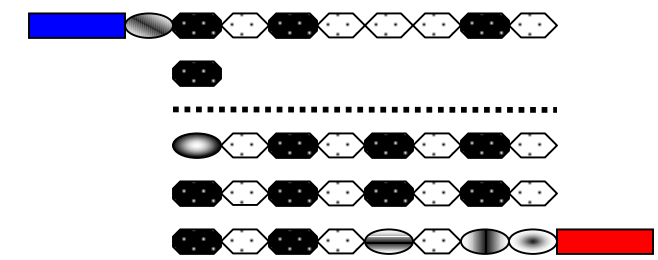
Supporting Information; Figure 1B



Supporting Information; Figure 2A

Figure 2. Arrangement of N-terminal, repetitive, and C-terminal domains within the major ampullate silk protein. (A) At.MaSp2, (B) Lg.MaSp1, and (C) Nim.MaSp2. Silk proteins are divided into non-repetitive N-terminal regions (blue), repetitive units, and C-terminal regions (red). Types of repetitive units (short blocks ending with $(A)_n$) are designated as repeat I, repeat II, and repeat III. The consensus sequence of each of these units was generated using the MacVector software (Accelerys, San Diego, CA). Silk-like sequences represent regions that did not have any significant similarity to any other sequences, thus considered as non-repetitive units. Dotted lines represent missing protein sequences.

Supporting Information; Figure 2B



N-terminus
MTAPSQSGAFSSDQIDDMSVISNTLMAAMDNMGGRITPTK
LQALDMAFASSVAEIAAVEGQNIGVTTNAISDALTSAFYQ
TTGVVNNKFISEIRSLINMFAQASANDVY

Repeat I
$PGQGGYGQGGFGQGASG(A)_7$

Repeat II
$GGAGQGGYLQRGAGQG(A)_7X$ X=A,E,L,P

Other silk-like sequences
$SSAGASSGAKGYGSVSSSVSAAQRSGMI$
$YQAPAKAQVTFSMTRQQQQGGSG(A)_6$

$GGAGQGAG(A)_5$

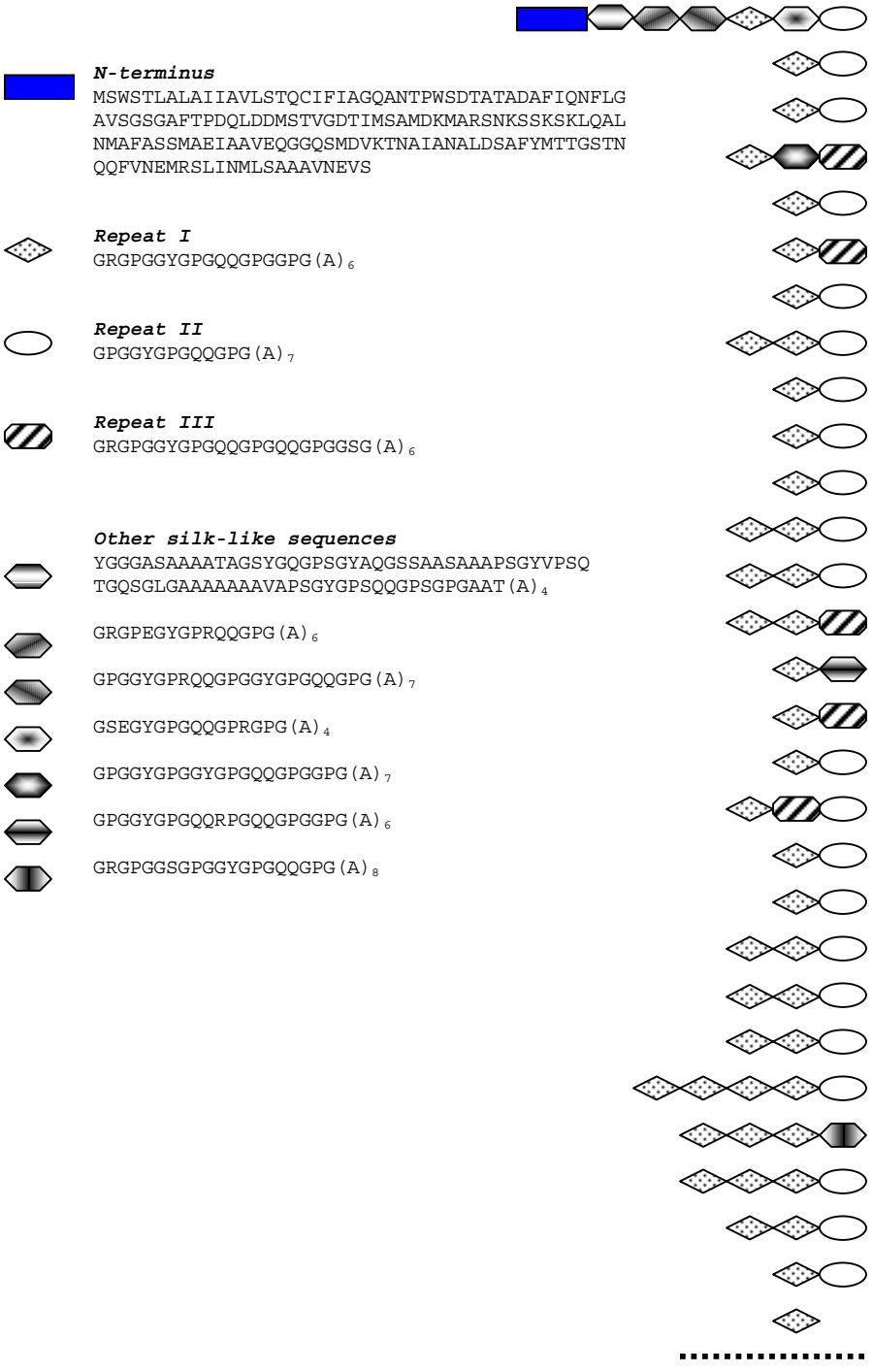
$GSGQGGQGGYGQGYRQGSVAQGG(A)_8$

$GGAGQGGAG(A)_9$

GPGQVYYGPQSVAAPAA

C-terminus
SALAAPATSARISSHALTLLSNGPTNPASISNVISNAVSQI
SSSNPGYSSCDILVQALLELVTALLTIIGSSNVNDINYGSS
GQYAQMVSQSVQNVFG

Supporting Information; Figure 2C



Supporting Information; Figure 3

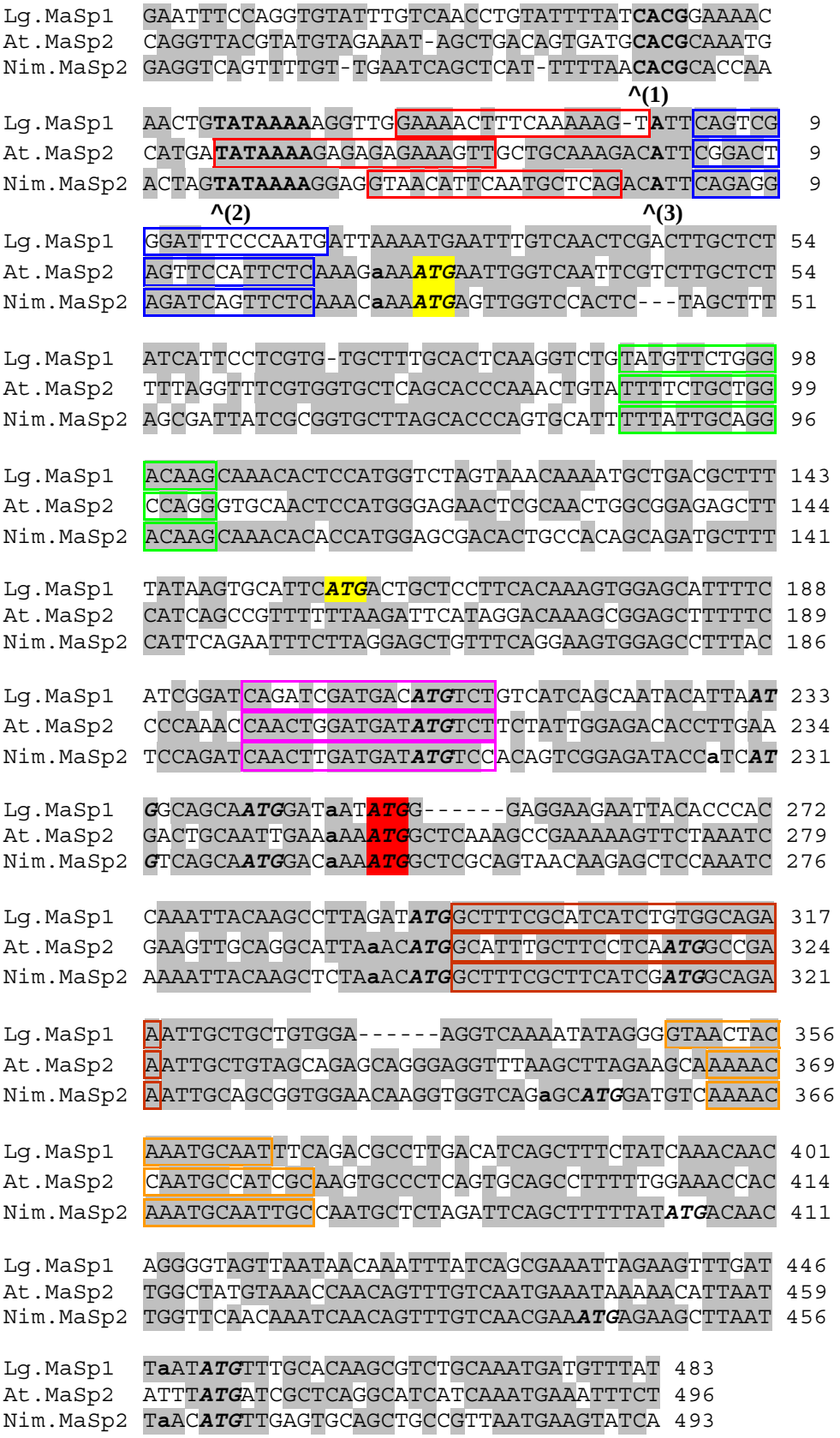


Figure 3. Alignment of genomic DNA sequences. Aligned DNA ranges from upstream of the predicted transcription start site (approximately 80 nucleotides) to the downstream of the predicted transcription start site (almost 500 nucleotides). The entire N-termini coding regions are included in this alignment. Start codons of proposed translation start sites of long isoforms are shaded in yellow. Start codons of the predicted translation start sites of short isoforms are shaded in red. The remaining start codons in frame are bold and italicized. Adenines at -3 position in respect to the ATG are in a lower case and bold. Carat symbols point at the first nucleotide of the following: (1) CACG motif; (2) TATA motif; (3) predicted transcription start site that was designated as +1. Boxed sequences correspond to regions used to design oligonucleotides that were utilized in the Northern blotting experiment. Colors of boxes represent numbers following species names (see Materials section): 1, red; 2, blue; 3, green; 4, pink; 5, brown; 6, orange.