

## **Supporting Information**

**Title:**

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

**Authors and affiliations:**

Dagmara Motriuk-Smith<sup>1\*</sup>, Alyson Smith<sup>1</sup>, Cheryl Y. Hayashi<sup>2</sup>, Randolph V. Lewis<sup>1</sup>

<sup>1</sup>University of Wyoming, Department of Molecular Biology, Laramie WY 82071

<sup>2</sup>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

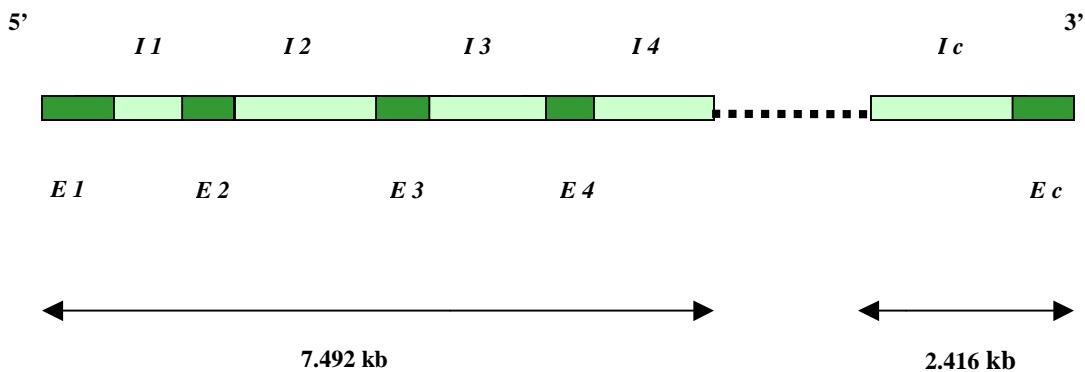


Figure1. Molecular architecture of *At.MaSp2* major ampullate silk fibroin. (A) Arrangement of exons (dark green) interrupted by introns (light green) in the *At.MaSp2* fibroin gene. (B) Alignment of intron sequences showing a high level of conservation between introns *I2* and *I3*. Abbreviations used: 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

12    1   G T G A T T G C T A T T T C C A T T G A T A C T G A T T T C C G T T T G A A A G T A C T G G G A T G T T C A T T T A G G 60  
 13    1   G T G A T T G C T A T T T C C A T T G A T A C T G A T T T C C G T T T G A A A G T A C T G G G A T G T T C A T T T A G G 60

12    61   T T C T G G T C A T T T A A G A A A G T C A C A C C C A A A C G T T G T T A A T C C T C A A C A T T T A T T A A C 120  
 13    61   T T C T G G T C A T T T A A G A A A G T C A C A C C C A A A C G T T G T T A A T C C T C A A C A T T T A T T A A C 120

12    121   T T T T T A T G T T A C A T G T C A T T C G T T C T C G A A A A T T G A C A C T C T C T C G A A C A T A T A A C T G A C T T 180  
 13    121   T T T T T A T G T T A C A T G T C A T T C G T T C T C G A A A A T T G A C A C T C T C T C G A A C A T A T A A C T G A C T T 180

12    181   G T A G G A A T T C G T G T T T G A A A T C C A A C A A T G C C G T G G C T T A A A A A C C A A T T G C C A T T A C A C T A 240  
 13    181   G T A G G A A T T C G T G T T T G A A A T C C A A C A A T G C C G T G G C T T A A A A A C C A A T T G C C A T T A C A C T A 240

12    241   G A C C A T T T A A A A A C A T T C A T T C G A G A A T T T A A A A G T A A T G A G A C A A A G A G T C T T A G T T 300  
 13    241   G A C C A T T T A A A A A C A T T C A T T C G A G A A T T T A A A A G T A A T G A G A C G G A G A G T C T T A G T T 300

12    301   A T A A A A C A T C A G T T A T T T G T A C A C T T T T A A G A C A T C A A T T T C A T G G G A A T T A A C A A T G 360  
 13    301   A T A A A A C A T C A G T T A T T T G T A C A C T T T T A A G A C A T C A A T T T C C A T G G G A A T T A A C A A T G 360

12    361   T C A C T G G C T T A A A T T T C T T T A A T C C T G T T T C T A C A A A T A G A A T A A G G A A A A G C G A A T G G T A 420  
 13    361   T C A C T G G C T T A A A T T T C T T T A A T C C T G T T T C T A C A A A T A G A A T A A G G A A A A G A G A A T G G T A 420

12    421   T A T G A T T A T T G A T T A A A A G T G C C C A T A T T T T A A T G T T A C A C A G A G G T T C T A G A T T G C A A 480  
 13    421   T A T G A T T A T T G A T t A A A A G T G C C C A T A T T T T A A T G T T A C A C A G A G G T T C T A G A T T G C A A 480

12    481   C T C A A T T T T T A A A T A A C A A G G G A A T T T A A T A A A T G C A T G A A T A G G C G T G A G G G T A T T A A T 540  
 13    481   C T C A A T T C T T A A A T A A C A A G G G A A T T T A A A T A A A T G C A T G A A T A G G C G T G A G G G T A T T A A T 540

12    541   A T C A T T G C G A A T A A T A A G A A A T T T A C A A T G G A C T A T C A A T A A A A A T T T C T G C T T T G T 600  
 13    541   A T C A T T G C G A A T A A T A A G A A A T T T A C C A T G A A C T A T C A A T A A A A A T T T C C T G C T T T G T 600

12    601   T T A A A A G A A A T T T A G G T T G G G A A A T A A T A A T A G T G G G A A A C T T T A T T G T G A A T A A G T A A A A G T C T 660  
 13    601   T T A A A A G A A A T T T A G G T T G G G A A A T A A T A G T G G G A A A C T T T A T T G T g A A t A A G T A A A A G T C T 660

12    661   A T T A A T A A A A A T T A A A T T T A A A A A G A C A G C G A T A G C C T T C G A A A A T A G C G A A G G A A G T A G A A A 720  
 13    661   A T T A A T A A A A A T T A A A T T T A A A A A G A C A G C G A T A G C C T T C G A A A A T A G C G A A G G A A G T A G A A A 720

12    721   A T A C C A T T G A A T A G T T T T A G G T T C G C T C A T T T A T T T G C T A G T C A T G G G A A G G T A T T T A A T 780  
 13    721   A T A C C A T T G A A T A G T T T T A G G T T C G C T C A T T T A T T T G C T A G T C A T G G G A A G G T A T T T A A T 780

12    781   A T T A A T A C A G A T T T G A A T G T A T A A A T A A A T T A C T A T T T A A A T G G A A A T T T C G A T C A T T C 840  
 13    781   A T T A A T A C A G A T T T G A A T G T A T A A T A A A T T A C T A T T T A A A T G G A A A T T T C G A T C A T T C 840

12    841   A T A T A T T T T T A C A C C A G C A T A T A G T T C A T T A C A T T T T C A A A A T A C T T G C T C T A A T T T C A 900  
 13    841   A T A T A T T T T T A C A C C A G C A T A T A G T T C A T T A C A T T T T C A A A A T A C T T G C T C T A A T T T C A 900

12    901   C G A A G C G A A T A G A T A T G T A T A T T A A G G G A A A C A T G T T T A G G G A T A T G A A C T C A A A G T T T T A A 960  
 13    901   C G A A G C G A A T A G A T A T G T A T A T T A A G G G A A A C A T G T T T A G G G A T A T G A A C T C C A A A G T T T T A A 960

12    961   T T C C T A G G G A A G C T A T A T T C T G T A T T A A G T G C T T A T T C T G T A T T A A T T A T T T C C T C A A A T T 1020  
 13    961   T T C C T A G G G A A G C T A T A T T C T G T A T T A A G T G C T T A T T C T G T A T T A A T T A T T T C C T C C A A A T T 1020

12    1021   T G A A A A A A A T T T G G A G G G A A T G T T T T T G A G A T A A C C G A T A A A T A A T A C A G G C T C A C T G T 1080  
 13    1021   T G A A A A A A A T T T G G A G G G A A A T G T T T T T G A G A T A A C C G A T A A A T A A T A C A G G C T C A C T G T 1080

12    1081   T C A G G T T A G A A T C T G G A A T T T A T C G C T T T C T C A A A A A T A A T T A C G G A T A A T A C T A T T T C C A 1140  
 13    1081   T C A G G T T A G A A T C T G G A A T T T A T C G C T T T C T C A A A A A T A A T T A C G G A T A A T A C T A T T T C C A 1140

12    1141   G A A T T A A T A T T A A T G C G A T A T T T T T G A A C G G A A T G C A C T T A T A G G A A A G G T A T T T G A A T G A 1200  
 13    1141   G A A T T A A T A T T A A T G C G A T A T T T T T G A A C G G A A T G C A C T T A T A G G A A A G G T A T T T G A A T G A 1200

12    1201   A A C T T T A T T T C A A A A A G G A T A T C T T A A T T A A G T C G G T T A T T T T A T T T G T A G A T T G G C T G 1260  
 13    1201   A A C T T T A T T T C A A A A A G G A T A T C T T A A T T A A G T C G G T T A T T T T A T T T G T A G A T T G G C T G 1260

12    1261   A C A A A T A C T T C A T A T A T G T A T A T T G A C A A A T T A T T C T G C A G A A A A A A A G T T A T T C T A G A 1320  
 13    1261   A C A A A C A T T C A T A T A T G T A T A T T G A C A A A T T A T T C T G C A G A A A A A A A G T T A T T C T A G A 1320

12    1321   G A A T T G C A T C A C A A T A T C C A A A T C A T G C A T A A C T A A G G A G A T C T C A A T C C A G A T T T C C A G 1380  
 13    1321   G A A T T G C A T C A C A A T A T C C A A A T C A T G C A T A A C T A A G G A G A T C T C A A T C C A G A T T T C C A G 1380

12    1381   A G G A A T G A A A A G T C T C C T G T A C A T G A A T A G A T G A A A T T C G A T C G T A T C A A T T A T T C G A T T G 1440  
 13    1381   A G G A A T G A A A A G T C T C C T G T A C A T G A A T A G A T G A A A T T C G A T C G T A T C A A T T A T T C G A T T G 1440

12    1441   A T A T G G G T T T T T A A T A T T A A A G A A A G A T G A A A T T T C T A A T T A T T T C A A T T A T T C A C A T A T A G T T A T 1500  
 13    1441   A T A T G G G T T T T T A A T A T T A A A G A A A G A T G A A A T T T C T A A T T A T T T C A A T T A T T C A C A T A T A G T T A T 1500

12    1501   T A A C G T T G A A C G C T T A A T G A A G G A T T T G G A A C A T G A A A T A T T T G A G A G A G G G T G T T C A T T T 1560  
 13    1501   T A A C G T T G A A C G C T T A A T G A A G G A T T T G G A A C A T G A A A T A T T T G A G A G A G G G T G T T C A T T T 1560

12    1561   A A A A G A A A A A T C C A A C G T A C C C T T C A A G G C G T G T T C T A T T G T G G G A A G T A C G A A G T A C T T C C 1620  
 13    1561   A A A A G A A A A A T C C A A C G T A C C C T T C A A G G C G T G T T C T A T T G T G G G A A G T A C G A A G T A C T T C C 1620

12    1621   A T A T T C G A T A C C A T A T T C A A A A A T T C G A T T T C C T T G A A A A T A A C A T G A A A T T A T C A A T C 1680  
 13    1621   A T A T T C G A T A C C A T A T T C A A A A A T T C G A T T T C C T T G A A A A T A A C A T G A A A T T A T C A A T C 1680

12    1681   T G A A T T A T T A A T T C T A A T T C C T T A A A A A T A G T G A A T A A T A A A T A A G T G G T T C T T T T - A A A T 1739  
 13    1681   T G A A T T A T T A A T T C T A A T T C C T T A A A A A T A G T G A A T A A T A A A T A A G T G G T T C T T T T T A A A T 1740

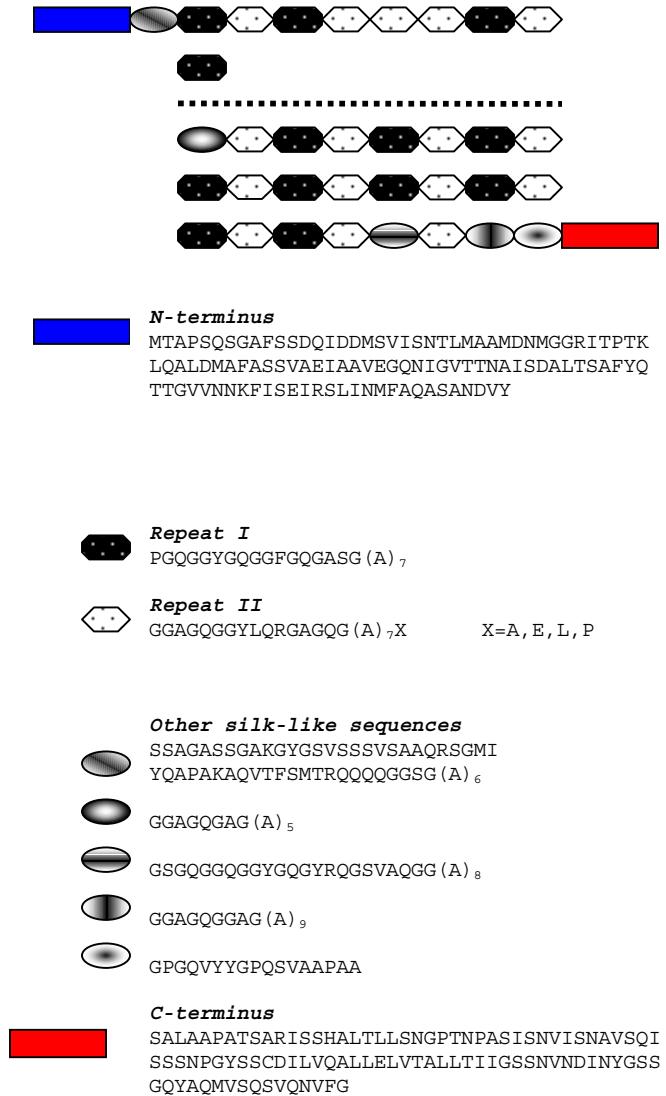
12    1740   G A G T T T T A A T A T A G A G G A A T T T A A A A A T C T A T C T C C C A A A T T T G T A A T G G T C T A G 1792  
 13    1741   G A G T T T T A A T A T A G A G G A A T T T A A A A A T C T A T C T C C C A A A T T T G T A A T G G T C T A G 1793

Supporting Information; Figure 2A

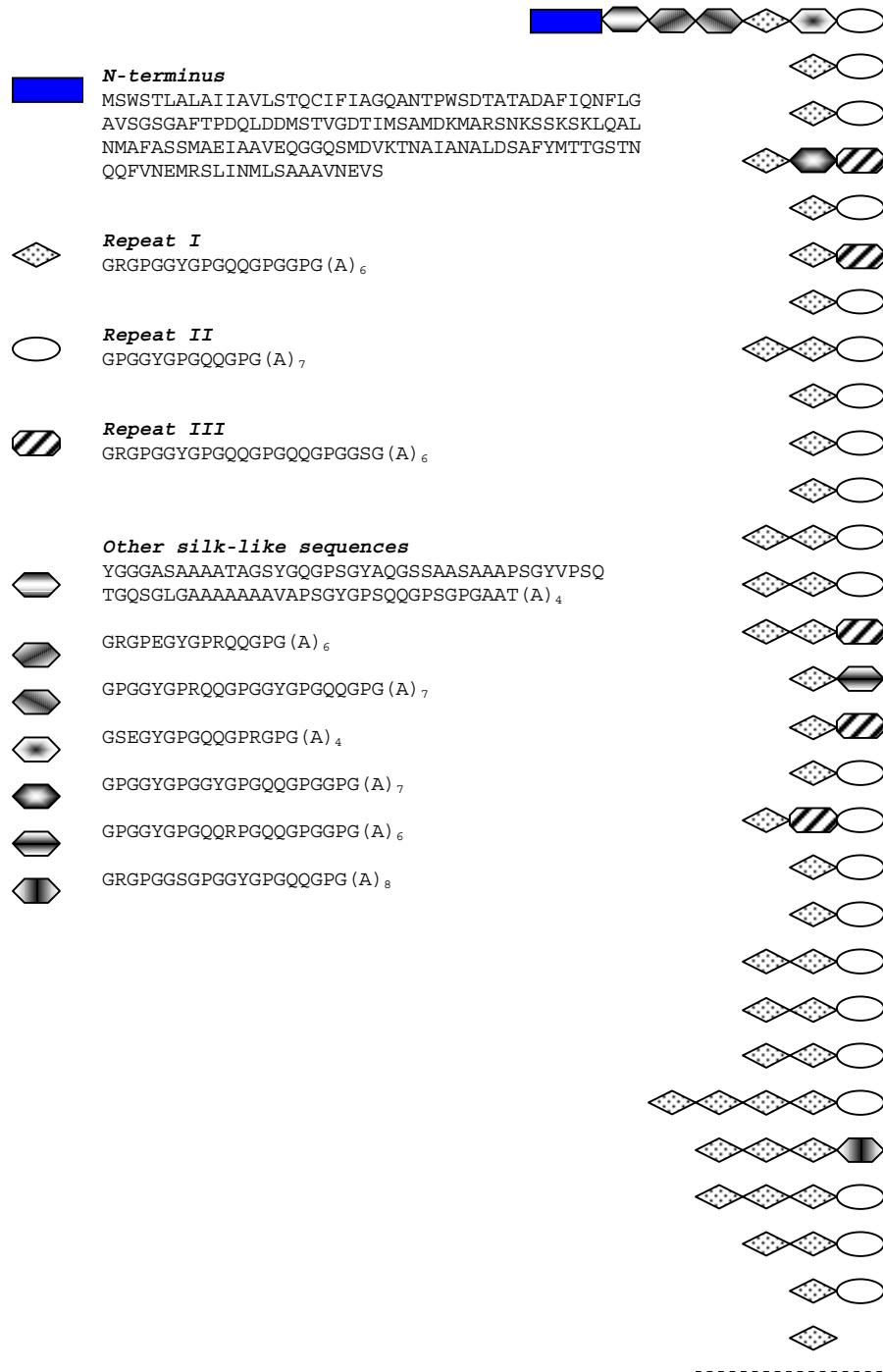


Figure 2. Arrangement of N-terminal, repetitive, and C-terminal domains within the major ampullate silk protein. (A) At.MaSp2, (B) Lg.MaSp1, and (C) Nim.MaSp2. Silk proteins are divided into non-repetitive N-terminal regions (blue), repetitive units, and C-terminal regions (red). Types of repetitive units (short blocks ending with (A)<sub>n</sub>) are designated as repeat I, repeat II, and repeat III. The consensus sequence of each of these units was generated using the MacVector software (Accelrys, 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



Supporting Information; Figure 2C



### Supporting Information; Figure 3

Lg .MaSp1	GAATTCCAGGTGTATTGTCAACCTGTATTTATCACGGAAAAAC	
At .MaSp2	CAGGTTACGTATGTAGAAAAT - AGCTGACAGTGATGCACGCAAATG	
Nim .MaSp2	GAGGTCAAGTTTGT - TGAATCAGCTCAT - TTTAACACGCACCAA	
	^ (1)	
Lg .MaSp1	AACTGTATAAAAAGGTTG GAAAACTTCAAAAAG - TATT CAGTCG	9
At .MaSp2	CATGA <b>TATAAAAAGAGAGAGAAAAGTTG</b> GCTGCAAAGACATT CGGACT	9
Nim .MaSp2	ACTAGTATAAAAAGGAG GTAACATTCAATGCTCAG ACATT CAGAGG	9
	^ (2) ^ (3)	
Lg .MaSp1	GGATTTCCAATGATTAAAATGAATTGTCAACTCGACTTGCCT 54	
At .MaSp2	AGTTCCATTCTCAAAGaAAATGAATTGGTCAATTGCTCTTGCCT 54	
Nim .MaSp2	AGATCAGTTCTCAAACaAAATGAGTTGGTCAACTC -- TAGCTTT 51	
Lg .MaSp1	ATCATTCCCTCGTG - TGCTTTGCACTCAAGGTCTG TATGTTCTGGG	98
At .MaSp2	TTTAGGTTCTGGTGGCTCAGCACCCAAACTGTATTTCTGCTGG	99
Nim .MaSp2	AGCGATTATCGCGGTGCTTAGCACCCAGTGCATT TTTATTGCAGG	96
Lg .MaSp1	ACAAG CAAACACTCCATGGCTAGTAAACAAAATGCTGACGCTTT 143	
At .MaSp2	CCAGGGTGCAACTCCATGGGAGAACTCGCAACTGGCGAGAGCTT 144	
Nim .MaSp2	ACAAG CAAACACACCATGGAGCAGACTGCCACAGCAGATGCTTT 141	
Lg .MaSp1	TATAAGTGCATTC <b>ATG</b> ACTGCTCC TTCAAAAGTGGAGCATTTC	188
At .MaSp2	CATCAGCCTTTTAAAGATTCA TAGGACAAGCGGGAGCTTTTC	189
Nim .MaSp2	CATTCAGAATTCTTAGGAGCTTTCAAGGAAGTGGAGCCTTAC	186
Lg .MaSp1	ATCGGAT CAGATCGATGAC <b>ATG</b> TCT GTCATCAGCAATACATTA <b>AT</b>	233
At .MaSp2	CCCAAAC CAACTGGATGAT <b>ATG</b> TCT TCTATTGGAGACACCTTGAA	234
Nim .MaSp2	TCCAGAT CAACTTGATGAT <b>ATG</b> TCC ACAGTCGGAGATACC <b>AT</b>	231
Lg .MaSp1	GGCAGCA <b>ATGG</b> AAT <b>ATGG</b> ----- GAGGAAGAATTACACCCAC	272
At .MaSp2	GACTGCAATTGAA <b>AAAATGG</b> CTCAAAGCCGAAAAAGTTCTAAATC	279
Nim .MaSp2	GTCAGCA <b>ATGG</b> A <b>AAAATGG</b> CT CGCAGTAACAAGAGCTCCAAATC	276
Lg .MaSp1	CAAATTACAAGCCTTAGAT <b>ATGG</b> GCTTCGCATCATCTGTGGCAGA	317
At .MaSp2	GAAGTTGCAGGCATTA <b>aACATG</b> GCATTGCTTCCTCA <b>ATGGCCGA</b>	324
Nim .MaSp2	AAAATTACAAGCTCTA <b>aACATG</b> GCTTCGCTTCATCG <b>ATGGCAGA</b>	321
Lg .MaSp1	AATTGCTGCTGTGGA ----- AGGTCAAAATATAGGG <b>GTA</b> ACTAC	356
At .MaSp2	AATTGCTGTAGCAGAGCAGGGAGGTTAAGCTTAGAAGCA <b>AAAAC</b>	369
Nim .MaSp2	AATTGCAAGCGGTGGAACAAGGTGGTCAG <b>aGCATGG</b> ATGTC <b>AAAAC</b>	366
Lg .MaSp1	AAATGCAAT <b>TTC</b> CAGACGCCCTGACATCAGCTTCTATCAAACAA	401
At .MaSp2	CAATGCCATCGC <b>AAAGT</b> GCCCTCAGTGCAGCCTTTTGAAACAC	414
Nim .MaSp2	AAATGCAATTGCCAATGCTCTAGATTCA <b>GCTT</b> AT <b>ATG</b> ACAAC	411
Lg .MaSp1	AGGGTAGTTAATAACAATTATCAGCGAAATTAGAAGTTGAT	446
At .MaSp2	TGGCTATGTAAACCAACAGTTGTCAATGAAATAAAACATTAAT	459
Nim .MaSp2	TGGTTCAACAAATCAACAGTTGTCAACGAA <b>ATG</b> AGAAGCTTAAT	456
Lg .MaSp1	<b>Ta</b> AT <b>ATG</b> TTTGCACAAGCGTCTGCAAATGATGTTAT	483
At .MaSp2	ATT <b>TATG</b> ATCGCTCAGGCATCATCAAATGAAATTCT	496
Nim .MaSp2	<b>Ta</b> AC <b>ATG</b> TTAGTGCAGCTGCCGTTAATGAAGTATCA	493

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.