Title

Molecular Evolution of Silk Genes in Mesothele and Mygalomorph Spiders, With Implications for the Early Evolution and Functional Divergence of Silk

Author
Starrett, James Richard

Publication Date
2012

UNIVERSITY OF CALIFORNIA
RIVERSIDE

Molecular Evolution of Silk Genes in Mesothele and Mygalomorph Spiders, With Implications for the Early Evolution and Functional Divergence of Silk

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Genetics, Genomics, and Bioinformatics

by

James Richard Starrett

September 2012

Dissertation Committee:
Dr. Cheryl Y. Hayashi, Chairperson
Dr. Renyi Liu
Dr. Mark Springer

The Dissertation of James Richard Starrett is approved:

____

____

____

Committee Chairperson

University of California, Riverside

Acknowledgements

The first chapter of this dissertation is a reprint of the material as it appears in *PLoS ONE* 7(6): e38084. doi:10.1371/journal.pone.0038084, published 22 June 2012. It is reproduced with permission from James Starrett, Jessica E. Garb, Amanda Kuelbs, Ugochi O. Azubuike, and Cheryl Y. Hayashi and is an open-access article distributed under the terms of the Creative Commons Attribution License. Co-authors Jessica E. Garb, Amanda Kuelbs, and Ugochi O. Azubuike provided research assistance. Co-author Cheryl Y. Hayashi supervised the research and provided lab materials.

Research was supported by National Science Foundation (NSF) Doctoral Dissertation Improvement Grant DEB-0910365 to James Starrett and Cheryl Y. Hayashi, NSF DEB-0236020 and Army Research Office W911NF0610455 grants to Cheryl Y. Hayashi, and NSF DEB-0108575 grant to Marshal Hedin and Jason Bond. Funding was also provided by a Biology Newell Travel Grant Award, UCR Distinguished Graduate Fellowship in Biology, UCR Dissertation-Year Fellowship Award, and UCR Genetics, Genomics, and Bioinformatics Graduate Fellowship.

I would like to thank Peter Arensburger, Nadia Ayoub, John Brown, Matthew Collin, Chinonyerem Oguguo, and Yonghui Zhao for providing help in the lab and with analyses. Nadia Ayoub, Matthew Collin, Jessica Garb, Marshal Hedin, Seung-Chul Kim, Dmitri Maslov, Robert Meredith, Morris Maduro, and Jordan Satler provided valuable comments on earlier versions of the dissertation chapters. John Gatesy offered comments that improved the dissertation considerably. Jason Bond, Dave Carlson, Frederick Coyle,

James Cokendolpher, John Gatesy, Helene Lee, Norman Platnick, and Casey Richart helped with specimen collecting. Marshal Hedin, Dean Leavitt, Jordan Satler, and Steven Thomas were especially helpful in the field.

I would like to thank my dissertation committee: Cheryl Hayashi, Mark Springer, and Renyi Liu for their valuable advice, support, and considerable flexibility.

I am grateful to Cheryl Hayashi and Marshal Hedin for their significant contributions to the successes I have had so far in my academic pursuits. Both have dedicated a tremendous amount of time, energy, resources, and ideas to my projects. I would not be where I am without the opportunities they have provided me, and I have always been encouraged by their consistent enthusiasm and willingness to discuss all things spider, natural history, and molecular evolution related.

Dedication

To my friends and family, for all of your support. If you ever need a kidney...

ABSTRACT OF THE DISSERTATION

Molecular Evolution of Silk Genes in Mesothele and Mygalomorph Spiders, With Implications for the Early Evolution and Functional Divergence of Silk

by

James Richard Starrett

Doctor of Philosophy, Graduate Program in Genetics, Genomics, and Bioinformatics
University of California, Riverside, September 2012
Dr. Cheryl Y. Hayashi, Chairperson

The evolution of adaptively significant gene families is an important subject in the field of evolutionary genetics. For spiders, the gene families encoding silk proteins have received considerable attention due to the high performance capabilities of the fibers they produce. However, silk gene research has largely focused on spiders of the infraorder Araneomorphae, leaving much of the phylogenetic diversity of spiders unsampled for their silk genes. Here, I sample silk genes from spiders of the suborder Mesothelae and infraorder Mygalomorphae, which are distantly related to araneomorph spiders. Phylogenetic analyses of the spidroin genes indicate that numerous duplications occurred in the spidroin gene family after opisthotheles (mygalomorphs plus araneomorphs) split from mesotheles. However, while mesotheles appear to possess a single spidroin gene, they possess numerous copies of genes homologous to Egg Case Proteins, which are currently only known from one araneomorph species. Together these results indicate that the common ancestor of extant spiders possessed a diversity of silk genes. In addition to higher level species sampling of silk genes, I sequence repetitive and carboxy terminal regions of spidroins from all species of the trapdoor spider genus, *Aliatypus*. Gene tree analyses and tests of selection suggest that contrasting evolutionary forces influenced the different regions of the spidroin gene. I also investigate the expression of silk transcripts from the tarsi and silk glands of tarantulas. Tarantulas exude silk-like secretions from their tarsi, which is hypothesized to increase surface adhesion. I discovered that while known spidroin silk genes are not expressed in the tarsi, novel silk-like genes are expressed in tarantula silk glands and tarsi. Gene families of adaptive significance may also show phylogenetic signal. Here, I sample hemocyanin gene family sequences from a phylogenetically diverse sample of spider species and infer gene trees and species trees. Phylogenetic analyses reveal that despite instances of lineage specific duplication and loss of hemocyanin paralogs, hemocyanins have phylogenetic utility for most spider groups. This dissertation shows the importance of research on gene family evolution, the roles of gene families in adaptation, and the utility of gene families in phylogenetics.

Table of Contents

Chapter 3

Chapter 4

List of Tables

List of Figures

Introduction

Evolutionary analysis of molecules of adaptive significance is essential for a thorough understanding of biological diversity. At more than 42,000 described species, Araneae (spiders) is a highly diverse taxonomic groups (Platnick, 2011), and this incredible species richness is partly attributed to their ability to produce and utilize silk. Spiders are composed of three major lineages, the Mesothelae (segmented spiders), Mygalomorphae (tarantulas, trapdoor spiders) and Araneomorphae ('true spiders'), and most members of these lineages use silk throughout their lifetime for multiple essential purposes (Coddington and Levi, 1991).

Silk fibers are composed mostly of proteins known as spidroins (Hinman and Lewis, 1992), which are encoded by members of a multigene family (Guerette et al., 1996; Gatesy et al., 2001). The spidroin gene family has been studied extensively in the araneomorph group Orbiculariae (orbweavers and cobweb spiders). Orbicularians produce up to seven different spidroins, many of which are task specific (Hu et al., 2006). Molecular evolution studies have shown that these different spidroins are related via gene duplication events (Gatesy et al., 2001), and subsequent divergence in sequence has resulted in silk fibers with different mechanical and functional properties (Hu et al., 2006). However, focus of silk gene evolution in orbicularians has limited inferences of the duplication history of the spidroin gene family, and little is known of the silk gene diversity of non-araneomorph spiders.

Mesothele and mygalomorph spiders have received little attention for their silk genes, yet characterizing silk genes from these spiders is essential for a full understanding of silk evolution. Characterization of silk genes from mesothele and mygalomorph spiders is not only important for understanding spidroin diversity and duplication history, but also for uncovering putative novel silk-associated genes. For example, studies have shown that theraphosids (tarantulas) secrete a silk-like substance from their tarsi (Gorb et al., 2006). These silk-like secretions may be encoded by the same spidroins that are expressed in the abdominal silk glands. Alternatively, genes that are not homologous to spidroins may encode tarsal silk-like secretions.

Hemocyanin proteins also represent a molecule of adaptive significance in spiders. Hemocyanins play an essential role in oxygen storage and transport (Burmester, 2002). The multi-gene family that encodes hemocyanins underwent numerous duplications prior to spider speciation (Voit et al., 2000; Averdam et al., 2003), resulting in an ancestral condition of seven hemocyanin paralogs (Markl, 1986; Markl et al., 1986). While some spider groups, such as mygalomorphs, have retained the ancestral condition of seven paralogs, some araneomorph groups have experienced lineage specific duplication and loss of hemocyanin paralogs (Ballweber et al., 2002; Averdam et al., 2003). These losses of certain paralogs and extensive duplication of others may be associated with changes in respiratory morphology and activity levels. Despite having a complex history of duplication and loss, hemocyanin sequences are highly conserved, indicating they have potential utility in spider phylogenetics. Currently, there are few available molecular markers for estimation of spider species trees. Thus, investigation of hemocyanin molecular evolution may have dual benefit: understanding respiratory evolution in spiders and inferring phylogenies.

In my dissertation, I address the evolution of the adaptively significant spidroin and hemocyanin gene families. My first chapter focuses on isolating silk genes from Mesothelae, Mygalomorphae, and a non-orbicularian araneomorph in order to assess the early evolution of silk in spiders. In the second dissertation chapter, spidroin genes are sampled from congeneric species of the trapdoor spider, *Aliatypus*, and evolutionary analyses are conducted to determine how selection has shaped different regions of spidroin genes. The focus of the third chapter of the dissertation is on the genes expressed in the abdominal silk glands and tarsi of tarantulas and determining whether tarantula tarsal silk-like secretions are encoded by spidroin genes or potentially novel silk genes. In the final chapter of the dissertation, I sample hemocyanin sequences from a broad taxonomic range of spiders and infer gene trees, species trees, and spider divergence dates to test the phylogenetic utility of a gene family with adaptive significance.

References

Averdam A, Markl J, Burmester T (2003) Subunit sequences of the 4 X 6-mer hemocyanin from the golden orb-web spider, *Nephila inaurata*. Eur. J. Biochem. 270: 3432-3439.

Ballweber P, Markl J, Burmester T (2002) Complete hemocyanin subunit sequences of the hunting spider *Cupiennius salei*. J. Biological Chemistry. 277: 14451-14457.

Burmester T (2002) Origin and evolution of arthropod hemocyanins and related proteins. J Comp Physiol B. 172: 95-107.

Coddington JA, Levi HW (1991) Systematics and Evolution of Spiders (Araneae). Annu. Rev. Ecol. Syst. 22: 565-592.

Gatesy J, Hayashi CY, Motriuk D, Woods J, Lewis RV (2001) Extreme Diversity, Conservation, and Convergence of Spider Silk Fibroin Sequences. Science. 291: 2603-2605.

Gorb SN, Niederegger S, Hayashi CY, Summers AP, Votsch W, et al. (2006) Silk-like secretion from tarantula feet. Nature. 443: 407.

Guerette PA, Ginzinger DG, Weber BH, Gosline JM (1996) Silk properties determined by gland-specific expression of a spider fibroin gene family. Science. 272: 112-115.

Hinman MB, Lewis RV (1992) Isolation of a Clone Encoding a Second Dragline Silk Fibroin. J. Biol. Chem. 267(27): 19320-19324.

Hu X, Vasanthavada K, Kohler K, McNary S, Moore AMF, et al. (2006) Molecular mechanisms of spider silk. Cell. Mol. Life Sci. 63: 1986-1999.

Markl J (1986) Evolution and function of structurally diverse subunits in the respiratory protein hemocyanin from arthropods. Biol. Bull. (Woods Hole, Mass). 171: 90-115.

Markl J, Stöcker W, Runzler R, Precht E (1986) Immunological correspondences between the hemocyanin subunits of 86 arthropods: evolution of a multigene protein family. in: Linzen, B. (Ed), Invertebrate Oxygen Carriers. Springer, Berlin Heidelberg New York, pp 281-292.

Platnick NI (2011) The world spider catalog, version 12.0. American Museum of Natural History, Available: http://research.amnh.org/iz/spiders/catalog. DOI: 10.5531/db.iz.0001. Accessed 2011, August 8.

Voit R, Feldmaier-Fuchs G, Schweikardt T, Decker H, Burmester T (2000) Complete sequence of the 24-mer hemocyanin of the tarantula *Eurypelma californicum.* The Journal of Biological Chemistry. 275: 39339-39344.

Chapter 1

Early Events in the Evolution of Spider Silk Genes

Abstract

Silk spinning is essential to spider ecology and has had a key role in the expansive diversification of spiders. Silk is composed primarily of proteins called spidroins, which are encoded by a multi-gene family. Spidroins have been studied extensively in the derived clade, Orbiculariae (orb-weavers), from the suborder Araneomorphae (‘true spiders’). Orbicularians produce a suite of different silks, and underlying this repertoire is a history of duplication and spidroin gene divergence. A second class of silk proteins, Egg Case Proteins (ECPs), is known only from the orbicularian species, *Lactrodectus hesperus* (Western black widow). In *L. hesperus,* ECPs bond with tubuliform spidroins to form egg case silk fibers. Because most of the phylogenetic diversity of spiders has not been sampled for their silk genes, there is limited understanding of spidroin gene family history and the prevalence of ECPs. Silk genes have not been reported from the suborder Mesothelae (segmented spiders), which diverged from all other spiders >380 million years ago, and sampling from Mygalomorphae (tarantulas, trapdoor spiders) and basal araneomorph lineages is sparse. In comparison to orbicularians, mesotheles and mygalomorphs have a simpler silk biology and thus are hypothesized to have less diversity of silk genes. Here, we present cDNAs synthesized from the silk glands of six mygalomorph species, a mesothele, and a non-orbicularian araneomorph, and uncover a surprisingly rich silk gene diversity. In particular, we find ECP homologs in the mesothele, suggesting that ECPs were present in the common ancestor of extant spiders, and originally were not specialized to complex with tubuliform spidroins. Furthermore, gene-tree/species-tree reconciliation analysis reveals that numerous spidroin gene duplications occurred after the split between Mesothelae and Opisthothelae (Mygalomorphae plus Araneomorphae). We use the spidroin gene tree to reconstruct the evolution of amino acid compositions of spidroins that perform different ecological functions.

Key Words: Concerted evolution, Egg Case Protein, Gene duplication, Mesothelae, Mygalomorphae, Spidroin

Introduction

Silk is vital to the ecology of spiders, being used throughout their lifetime for a wide array of essential functions. There are over 42,000 described species of spiders (Platnick, 2011), and they are not only taxonomically diverse but also ecologically diverse in their silk biology. Yet few species have been sampled for their silk genes. While most silk research has focused on derived members of Araneomorphae ("true spiders"), we present silk genes from Paleocribelletae (a basal araneomorph clade), increase sampling for Mygalomorphae (trapdoor spiders, tarantulas, and their kin; the sister group to Araneomorphae), and record silk sequences from Mesothelae (segmented spiders; the sister suborder to all other spiders; Figure 1.1; Coddington and Levi, 1991). Mesotheles and mygalomorphs exhibit profound differences in silk use compared to most araneomorph spiders (Coyle, 1986; Haupt, 2003). Mesotheles and mygalomorphs produce general-purpose fibers and apply silk in a sheet-like manner to a burrow or other substrate, which is believed to be most similar to silk use in the common ancestor of extant spiders that lived >380 million years ago (Coddington and Levi, 1991; Shear et al., 1989; Haupt and Kovoor, 1993; Foelix, 1996; Vollrath and Selden, 2007; Ayoub et al., 2007a; Ayoub and Hayashi, 2009; Blackledge et al., 2009).

Spider silk is known for its extraordinary mechanical properties, rivaling most natural and synthetic materials in strength, flexibility, and toughness (Griffiths and Salantiri, 1980; Gosline et al., 1986; Blackledge and Hayashi, 2006; Agrarsson et al., 2010). Silk is chiefly composed of proteins known as spidroins (a contraction of spider fibroins; Hinman and Lewis, 1992), which are encoded by members of a multigene family (Guerette et al., 1996; Gatesy et al., 2001; Hayashi et al., 2004; Garb and Hayashi, 2005; Garb et al., 2007; Blasingame et al., 2009; Garb et al., 2010). Studies on the spidroin gene family in orbicularian spiders show that these proteins are very long (up to 15 kb) and highly repetitive (Xu and Lewis, 1990; Hayashi and Lewis, 1998; Hayashi and Lewis, 2000; Hu et al., 2005a; Ayoub et al., 2007b). The composition of the repetitive protein-coding region is often dominated by a few amino acids - particularly alanine, glycine, and serine. The amino acid composition of the repeat regions varies considerably across different spidroin gene family members, and plays an important role in the mechanical properties of the different silk fibers (Hu et al., 2006a).

Egg Case Proteins (ECPs) comprise a second class of proteins found in spider silk but have been identified only in the egg cases of the Western black widow, *Latrodectus hesperus* (Hu et al., 2005b; Hu et al., 2006b). Unlike spidroins, ECPs are rich in cysteine. The cysteines are hypothesized to form disulfide bonds with tubuliform spidroins, the major component of *Latrodectus* egg cases (Hu et al., 2006b). Since ECPs are only known from a single species, the evolutionary history of this gene family is not clear. The phylogenetic distribution of ECPs suggests that the genes that encode ECPs are a recent evolutionary innovation restricted to black widow spiders.

Silk gland morphology and silk fiber use in mesothele, mygalomorph, and paleocribellate spiders are relatively simple in comparison to that of orbicularian (orb-web weaving) spiders (Haupt and Kovoor, 1993; Glatz, 1972; Glatz, 1973; Palmer et al., 1982; Palmer, 1985; Kovoor, 1987). Orb-weavers produce individual silk fibers that are task-specific, such as major ampullate silk, which is used in draglines and aerial orb-web frames, and tubuliform silk, which is incorporated into egg cases. Orb-weaver spiders produce up to seven silk types with unique functions that are synthesized in different morphologically distinct glands (Hu et al., 2006a). In contrast, mesotheles and mygalomorphs generally have morphologically indistinct glands that do not produce task specific fibers. Therefore, characterizing silk transcripts in mesotheles, mygalomorphs, and a basal araneomorph lineage allows for a better understanding of the evolutionary transition from substrate-borne, general-use silk fibers to aerial webs with task specific fibers spun by orb-weavers.

Despite the simplicity of their silk gland morphology and fiber types, mesothele and mygalomorph spiders rely heavily on their silk. Silk is crucial for extending the prey detection sensory area (Coyle, 1986). Additionally, these spiders are long lived and may inhabit a single burrow for their entire life (10-20 years; Ubick et al., 2005), making durable silk important for burrow maintenance. Different species of mesotheles and mygalomorphs construct a variety of web types (e.g., sheet-webs and purse-webs) and burrow entrance architectures (e.g., trip-lines, turrets, and trapdoors) indicating the potential for the discovery of silk proteins with unique mechanical properties. Further, histological studies of silk glands from representatives of basal spider lineages suggest the production of multiple protein types, raising questions regarding the silk gene diversity in these spiders (Haupt and Kovoor, 1993; Glatz, 1972; Glatz, 1973; Palmer et al., 1982; Palmer, 1985).

Spiders have received considerable attention because of the high-performance silks that they produce and the variety of ways that these silks are deployed in different ecological and behavioral contexts; yet, the understanding of the origin and early evolution of spidroins and silk remains limited (Shultz, 1987). Additionally, little is known about the diversity of silk encoding genes across spider phylogeny. Characterizing silk genes from Mesothelae, the sister group to all other extant spiders, is essential for this purpose. The few studies that have characterized mygalomorph silk genes indicate that spidroins diversified prior to the mygalomorph/araneomorph split, and mygalomorphs have the potential for producing multiple spidroins (Garb et al., 2007; Bittencourt et al., 2010; Prosdocimi et al., 2011). The recent controversy regarding silk production in the tarsi (terminal leg segments) of tarantulas also highlights the need for further investigation into the diversity of silk proteins in these spiders Gorb et al., 2006; Rind et al., 2011; Peattie et al., 2011; Pérez-Miles et al., 2009; Foelix et al., 2012).

We constructed cDNA libraries from the silk glands of spiders from Mesothelae, Mygalomorphae, and Paleocribelletae, for the purpose of characterizing the genes encoding their silk proteins. We found a considerable diversity of silk associated cDNAs in the mesothele species, *Liphistius malayanus*; in particular, we discovered homologs to ECPs that are otherwise only known from the orbicularian species, *Latrodectus hesperus*. Also, we infer from a reconciliation analysis of our spidroin gene tree that gene duplications occurred in the common ancestor of opisthotheles, after they split from mesotheles. Ancestral state reconstruction of spidroin repetitive region characteristics on the spidroin gene tree was used to infer evolutionary transitions in repeat sequence that have led to specialized and functionally diverse fibers in spiders.

Materials and Methods

Taxonomic Sampling

Our taxonomic sampling was aimed at covering phylogenetic diversity and surveying a variety of web architectures. The mesothele representative, *Liphistius malayanus*, constructs a subterranean burrow with a trapdoor and radiating sensory lines. Six species of mygalomorphs were sampled. From the Atypoidea clade, which is the sister group to remaining mygalomorphs (Ayoub et al., 2007a; Hedin and Bond, 2006), the following species were selected, with web constructs in parentheses: *Megahexura fulva* (sheet-web), *Hexura picea* (sheet-web), *Sphodros rufipes* (purse-web), and *Antrodiaetus riversi* (burrow with turret-like entrance). We sampled two non-atypoid mygalomorphs from the family Theraphosidae, *Aphonopelma seemanni*, a ground dweller (burrow/sheet-web), and *Poecilotheria regalis*, an arboreal species (sheet-web). Finally, we included the lamp-shade web spider, *Hypochilus thorelli*, which is a member of the basal araneomorph lineage, Paleocribellatae.

All specimens used in our study were obtained from pet stores or were collected on public, unprotected lands. Additionally, no species used in this study is protected or endangered. Thus, no specific permits were required for the described field studies.

cDNA Library Construction and Screening

We followed the cDNA library construction methods described in Garb et al. (2007). Briefly, each spider was anesthetized with CO_2 and then the entire set of silk glands was removed intact. The silk glands were frozen in liquid nitrogen and stored at -80º C. With the exception of the two theraphosids, glands from multiple individuals of the same species were combined to obtain sufficient tissue. Total RNA was extracted using TRIzol (Invitrogen, Carlsbad, CA) and the RNeasy Minikit (Qiagen, Valencia, CA). We isolated mRNA from total RNA using Dynal magnetic beads with oligo-(dT) anchors (Invitrogen). Double-stranded cDNA was constructed using the Superscript Choice protocol (Invitrogen), and then size selected for large fragments using Chroma Spin 1000 columns (Clontech, Mountain View, CA). The size-selected cDNA was ligated into pZErO 2.0 vectors that had been digested with EcoRV, and then transformed into TOP10 *Escherichia coli* (Invitrogen). For each species, we arrayed ~1400-1700 cDNA clones into 96-well microtiter plates. The libraries were stored at -80º C.

We screened approximately one third of each library using the method of Beuken, Vink, and Bruggeman (1998) and sequenced clones containing inserts ≥500 base pairs with T7 and Sp6 universal primers. Sequences were compared to the NCBI nr database using BLASTX (Altschul et al., 1990) to identify potential silk homologs. Libraries were also replicated onto nylon filters and probed with γ^{32} P-labeled oligonucleotides. All libraries were screened with GCDGCDGCDGCDGCDGC and CCWGCWCCWGCWCCWGCWCC, which were designed based on motifs common to spidroins (Gatesy et al., 2001; Garb and Hayashi, 2005; Garb et al., 2006). Additionally, libraries were screened with taxon specific probes designed from the end sequences of the size-selected clones. For putative *Liphistius* Egg Case Proteins (ECPs), the following probes were developed: 1) TAGTAATAAGTTCCATCGCA, 2) GCAAGGATTATAAGGATG, 3) CTTACCCTCTCCACATTCAGT, 4) GGTTTAACTTTGTTGGCGTC, 5) GGGGTCGTAAAATGATTGATA, 6) ACATTGGTTCTTTTTGTAGCA, and 7) GTTCTTGTCGTAGCATTTGTA. Probes designed from putative spidroins were 8) AAAAGCAGTGGCAGTGGCTTC, 9) CCCCTAAAATAGGTATTCTGATA (8, 9 for *Liphistius*); 10) GCCGTATGATGCTGACTGTAG, 11) TGCTGATGCGGCGGCTTG, 12) GCTTGCATAGGCTGAGGC (10-12 for *Megahexura*); 13) TATATCAGTTCCATATGGTCC, 14) GGATCGAAAACGTTGTGAAA, 15) AGCTGCTTCATTTGCTGTGTT, 16) CTTACCACAGGCGTAACC (13-16 for *Hexura*); 17) GCCGCTGCATCGGCGTAGGC, 18) AATGCAAATGCGATGGCATA, 19) CAACACACCACTCAATCCAGA (17-19 for *Sphodros*); 20) GCTCCTTCWCTMCCATATCCTCC, 21) GCTTCAGCATAYGCTTTTGC, 22) TCTRGCATAACTAGCGGCATC, 23) GTAAACTGATTCGAATTCGTC (20-23 for *Antrodiaetus*); 24) TTATCACACATCATTTTTCC (24 for *Aphonopelma*); 25) CATGGCAGAGGGTATCAGGT, 26) AGTGTAATTTGCAATGCC, 27) GCAAGAGCAATGGCGTTTCC, 28) ATAGGCATAAGCACCAGCGTT, 29) GTAAGCATAAGCCTCGGCTCC (25-29 for *Poecilotheria*); 30) AGCTCCWGCACTTGCNCCACT (30 for *Hypochilus*).

All positive clones were sequenced using T7 and Sp6 universal primers. Based on these sequences, clones that had the same translated carboxyl (C) terminal region were grouped with each other. For each group, the clone with the longest insert was selected for complete characterization. Because the inserts contained repetitive nucleotide sequence, a primer walking approach was not feasible. Instead, each selected clone was bidirectionally sequenced in its entirety using the transposon-based GPS-1 Genome Priming System (NEB, Ipswich, MA) or EZ-Tn5 Kit (Epicentre, Madison, WI).

Alignment of Egg Case Proteins

Putative *Liphistius malayanus* ECPs were aligned with *Latrodectus hesperus* ECP-1 (AY994149) and ECP-2 (DQ341220) using MUSCLE with default settings (Edgar, 2004). The alignment was imported into GeneDoc 2.7.0 (Nicholas and Nicholas, 1997) and physiochemically conserved sites were highlighted.

Phylogenetic Analyses

Phylogenetic analyses were conducted on a dataset of C-terminal encoding regions from published spidroins and those reported here. Spidroins from GenBank were selected to represent different silk glands and phylogenetic diversity. From Araneomorphae, we included *Argiope trifasciata AcSp1* (accession number AY426339), *Flag* (AF350264), and *pyriform* (GQ980328; referred to as *PySp1* in this paper); *Deinopis spinosa Flag* (DQ399325), *fibroin 1a* (DQ399326), *fibroin 1b* (DQ399327), *fibroin 2* (DQ399323), *MaSp2a* (DQ399328), *MaSp2b* (DQ399329), *MiSp1* (DQ399324), and *TuSp1* (AY953073); *Diguetia canities MaSp-like* (HM752567) and MaSp (HM752565; referred to as *MaSp-like2* in this paper); *Dolomedes tenebrosus Dtfib1* (AF350269) and *Dtfib2* (AF350270); *Latrodectus hesperus AcSp1* (EU025854), *MaSp1* (DQ409057), *MaSp2* (EF595245), *PySp1* (FJ973621), and *TuSp1* (AY953070); *Nephila clavipes Flag* (AF027973), *MaSp1* (AY654292), *MaSp2* (M92913), *MiSp1* (AF027735), *pyriform* (GQ980330; referred to here as *PySp1*), and *TuSp1* (AY855102); *Peucetia viridans MaSp1* (GU306168); *Plectreurys tristis fibroin 1* (AF350281), *fibroin 2* (AF350282), *fibroin 3* (AF350283), and *fibroin 4* (AF350284); and *Uloborus diversus AcSp1* (DQ399333), *MaSp1* (DQ399331), *MaSp2* (DQ399334), *MiSp* (DQ399332), and *TuSp1* (AY953072). From Mygalomorphae, we included *Aliatypus gulosus fibroin 1* (EU117159); *Aptostichus* sp. *fibroin 1* (EU117160) and *fibroin 2* (EU117161); *Avicularia juruensis spidroin 1a* (EU652181; referred to as *fib1a* in this study), *1b* (EU652182; referred to as *fib1b* in this study), and *1c* (EU652183; referred to as *fib1c* in this study); *Bothriocyrtum californicum fibroin 1* (EU117162), *fibroin 2* (EU117163), and *fibroin 3* (EU117164); and *Euagrus chisoseus fibroin 1* (AF350271). The C-terminal regions were aligned using MUSCLE under default parameters (Edgar, 2004) followed by manual adjustment. C-terminal encoding DNA sequences were aligned according to the amino acid alignment with PAL2NAL (Suyama et al., 2006).

We did not include *Avicularia juruensis spidroin 2* (EU652184) in our final analyses, as it is potentially an experimental artifact. A BLASTN search of the *Avicularia spidroin 2* C-terminal region resulted in only two hits. These hits (accessions AF350267, AY365020) were *MaSp2* sequences from two species of the orbicularian, *Argiope,* and had extremely small, highly significant E values (<1e-63). Phylogenetic analysis grouped this sequence with araneoid major ampullate sequences (Bittencourt et al., 2010). This result could not be corroborated as close relatives of *Avicularia spidroin 2* were not found in any of 10 mygalomorph cDNA libraries (Gatesy et al., 2001; Garb et al., 2007; this study); nor did Prosdocimi et al. (2011) recover major ampullate-like spidroins from the silk gland transcriptome of the mygalomorph, *Actinopus sp*.

We conducted phylogenetic analyses using ML and Bayesian methods. Analyses were conducted on DNA data with gaps coded using the ‘Simple’ method following Simmons and Ochoterena (Simmons and Ochoterena, 2000). Analyses were conducted through the CIPRES web server (Miller et al., 2010). Likelihood searches for the best tree and bootstrap were performed simultaneously with 1000 replicates using RAxML v. 7.2.8 (Stamatakis, 2006a; Stamatakis, 2006b; Stamatakis et al., 2008). Analyses were performed with the data partitioned by codon position, using the GTR+γ model for each partition, following RAxML program author recommendations. Coded gaps were treated as binary data and as a separate data partition.

Bayesian analyses were conducted using MrBayes v. 3.1.2 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). DNA substitution models were determined for each codon position (position 1: HKY+I+γ, position 2: GTR+I+ γ, position 3: GTR+ γ) using MrModeltest v. 2.3 (Nylander, 2004). The restriction site (binary) model with variable ascertainment bias was used for the coded gap characters (Ronquist et al., 2005). Two simultaneous searches were run for at least 10 million generations, with trees and parameters sampled from four MCMC chains every 1000th generation. Partitions (codon positions and binary characters) were unlinked and substitution rates of evolution among partitions were allowed to vary. Analyses were considered complete when the standard deviation of split frequencies between the two searches was below 0.01 (Ronquist et al., 2005). The first forty percent of samples were treated as burnin and discarded. Bayesian posterior probabilities (PP) were used to assess clade support.

Likelihood and Bayesian analyses were also conducted with constraints placed for each gland-associated spidroin group (i.e., minor ampullate, major ampullate, flagelliform, tubuliform, pyriform, and aciniform gland types; Table 1.1). Our higher-level sampling was not intended to establish monophyly of each of the gland associated spidroin groups; rather we aimed to determine the phylogenetic placements of the gland associated spidroin groups among spidroins from across the spider phylogeny. For minor ampullate, flagelliform, tubuliform, pyriform, and aciniform glands, spidroins have been reported from only a few species, while major ampullate spidroins are more widely known. Our sample of major ampullate spidroins is not comprehensive because we focused on sampling species for which multiple spidroins had been characterized. Using N and C-terminal sequences, Garb et al. (2010) recovered monophyletic groups for each of tubuliforms, flagelliforms, and minor ampullates in parsimony and Bayesian analyses. Entelegyne major ampullates spidroins were also recovered as monophyletic in their Bayesian analysis. N-terminal sequences have not been reported for aciniform and pyriform gland associated spidroins, or from any mygalomorph spidroins except for one (*Bothriocyrtum californicum* fib1). We did not recover N-terminal sequences in any of our libraries; thus we did not include published N-terminal sequences in our analyses. An SH test (Shimodaira and Hasegawa, 1999) using RAxML with the log likelihood values from the ML analyses was preformed to compare the constrained and unconstrained tree topologies.

The constrained ML spidroin gene tree was reconciled with a species tree based on hypothesized phylogenetic relationships (Ayoub et al., 2007a; Hedin and Bond, 2006; Coddington et al., 2004) using the program GeneTree 1.3 (Page, 1998). Spidroins lack a non-spider outgroup. Thus, rooting of the spidroin gene tree was based on the minimization of total gene duplications plus losses.

Characterization of Spidroin Non-Terminal Regions

Tandem repeats in spidroin protein sequences were identified using XSTREAM under default settings (Newman and Cooper, 2007) and by eye. Consensus repeat sequences and their lengths for each spidroin were determined based on 50% majority rule with ambiguities indicated by an X. We also determined the amino acid compositions of spidroin repetitive regions with MacVector 7.2 (Accelrys Inc., San Diego, CA; Table 1.2).

Using the ML C-terminal tree from the analysis with gland associated spidroins constrained, we performed continuous character, ancestral state reconstructions for amino acid compositions. Reconstructions were done using parsimony under the linear cost assumption in Mesquite v. 2.74 (Maddison and Maddison, 2010). Additionally, the Mesquite module, CoMET, was used to calculate the likelihood of observing the continuous data given the entire C-terminal tree (all branching events and branch lengths) under nine different models of evolution (Lee et al., 2006). These models include pure phylogenetic, non-phylogenetic, or punctuated average, in combination with distance, equal, or free (Oakley et al., 2005). The best fitting model was determined by the Akaike Information Criterion (Akaike, 1973). CoMET analyses were run with thresholds of 100 and 1000 for comparison of the pure phylogenetic and punctuated average models. The punctuated average model was favored if the data was indicated to have evolved from branching events where the branch lengths were 100 or, more conservatively, 1000 times longer than their corresponding sister branch lengths (CoMET User's Guide, Feb. 2006).

Results

Liphistius Egg Case Protein Homologs

BLASTX searches (Altschul et al., 1990) of cDNA clones identified six *Liphistius malayanus* transcripts with top hits to *Latrodectus hesperus* ECP1 (AY994149) and ECP2 (DQ341220). Thus, these *Liphistius* transcripts were named ECP-like (ECPL; GenBank accessions JX102548-JX102553). No ECP-like transcripts were detected in any of the mygalomorph cDNA libraries or the *Hypochilus thorelli* cDNA library. *Liphistius* ECPL names and cDNA lengths in base pairs (bp) in parentheses are as follows: ECPL1 (836), ECPL2 (724), ECPL3 (967), ECPL4 (969), ECPL5 (800), ECPL6 (950). With the exception of ECPL5, all of the *Liphistius* ECPL mRNA sequences included full length coding sequence. *Liphistius* ECPL transcripts are significantly shorter than *Latrodectus*

ECP1 and ECP2 transcripts, which are 2799 bp (coding, 932 amino acids (AA)) and 2478 bp (coding, 825 AA), respectively. The *Liphistius* ECPLs align to the non-repetitive, cysteine rich, N-terminal region, and lack most of the repetitive region of the *Latrodectus* ECPs (Figure 1.2a). The average pairwise similarity for amino acid sequences (gaps treated as missing) among *Liphistius* ECPLs is 58.26%, and 33.53% between *Liphistius* ECPLs and *Latrodectus* ECPs (Figure 1.2b).

Spidroin Gene Tree

One or more spidroins were identified in the cDNA libraries for each taxon in our study, for a total of 13 new spidroins (GenBank accessions JX102554-JX102566). All of the spidroin cDNAs were partial length transcripts, lacking 5' untranslated sequence, a start codon, N-terminal region sequence, and an unknown amount of repeat region sequence. Spidroin names with cDNA lengths (bp) in parentheses are as follows: *Liphistius* fib1 (3513); *Hypochilus* fib1 (2063) and fib2 (2190); *Aphonopelma seemanni* fib1 (1904), fib2 (1634), and fib3 (1464); *Poecilotheria regalis* fib1 (4617) and fib2 (2437); *Antrodiaetus riversi* fib1 (1833) and fib2 (5023); *Sphodros rufipes* fib1 (2460); and *Hexura picea* fib1 (409). *Megahexura fulva* fib1 (1257) contained a C-terminal encoding region but lacked a complete repeat; therefore, an additional clone (4897) of exclusively repetitive region was sequenced. Comparison of the repeat regions of these two clones confirmed that they likely represent parts of the same transcript. The two *Megahexura* fib1 clone sequences were combined in GenBank accession JX102566.

We used the tree based on the maximum likelihood (ML) analysis with constraints (Figure 1.3; Table 1.1) for reconciliation analysis and reconstruction of the evolution of continuous characters (Table 1.2). While tubuliform, aciniform, pyriform, and flagelliform spidroins were each recovered as monophyletic in all ML and Bayesian analyses, without these constraints, monophyletic groupings of neither major ampullate spidroins nor minor ampullate spidroins were recovered. However, monophyly of both major ampullates and minor ampullates is supported by a previous Bayesian analysis of combined N and C-terminal data (Garb et al., 2010). The ML constrained and unconstrained trees were identical at 46 of 58 nodes (Figure 1.3). Conflicting relationships were restricted to weakly supported nodes (Table 1.1). The Shimodaira-Hasegawa (SH) test (Shimodaira and Hasegawa, 1999) determined that the constrained topology was not significantly worse than the unconstrained topology. Both the constrained and unconstrained Bayesian consensus trees were unresolved at many nodes (Table 1.1). The ML and Bayesian constrained trees conflicted at only one node, where MiSps were placed sister to Flags in the ML analysis but sister to MaSps in the Bayesian analysis. The bootstrap percentage and posterior probability were weak for either relationship.

The modest support at many nodes on the spidroin gene tree is not surprising given the small character set available (only C-terminal encoding regions) and the deep divergences among the taxa sampled. Support values for nodes of the spidroin gene tree will likely be improved in the future with inclusion of N-terminal regions, which are available for only a limited subset of published spidroins (Garb et al., 2010). Our spidroin gene tree is generated from the broadest phylogenetic sampling of spider lineages to date and thus is the best available topology for reconciliation and ancestral character state reconstruction analyses.

Reconciliation analysis of the spidroin gene tree with the species tree supported the *Liphistius* spidroin as sister to all other spidroins (100 events=31 duplications+69 losses; Figure 1.4; Table 1.2). Twenty-five other rootings implied the same number of duplications, but at an increased loss cost. Alternative rootings with *Hypochilus* fib1, or *Hypochilus* fib1 plus *Liphistius* fib1, resulted in the next best reconciliation score (101 events=31 duplications+70 losses) compared to the optimal score (rooting with *Liphistius* fib1).

Reciprocally monophyletic araneomorph and mygalomorph spidroin groups were never recovered in the phylogenetic analyses. Based on the most parsimonious rooting, *Hypochilus* fib1 was found to be sister to all remaining opisthothele spidroins, while *Hypochilus* fib2 was placed sister to the orbicularian aciniform spidroins (Figure 1.4). Mygalomorph spidroins fell into two groups. The most basal mygalomorph group consisted of a tarantula spidroin (*Aphonopelma* fib1) and an atypoid spidroin (*Sphodros* fib1), and this clade of genes was sister to spidroins from the haplogynes, *Plectreurys* and *Diguetia*. Most mygalomorph spidroins clustered in a group that was sister to an araneomorph clade consisting of *Plectreurys* fib4 and all of the orbicularian pyriform spidroins. This second mygalomorph clade is characterized by a basal split between atypoid spidroins and non-atypoid sequences; however, relationships within these two groups did not necessarily follow accepted species relationships (Hedin and Bond, 2006).

Spidroin Repeats

XSTREAM (Newman and Cooper, 2007) analyses identified repeat sequences in 9 of the 13 newly characterized spidroins (spidroin sequences *Aphonopelma* fib1, *Aphonopelma* fib2, *Aphonopelma* fib3 and *Hexura* fib1 were too short to record iterated repeats). Consensus repeats and their lengths are displayed in Figure 1.5. Most consensus repeat lengths are between 140 and 200 AA. *Hypochilus* fib1 and *Antrodiaetus* fib1 are significantly shorter at 34 and 50 AA, respectively. XSTREAM identified two repeat types in *Hypochilus* fib2. The consensus length of the first type, corresponding to repeats found in residues 1-309, is 141 AA. In contrast, the consensus repeat length of type two is 8 AA, and corresponds to repeats within residues 350-519. The *Megahexura* fib1 consensus repeat, at 365 AA, was significantly longer than the repeats from the other newly characterized spidroins described here. Unlike *Euagrus* fib1, which has a repeat of similar length (342 AA; Gatesy et al., 2001; Garb et al., 2007), the *Megahexura* fib1 repeat could not be broken down into sub-repeats of approximately ~180 AA in length.

Repeat regions of most spidroins reported here are rich in alanine and serine, but low in glycine (Table 1.2). Proline, which is implicated in the extensibility of orb-weaver major ampullate and flagelliform silks (Savage and Gosline, 2008), is rare in the spidroins reported here as well as in previously reported mygalomorph spidroins (0-4.13%). Alanine and serine tandem repeats occur in all of the newly generated spidroin sequences, whereas iterations of other amino acids are less common (Figure 1.5). The repeat region compositions of alanine, glycine, and serine for all spidroins analyzed in this study are summarized in Table 1.2. The individual contributions of alanine, glycine, and serine relative to the total composition for each spidroin are displayed in a heat map (Figure 1.6). Alanine levels are variable across spidroins. Glycine and serine levels appear to trade-off with each other in that they exhibit large and opposite changes. Glycine deficiency and high serine levels are primarily found in *Liphistius*, mygalomorph, and haplogyne spidroins, as well as tubuliform, aciniform, and pyriform gland-associated spidroins. By contrast, *Deinopis* fib1a and fib1b along with major ampullate, minor ampullate, and flagelliform gland-associated spidroins, have high glycine levels but are deficient in serine.

Continuous character modeling of alanine, glycine, and serine amino acid compositions, given our preferred tree (Figure 1.3), were executed using CoMET (Lee et al., 2006). Optimal models (pure phylogenetic, non-phylogenetic, or punctuated average, in combination with distance, equal, or free; Oakley et al., 2005) were chosen by the Akaike Information Criterion (Akaike, 1973). For alanine composition, the punctuated average/equal model was selected under the asymmetry threshold of 100, but the pure-phylogenetic/distance model was selected under the asymmetry threshold of 1000. The punctuated average/equal model was selected for glycine composition under thresholds of 100 and 1000. The model selected for serine composition was pure-phylogenetic/distance under both thresholds.

For each of the newly characterized spidroins, comparison of DNA sequences across repeats of a particular molecule reveal a high degree of sequence similarity among repeats. *Hypochilus* fib1, fib2 repeat 1, and fib2 repeat 2 showed the lowest average percent identities across repeat types at 85%, 79% and 77%, respectively. Repeats in the mygalomorph spidroin, *Antrodiaetus* fib1, shared 87% identity. Repeats within each of the six other new spidroins with identifiable repeats were >98% identical. A very low total of 13 non-synonymous differences and 3 synonymous differences occur across the 546 bp long alignment of *Liphistius* fib1 repeats (Figure 1.7).

Discussion

Liphistius Silk Gene Diversity

The common ancestor of mesotheles and all other spiders is estimated to have existed more than 380 million years ago (Ayoub and Hayashi, 2009). This deep divergence and distant phylogenetic relationship with other spiders makes characterization of silk genes from Mesothelae crucial for obtaining a complete understanding of silk evolution. Mesotheles retain a number of plesiomorphic morphological characters associated with silk spinning (e.g., four pairs of anteriorally-placed spinnerets and single spigot types), and these spiders exhibit little variation in silk fiber types (Haupt, 2003; Haupt and Kovoor, 1993). However, mesotheles use silk for a variety of functions such as construction of their egg cases, burrow, trapdoor, and sensory lines (Haupt and Kovoor, 1993; Shultz, 1987). This combination of silk-spinning traits raises questions about the underlying diversity and function of silk genes and proteins from Mesothelae.

The *Liphistius* cDNA library included a considerable diversity of silk protein transcripts. In total, seven silk associated cDNAs were recovered, which approaches the number of different ortholog groups described from a single orb-weaver species and surpasses the number reported from most non-orbicularian araneomorph species (Gatesy et al., 2001; Blasingame et al., 2009; Garb et al., 2010). This diversity is surprising given the much simpler silk gland morphologies of *Liphistius* compared to araneomorph spiders. Six of the seven *Liphistius* silk cDNAs shared substantial sequence similarity to the ECPs (egg case proteins; BLASTX E values <1e-05), which have thus far only been reported from the Western black widow, *Latrodectus hesperus* (Hu et al., 2005b; Hu et al., 2006b). The six *Liphistius* egg case protein-like (ECPL) sequences group into three clusters. DNA sequence percent similarities across these three groups range from 49-57%. Within groups, percent similarities (gaps treated as missing) range from 96-100%. All of these sequences exhibit length differences in the protein-coding region, and for one of the groups, the only difference between members was a three-base pair indel. It is possible that some of the ECPL sequences represent allelic differences and/or splice variants.

The phylogenetic distribution of ECPs and ECPLs implies that egg case proteins either convergently evolved in *Liphistius* and *Latrodectus*, or that ECPs were present in the common ancestor of all extant spiders. Given the striking similarity of amino acids over a long region (~200 residues) and lack of significant similarity to any other proteins in the NCBI nr database, it seems unlikely that ECPs evolved convergently in mesotheles and in theridiid araneomorphs (Figure 1.2). Thus, we propose homology of *Latrodectus*

ECPs and *Liphistius* ECPLs. However, a recent study on silk gland transcriptomes from the mygalomorph, *Actinopus sp*., and an orbicularian araneomorph, *Gasteracantha cancriformis*, also did not report ECPs (Prosdocimi et al., 2011). If our hypothesis of homology is correct, ECPs must have been lost independently in many spider lineages. Alternatively, ECPs may be highly restricted in their timing of expression, eluding detection in most cDNA libraries. With the completion of spider genome sequences in the future, it will be possible to discern the presence, absence, or pseudogenization of ECPL genes in various spider taxa, and test the hypothesis of homology between the distantly related ECP and ECPL genes. In particular, synteny could provide additional evidence for orthology of ECPs from *Latrodectus* and ECPLs from *Liphistius*.

Both *Latrodectus* ECPs and *Liphistius* ECPLs are cysteine rich, with many cysteine positions conserved within and across species (Figure 1.2; Hu et al., 2005b; Hu et al., 2006b). However, *Liphistius* ECPLs are significantly shorter than *Latrodectus* ECPs, lacking most of the extensive repetitive region seen in *Latrodectus* ECPs. While the timing and specificity of ECPL expression in *Liphistius* is uncertain, the physiochemical conservation of 73% of amino acids at sites that are present in at least one ECPL and ECP suggests that these ECPLs have a cross-linking role in silk fiber formation similar to that proposed for ECPs (Hu et al., 2006b).

While mesotheles have high ECPL diversity, our cDNA screen suggests a low spidroin diversity, as only a single spidroin type (fib1) was detected in our *Liphistius* cDNA library. The presence of a spidroin in a mesothele confirms that the spidroin gene family evolved very early in Araneae and has an important role in silk production for all major spider groups that have been studied to date. Whether the *Liphistius* spidroin forms complexes with the ECPLs is currently unknown. In *Latrodectus*, ECPs form trimeric complexes with the N-terminal region of tubuliform spidroins (TuSp1) to make the outer silk wrapped around eggs (Hu et al., 2006b). The N-terminal region of *Liphistius* fib1 has not been characterized, but there are three cysteines in the C-terminal region that may allow for disulfide bonds with the ECPLs, as well as between fib1 monomers. Phylogenetic analyses did not recover a close relationship between *Liphistius* fib1 and TuSp1, indicating that TuSp1 is the result of spidroin duplication after the split of Opisthothelae from Mesothelae (Figures 1.1, 1.4). This implies that ECPs evolved prior to TuSp1. Thus, ECPs likely first were incorporated into silk fibers made with spidroins that were serving a more general purpose, and later became incorporated into *Latrodectus* tubuliform silk fibers, which are specialized for egg case construction.

Spidroin Evolution

The most parsimonious rooting of the spidroin gene tree using reconciliation analysis indicates that *Liphistius* fib1 is sister to all other spidroins (Figure 1.4). Alternative less parsimonious rootings of the spidroin gene tree are consistent with spidroin gene family duplications occurring prior to the split of mesotheles and opisthotheles (Table 1.2). While mesotheles may have retained a single spidroin type, opisthotheles underwent an extensive diversification of spidroins very early in their history. Non-monophyly of araneomorph spidroins and of mygalomorph spidroins confirms that duplications occurred prior to the initial split of opisthotheles (Garb et al., 2007; Prosdocimi et al., 2011). The common ancestor of opisthotheles minimally had five spidroin paralogs (Figure 1.4). These five paralogous gene lineages are now represented by 1) *Hypochilus* fib1, 2) a clade consisting of two mygalomorph spidroins and four haplogyne spidroins, 3) a clade consisting of orbicularian aciniform spidroins plus *Hypochilus* fib2 and orbicularian tubuliform spidroins plus *Plectreurys* fib3, 4) a clade consisting of the remaining mygalomorph spidroins and orbicularian pyriform spidroins plus *Plectreurys* fib4, and 5) a clade consisting of major and minor ampullates, orbicularian flagelliforms, and three additional *Deinopis* spidroins (Figure 1.4).

The spidroin gene tree allows for inference of the duplication history of spidroins and how the origins of these different gene copies relate to the diversification of silk glands and to the evolution of spigot morphology. Mygalomorphs generally have a single spigot type and silk glands that are largely uniform and acinous in shape, which is thought to be the ancestral condition for spiders (Palmer et al., 2982; Palmer, 1985; Schultz, 1987). Given the diversity of spidroins hypothesized in the opisthothele common ancestor, spidroin diversification preceded the evolution of morphologically distinct silk glands (Figure 1.4).

The last common ancestor of araneomorphs is believed to have possessed ampullate, aciniform, pyriform, and cribellate silk glands and differentiated spigot types for each of these glands (Coddington and Levi, 1991; Glatz, 1972; Platnick et al., 1991). Spidroin ortholog groups associated with these glands are represented in the opisthothele common ancestor, with the exception of the cribellate spidroins, which to date have not been identified (Figure 1.4). Additionally, tubuliform and aciniform spidroins are inferred to have resulted from gene duplication before the diversification of Araneomorphae. Tubuliform spigots are a synapomorphy for entelygyne araneomorphs, yet both tubuliform and aciniform spidroins have non-entelegyne relatives, consistent with spidroin diversification preceding the evolution of the morphologically distinct tubuliform gland and spigot type (Platnick et al., 1991; Coddington, 1989; Griswold et al., 1999). Based on our gene tree, the flagelliform, major ampullate, and minor ampullate spidroins appear to have diversified within Entelegynae. For the cDNA libraries from non-entelgyne spiders screened in this study, fibroins closely related to ampullate and flagelliform fibroins were not found.

As in two recent studies, we did not recover monophyly of mygalomorph spidroins in our phylogenetic analyses (Garb et al., 2007; Prosdocimi et al., 2011). In contrast to these other studies, our increased taxonomic sampling reveals that both of the mygalomorph spidroin clades include atypoid and non-atypoid spidroins, indicating that ancient spidroin duplicates may be retained in different mygalomorph taxa, as seen in *Aphonopelma* (Figure 1.4). However, some mygalomorph taxa, such as *Bothriocyrtum*, retain spidroin copies that are very similar to each other, consistent with recent gene duplication or homogenization via concerted evolution in this mygalomorph lineage.

Mesothele and mygalomorph species have evolved a wide variety of web architectures, including sheet-webs, purse-webs, and trapdoors (Coyle, 1986). Assuming that the spidroins we have characterized from these taxa are those used to construct their webs, the relationship between different web shapes and the spidroins used to construct them appears to be highly variable. In many cases, closely related spidroin proteins may be used in the construction of very different web architectures. For example, *Aliatypus* spiders construct trapdoors, yet their spidroin is most closely related to the spidroins of *Hexura* and *Megahexura*, which construct sheet-webs (Figure 1.3). On the other hand, very similar architectures may be built from very divergent spidroins. *Liphistius*, *Aliatypus*, *Aptostichus*, and *Bothriocyrtum* have convergently evolved trapdoors, and the spidroins found from most of these spiders are not closely related. Thus, the ability of mesothele and mygalomorph species to produce different web architectures does not seem to be constrained by the silk proteins produced.

Evolution of Spidroin Repeat Regions

Our analyses reveal very low nucleotide sequence variability among repeat units within a particular spidroin gene. Even *Hypochilus* fib2 and *Plectreurys* fib3, which are the only reported spidroins composed of two different ensemble repeat types (a tandem array of a particular ensemble repeat followed by a tandem array of a different ensemble repeat), have high sequence similarity across ensemble repeats of the same type. Homogenization of repeats is consistent with concerted evolution via intragenic gene conversion or unequal crossing over, and is a pattern typical of spidroins reported from mygalomorph and araneomorph spiders (Gatesy et al., 2001; Hayashi et al., 2004; Garb and Hayashi, 2005; Garb et al., 2007; Hayashi and Lewis, 2000; Perry et al., 2010). The homogenization of repeats seen in *Liphistius* (Figure 1.7) indicates that a gene architecture of tandemly arranged, homogenized repeats is an ancestral feature for spidroins.

Liphistius fib1 and nearly all mygalomorph spidroin repeats described here are ~180 AA long (157-194 AA; Figure 1.5). The exceptions are *Antrodiaetus* fib1 (repeat length of 50 AA) and *Megahexura* fib1 (365 AA). The *Megahexura* fib1 repeat could have arisen from a doubling of the unit of homogenization (~180 to ~360 AA), which has been postulated for the large size of the *Euagrus* fib1 repeat (342 AA). The *Euagrus* fib1 repeat can be divided into two subrepeats of approximately equal size that are 56% identical (Garb et al., 2007). This suggests that the *Euagrus* fib1 repeat arose from a change of the unit of homogenization from ~170 to 342 AA. The *Megahexura* fib1 repeat (365 AA) cannot be divided into two subrepeats, suggesting that extensive sequence divergence has occurred between its putative ~180 AA ancestral subrepeats. Further studies are needed to determine whether ~180 AA is an optimal length for mygalomorph and mesothele silk production. At present, studies on recombinant silk production have focused on number of repeats and fiber formation, but not the influence of repeat size on fiber formation and mechanical properties (Brooks et al., 2008; An et al., 2011).

Alanine, glycine, and serine are three of the major amino acid components of spider silks and the silks of other arthropods (Hu et al., 2006a; Sutherland et al., 2010); for the spidroins analyzed here, these three amino acids account for, on average, 64% of the total amino acid content of the repetitive region. The percentages of these common amino acids vary considerably across the spidroin gene tree (Figure 1.6, Table 1.2). For most spidroins, alanine levels fall within the range of 20-35%. This is also exhibited by the *Liphistius* spidroin (26.5% alanine), and ancestral state reconstruction posits 26-36% as the primitive condition for spider silks. The best fitting model for alanine, under the most conservative asymmetry threshold in CoMET, indicates that the branching patterns in the spidroin C-terminal tree and DNA sequence divergence level between C-terminal encoding regions predicts the divergence level of alanine percentage in the repetitive regions (Oakley et al., 2005).

The heat map of the percent compositions of serine and glycine across the spidroin gene tree indicates that they contrast strikingly with each other (Figure 1.6). *Liphistius* fib1, most myaglomorph spidroins, and most non-ampullate and non-flagelliform araneomorph spidroins exhibit moderately high serine levels, but are deficient in glycine. In contrast, ampullate and flagelliform spidroins show high levels of glycine and low levels of serine. The best fitting CoMET model determined for glycine percentage suggests that at branching events in the spidroin gene tree, one spidroin retains the ancestral glycine level while the other descendant gene lineage diverges (Oakley et al., 2005). Punctuated evolution of glycine could be due to selection for sequence encoding glycine rich motifs, spread rapidly throughout the gene by concerted evolution, and maintained thereafter by stabilizing selection. Glycine rich motifs are known to contribute to the high tensile strength and extensibility of major ampullate and flagelliform silk fibers, respectively (Hu et al., 2006a). As was the case for alanine, the best fitting model selected for serine percentage indicates that change in serine composition more closely reflects the spidroin relationships and level of spidroin C-terminal sequence divergences (Oakley et al., 2005). Therefore, the CoMET models suggest that, given the spidroin tree, alanine and serine percentages change gradually, whereas glycine levels exhibit a pattern of large change followed by stasis.

Spider silks vary greatly in mechanical performance across species and among silks associated with different gland types (Blackledge and Hayashi, 2006; Swanson et al., 2009; Boutry and Blackledge, 2010). Tensile testing of silks from representatives of *Liphistius* and mygalomorphs has shown that these silks have lower tensile strength than major ampullate silks and lack the high extensibility of flagelliform silks (Blackledge and Hayashi, 2006; Swanson et al., 2009). Thus far, silk mechanical properties have only been tested on a few mesothele species and theraphosid mygalomorphs (tarantulas). Our study reveals mygalomorph silk proteins with distinct molecular architectures that may enable unique, and perhaps exceptional, mechanical properties. For example, *Antrodiaetus* expresses two silk encoding genes, one of which (fib1) encodes a protein with a glycine percentage of ~30%, which is more comparable to major ampullate and minor ampullate silks (~24-45%) than theraphosid silks (<10%). Also, the repeat length encoded by *Megahexura* fib1 is ~365 AA, which is well above the known range of repeat lengths encoded by theraphosid spidroin genes (157-186 AA). Thus, broader examination of silk mechanical properties in different mygalomorphs is warranted.

Mesotheles and mygalomorphs mostly use their silks to line their burrows, construct retreats, make egg sacs, and extend their sensory area. Exceptionally extensible or strong silk may not be advantageous for these purposes (Coyle, 1986). These spiders rely on their size, power, and robust fangs to capture ground dwelling prey, and there is little need for silks capable of absorbing kinetic energy from flying insects. Instead, selection in mesothele and mygalomorph lineages may favor durable silks that are optimized for stability in subterranean conditions or for sensitivity in detection of vibrations from prey. The new silk genes we have found can be used to further investigate silk mechanical and functional properties and how these relate to the subterranean lifestyle of mesotheles and mygalomorphs.

Conclusion

Analysis of silk gland expression libraries from mesothele, paleocribellate, and mygalomorph spiders greatly clarifies the evolutionary history of silk in Araneae. The discovery of mesothele ECPL sequences that share conserved regions with *Latrodectus* ECPs suggests that these loci comprise a gene family which has been associated with silk production in spiders for >380 million years. Further research is needed to determine the phylogenetic breadth of this gene family in spiders, as well as how ECPs functionally interact with members of the spidroin gene family. Phylogenetic analysis of our new data from Mesothelae, Mygalomorphae, and Paleocribelletae suggests that the most recent common ancestor of all extant spiders had a single spidroin, and that diversification of spidroins by gene duplication had already occurred prior to the divergence of mygalomorphs and araneomorphs. We also found that repeat regions vary considerably in amino acid composition across different spidroin types. The punctuated pattern of change in glycine percentage could be due to selection for improved mechanical properties enabled by these characteristics, facilitated by concerted evolution quickly spreading desirable protein coding motifs throughout a spidroin gene.

Mesotheles and mygalomorphs construct a wide variety of web shapes and burrow entrance architectures. Considering the ecological function of mygalomorph and mesothele silks, selection on silk from these spiders may have favored properties associated with the largely subterranean niche they fill, such as durability for burrow maintenance and vibration transmission for prey capture (Coyle, 1986). The diversity of silk genes we have uncovered in mesotheles and mygalomorphs highlights the need for further exploration into the phylogenetic diversity of spiders for silk genes that encode unique silk mechanical properties.

References

Agnarsson I, Kuntner M, Blackledge TA (2010) Bioprospecting Finds the Toughest Biological Material: Extraordinary Silk for a Giant Riverine Orb spider. PloS ONE. 5(9): e11234.

Akaike H (1973) Information theory as an extension of the maximum likelihood principle. Pp. 267-281 In: Petrov BN, Csaki F (Eds) Second International Symposium on Information theory. Akademiai Kiado, Budapest. Proceeding of the 2nd International Symposium on Information Theory, Supplement. Problems of control and information theory.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J. Mol. Biol. 215: 403-410.

An B, Hinman MB, Holland GP, Yarger JL, Lewis RV (2011) Inducing β-Sheets Formation in Synthetic Spider Silk Fibers by Aqueous Post-Spin Stretching. Biomacromolecules. 12: 2375-2381.

Ayoub NA, Garb JE, Hedin M, Hayashi CY (2007a) Utility of the nuclear protein-coding gene, elongation factor-1 gamma (*Ef-1γ*), for spider systematics, emphasizing family level relationships of tarantulas and their kin (Araneae: Mygalomorphae). Mol. Phylogenet. Evol. 42: 394-409.

Ayoub NA, Garb JE, Tinghitella R, Collin MA, Hayashi CY (2007b) Blueprint for a High-Performance Biomaterial: Full-Length Spider Dragline Silk Genes. PLoS ONE. 6: e514.

Ayoub NA, Hayashi CY (2009) Spiders (Araneae). In: Hedges SB, Kumar S (Eds), The Timetree of Life. Oxford: Oxford University Press. Pp. 255-259.

Beuken E, Vink C, Bruggeman C (1998) One-step procedure for screening recombinant plasmids by size. BioTechniques. 24: 748–750.

Bittencourt D, Dittmar K, Lewis RV, Rech EL (2010) A MaSp2-like gene found in the Amazon mygalomorph spider *Avicularia juruensis*. Comp. Biochem. Phys., B. 155: 419-426.

Blackledge TA, Hayashi CY (2006) Silken toolkits: biomechanics of silk fibers spun by the orb web spider *Argiope argentata* (Fabricius 1775). J. Exp. Biol. 209: 2452-2461.

Blackledge TA, Scharff N, Coddington JA, Szüts T, Wenzel JW, et al. (2009) Reconstructing web evolution and spider diversification in the molecular era. P. Natl. Acad. Sci. USA. 106 (13): 5229-5234.

Blasingame E, Tuton-Blasingame T, Larkin L, Falick AM, Zhao L, et al. (2009) Pyriform Spidroin 1, a Novel Member of the Silk Gene Family That Anchors Dragline Silk Fibers in Attachment Discs of the Black Widow Spider, *Latrodectus hesperus*. J. Biol. Chem. 284(42): 29097-29108.

Boutry C, Blackledge TA (2010) Evolution of supercontraction in spider silk: structure-function relationship from tarantulas to orb-weavers. J. Exp. Biol. 213: 3505-3514.

Brooks AE, Stricker SM, Joshi SB, Kamerzell TJ, Middaugh CR, et al. (2008) Properties of Synthetic Spider Silk Fibers Based on *Argiope aurantia* MaSp2. Biomacromolecules. 9: 1506-1510.

Coddington JA (1989) Spinneret Silk Spigot Morphology: Evidence for the Monophyly of Orbweaving Spiders, Cyrtophorinae (Araneidae), and the Group Theridiidae Plus Nesticidae. J. Arachnol. 17: 71-95.

Coddington JA, Giribet G, Harvey MS, Prendini L, Walter DE (2004) Arachnida. In: Cracraft J, Donoghue M (Eds) Assembling the Tree of Life. New York: Oxford University Press. P. 296-318.

Coddington JA, Levi HW (1991) Systematics and Evolution of Spiders (Araneae). Annu. Rev. Ecol. Syst. 22: 565-592.

Coyle FA (1986) The Role of Silk in Prey Capture by Nonaraneomorph Spiders. In: Shear WA (Ed) Spiders: webs, behavior, and evolution. Palo Alto, CA: Stanford Univ. Press. P. 310-363.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5): 1792-1797.

Foelix RF (1996) Biology of Spiders. 2nd Ed. Oxford: Oxford University Press.

Foelix RF, Rast B, Peattie AM (2012) Silk secretion from tarantula feet revisited: alleged spigots are probably chemoreceptors. J. Exp. Biol. 215: 1084-1089.

Garb JE, Ayoub NA, Hayashi CY (2010) Untangling spider silk evolution with spidroin terminal domains. BMC Evol. Biol. 10: 243.

Garb JE, DiMauro T, Lewis RV, Hayashi CY (2007) Expansion and Intragenic Homogenization of Spider Silk Genes since the Triassic: Evidence from Mygalomorphae (Tarantulas and Their Kin) Spidroins. Mol. Biol. Evol. 24(11): 2454-2464.

Garb JE, DiMauro T, Vo V, Hayashi CY (2006) Silk genes support the single origin of orb webs. Science. 312(5781): 1762.

Garb JE, Hayashi CY (2005) Modular evolution of egg case silk genes across orb-weaving spider superfamilies. P. Natl. Acad. Sci. USA. 102(32): 11379-11384.

Gatesy J, Hayashi CY, Motriuk D, Woods J, Lewis RV (2001) Extreme Diversity, Conservation, and Convergence of Spider Silk Fibroin Sequences. Science. 291: 2603-2605.

Glatz L (1972) Der Spinnapparat haplogyner Spinnen (Arachnida, Araneae). Z. Morphol. Tiere. 72: 1-25.

Glatz L (1973) Der Spinnapparat der Orthognatha (Arachnida, Araneae). Z. Morphol Tiere. 75: 1-50.

Gorb SN, Niederegger S, Hayashi CY, Summers AP, Votsch W, et al. (2006) Silk-like secretion from tarantula feet. Nature. 443: 407.

Gosline JM, DeMont ME, Denny MW (1986) The structure and properties of spider silk. Endeavor. 10(1): 37-43.

Griffiths JR, Salantiri VR (1980) The strength of spider silk. J. Mater. Sci. 15: 491-496.

Griswold CE, Coddington JA, Platnick NI, Foster RR (1999) Towards a Phylogeny of Entelegyne Spiders (Araneae, Araneomorphae, Entelegynae). J. Arachnol. 27: 53-63.

Guerette PA, Ginzinger DG, Weber BH, Gosline JM (1996) Silk properties determined by gland-specific expression of a spider fibroin gene family. Science. 272: 112-115.

Haupt J (2003) The Mesothelae – a monograph of an exceptional group of spiders (Araneae: Mesothelae). Zoologica. 154: 1-102.

Haupt J, Kovoor J (1993) Silk-gland system and silk production in Mesothelae (Araneae). Ann. Sci. Nat. Zool.13e. ser., 14: 35-48.

Hayashi CY, Blackledge TA, Lewis RV (2004) Molecular and Mechanical Characterization of Aciniform Silk: Uniformity of Iterated Sequence Modules in a Novel Member of the Spider Silk Fibroin Gene Family. Mol. Biol. Evol. 21(10): 1950-1959.

Hayashi CY, Lewis RV (1998) Evidence from Flagelliform Silk cDNA for the Structural Basis of Elasticity and Modular Nature of Spider Silks. J. Mol. Biol. 275: 773-784.

Hayashi CY, Lewis RV (2000) Molecular Architecture and Evolution of a Modular Spider Silk Protein Gene. Science. 287: 1477-1479.

Hedin M, Bond JE (2006) Molecular phylogenetics of the spider infraorder Mygalomorphae using nuclear rRNA genes (18S and 28S): Conflict and agreement with the current system of classification. Mol. Phylogenet. Evol. 41: 454-471.

Hinman MB, Lewis RV (1992) Isolation of a Clone Encoding a Second Dragline Silk Fibroin. J. Biol. Chem. 267(27): 19320-19324.

Hu X, Kohler K, Falick AM, Moore AMF, Jones PR, et al. (2005b) Egg Case Protein-1. J. Biol. Chem. 280(22): 21220-21230.

Hu X, Kohler K, Falick AM, Moore AMF, Jones PR, et al. (2006b) Spider Egg Case Core Fibers: Trimeric Complexes Assembled from TuSp1, ECP-1, and ECP-2. Biochemistry-US. 45: 3506-3516.

Hu X, Lawrence B, Kohler K, Falick AM, Moore AMF, et al. (2005a) Araneoid Egg Case Silk: A Fibroin with Novel Ensemble Repeat Units from the Black Widow Spider, *Latrodectus hesperus*. Biochemistry-US. 44: 10020-10027.

Hu X, Vasanthavada K, Kohler K, McNary S, Moore AMF, et al. (2006a) Molecular mechanisms of spider silk. Cell. Mol. Life Sci. 63: 1986-1999.

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 17: 754-755.

Kovoor J (1987) Comparative structure and histochemistry of silk-producing organs in arachnids. In: Nentwig W (Ed) Ecophysiology of Arachnids. Berlin: Springer-Verlag. Pp. 160-186.

Lee C, Blay S, Mooers AO, Singh A, Oakley TH (2006) CoMET: A Mesquite package for comparing models of continuous character evolution on phylogenies. Evol. Bioinform. 2: 193-196.

Maddison WP, Maddison DR (2010) Mesquite: a modular system for evolutionary analysis. Version 2.74. Mesquite website. Available: http://mesquiteproject.org. Accessed 2011, June 12.

Miller MA, Pfeiffer W, Schwartz T (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Proceedings of the Gateway Computing Environments Workshop (GCE), 14 Nov. 2010, New Orleans, LA pp 1 - 8.

Newman AM, Cooper JB (2007) XSTREAM: A practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. BMC Bioinformatics. 8: 382.

Nicholas KB, Nicholas Jr HB (1997) GeneDoc: a tool for editing and annotating multiple sequence alignments. Distributed by the author.

Nylander JAA (2004) MrModeltest v2. Program distributed by the author. Evolutionary Biology Centre, Uppsala University.

Oakley TH, Gu Z, Abouheif E, Patel NH, Li W-H (2005) Comparative Methods for the Analysis of Gene-Expression Evolution: An Example Using Yeast Functional Genomic Data. Mol. Biol. Evol. 22(1): 40-50.

Page RDM (1998) GeneTree: comparing genes and species phylogenies using reconciled trees. Bioinformatics. 14: 819-820.

Palmer JM (1985) The silk and silk production of the funnel-web mygalomorph spider *Euagrus* (Araneae, Dipluridae). J. Morphol. 186: 195-207.

Palmer JM, Coyle FA, Harrison FW (1982) Structure and cytochemistry of the silk glands of the mygalomorph spider *Antrodiaetus unicolor* (Araneae, Antrodiaetidae). J. Morphol. 174: 269-274.

Peattie AM, Dirks J-H, Henriques S, Federle W (2011) Arachnids Secrete a Fluid over Their Adhesive Pads. PLoS ONE. 6(5): e20485.

Pérez-Miles F, Panzera A, Ortiz-Villatoro D, Perdomo C (2009) Silk production from tarantula feet questioned. Nature. 461: E9.

Perry DJ, Bittencourt D, Siltberg-Liberles J, Rech EL, Lewis RV (2010) Piriform Spider Silk Sequences Reveal Unique Repetitive Elements. Biomacromolecules. 11: 3000-3006.

Platnick NI (2011) The world spider catalog, version 12.0. American Museum of Natural History, Available: http://research.amnh.org/iz/spiders/catalog. DOI: 10.5531/db.iz.0001. Accessed 2011, August 8.

Platnick NI, Coddington J, Forster RR, Griswold CE (1991) Spinneret Morphology and the Phylogeny of Haplogyne Spiders (Araneae, Araneomorphae). Am. Mus. Novit. 3016: 1-73.

Prosdocimi F, Bittencourt D, da Silva FR, Kirst M, Motta PC, et al. (2011) Spinning Gland Transcriptomics from Two Main Clades of Spiders (Order: Araneae) –

Insights on Their Molecular, Anatomical and Behavioral Evolution. PLoS ONE. 6(6): e21634.

Rind FC, Birkett CL, Duncan B-JA, Ranken AJ (2011) Tarantulas cling to smooth vertical surfaces by secreting silk from their feet. J. Exp. Biol. 214: 1874-1879.

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19: 1572-1574.

Ronquist F, Huelsenbeck JP, van der Mark P (2005) mrbayes 3.1 manual, draft 5/26/2005, MrBayes website. Available: http://mrbayes.csit.fsu.edu/manual.php. Accessed 2010, October 14.

Savage KN, Gosline JM (2008) The role of proline in the elastic mechanism of hydrated spider silks. J. Exp. Biol. 211: 1948-1957.

Shear WA, Palmer JM, Coddington JA, Bonamo PM (1989) A Devonian Spinneret: Early Evidence of Spiders and Silk Use. Science. 246: 479-481.

Shimodaira H, Hasegawa M (1999) Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. Mol. Biol. Evol. 16: 1114-1116.

Shultz JW (1987) The Origin of the Spinning Apparatus in Spiders. Biol. Rev. 62: 89-113.

Simmons MP, Ochoterena H (2000) Gaps as Characters in Sequence-Based Phylogenetic Analyses. Syst. Biol. 49(2): 369-381.

Stamatakis A (2006a) Phylogenetic Models of Rate Heterogeneity: A High Performance Computing Perspective. In Proceedings of 20th IEEE/ACM International Parallel and Distributed Processing Symposium (IPDPS2006), High Performance Computational Biology Workshop, Rhodos, Greece.

Stamatakis A (2006b) RAxML-VI-HPC: Maximum Likelihood-based Phylogenetic Analyses with Thousands of Taxa and Mixed Models. Bioinformatics. 22(21): 2688–2690.

Stamatakis A, Hoover P, Rougemont J (2008) A Rapid Bootstrap Algorithm for the RAxML Web Servers. Syst. Biol. 57(5): 758-771.

Sutherland TD, Young JH, Weisman S, Hayashi CY, Merritt DJ (2010) Insect Silk: One Name, Many Materials. Annu. Rev. Entomol. 55: 171-188.

Suyama M, Torrents D, Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34: W609-W612.

Swanson BO, Anderson SP, DiGiovine C, Ross RN, Dorsey JP (2009) The evolution of complex biomaterial performance: The case of spider silk. Integr. Comp. Biol. 49(1): 21-31.

Swanson BO, Blackledge TA, Summers AP, Hayashi CY (2006) Spider Dragline Silk: Correlated and Mosaic Evolution in High-Performance Biological Materials. Evolution. 60(12): 2539-2551.

Ubick D, Paquin P, Cushing PE, Roth V (Eds) (2005) Spiders of North America: an identification manual. American Arachnological Society. 377 pages.

Vollrath F, Selden P (2007) The Role of Behavior in the Evolution of Spiders, Silks, and Webs. Annu. Rev. Ecol. Evol. S. 38: 819-846.

Xu M, Lewis RV (1990) Structure of a protein superfiber: Spider dragline silk. P. Natl. Acad. Sci. USA. 87: 7120-7124.

Tables

Table 1.1. Node support (ML bootstrap percentage (BP) and Bayesian posterior probability (PP)) for phylogenetic analyses. Node numbers refer to the phylogeny in Figures 1.3 and 1.4. Dashes refer to nodes with <50 BS or 0.5 PP support.

Node#	ML			Bayes	
	Constrained (C)	Unconstrained		Constrained (C)	Unconstrained
1	-	-		-	-
3	-	-		-	-
5	-	-		0.78	0.8
6	-	-		-	-
7	65	67		0.97	0.98
10	83	79		1.0	0.99
11	92	89		1.0	1.0
14	87	87		1.0	1.0
17	-	-		-	-
18	-	-		-	-
19	-	-		0.57	0.66
21	C, 97	90		C, 1.0	1.0
23	-	53		-	0.56
26	-	-		0.54	0.5
28	C, 74	-		C, 1.0	0.97
30	-	-		0.51	0.53
32	68	56		0.94	0.93
35	-	-		0.74	0.77
36	-	-		-	-
37	-	-		0.59	0.62
39	C, 94	92		C, 1.0	1.0
41	75	75		0.94	0.95
44	-	-		-	-
45	-	-		-	-
47	-	-		0.65	0.59
49	-	-		-	-
51	55	52		0.91	0.9
54	-	-		0.67	0.67
56	-	-		-	0.5
58	-	-		-	-
59	82	82		1.0	1.0
61	83	80		1.0	1.0
63	93	93		1.0	1.0
65	56	-		-	-
68	-	-		-	-
69	70	72		0.93	0.93
72	-	-		0.94	0.95
74	59	54		0.97	0.97
76	98	97		1.0	1.0
79	-	-		0.98	0.98

80	-	-		0.84	0.61
82	100	100		1.0	1.0
85	-	-		0.61	0.5
86	C, -	-		C, 1.0	-
87	82	69		1.0	0.99
89	67	61		0.81	0.82
91	98	98		1.0	1.0
94	-	-		0.94	0.71
95	63	-		0.99	0.94
97	95	94		1.0	1.0
100	82	77		0.96	0.96
101	100	100		1.0	1.0
104	99	99		1.0	1.0
107	-	-		-	-
108	C, 55	-		C, 1.0	-
110	-	-		0.69	-
113	C, 91	-		C, 1.0	0.99
115	100	100		1.0	1.0

Table 1.2. Continuous character data and alternative reconciliation based outgroups for ML constrained tree. Ancestral state parsimony optimization was determined by Mesquite v. 2.74 (Maddison and Maddison, 2010). Node numbers refer to the phylogeny in Figures 1.3, 1.4, and 1.6.

Node#	Alanine %	Glycine %	Serine %	Alternative Rooting/ Outgroup
				duplications/ losses/ deep coalescence events
1	26.5-36.44	3.15-12.83	17.18-19.03	
2 (*Liphistius*_fib1)	26.5	3.15	19.03	31/69/81
3	26.5-36.44	9.69-12.83	17.18-18.94	
4 (*Hypochilus*_fib1)	40.28	23.7	9.72	31/70/82
5	26.5-36.44	9.69-12.83	17.18-18.94	31/70/82
6	26.5-36.44	9.69-12.83	17.18-18.94	31/71/83
7	26.5-36.44	7.13	18.94	
8 (*Aphonopelma*_fib1)	8.15	7.13	18.94	
9 (*Sphodros*_fib1)	39.07	4.52	24.78	
10	29.1-36.44	13.15-19.0	12.64-18.94	
11	29.1-36.44	13.15-19.0	12.64-18.94	
12 (*Plectreurys*_fib2)	29.1	13.15	20.09	
13 (*Plectreurys*_fib1)	41.19	27.1	12.64	
14	36.44	19	11.74	
15 (*Diguetia*_MaSplike)	36.44	30.97	11.74	
16 (*Diguetia*_MaSplike2)	54.75	19	9.95	
17	25.26-29.65	9.69-12.83	17.18-18.94	31/71/83
18	25.26-29.65	9.69-12.83	20.53-21.46	31/72/84
19	25.26-29.65	9.69-12.83	20.53-21.46	31/74/86
20 (*Hypochilus*_fib2)	33.68	13.51	20.53	
21	14.67	9.69-12.83	20.53-21.46	
22 (*Uloborus*_AcSp1)	14.67	7.29	26.63	
23	14.15	12.83	20.53-21.46	
24 (*Latrodectus*_AcSp1)	13.43	12.83	13.63	
25 (*Argiope*_AcSp1)	14.15	15.5	21.46	
26	25.26-29.65	9.69	22.11-29.91	31/74/86
27 (*Plectreurys*_fib3)	25.26	2.14	30.94	
28	29.21-29.65	9.69	22.11-29.91	
29 (*Uloborus*_TuSp1)	29.65	9.69	29.91	
30	29.21-29.65	9.69	22.11-26.86	
31 (*Deinopis*_TuSp1)	34.5	11.78	20.34	
32	29.21	8.52	22.11-26.86	
33 (*Latrodectus*_TuSp1)	26.01	6.75	26.86	
34 (*Nephila*_TuSp1)	29.21	8.52	22.11	
35	24.23-27.59	9.69-12.83	17.18-18.94	31/72/84
36	22.44-27.59	8.25-12.83	17.18-18.94	31/73/85

37	22.44-27.59	8.25-12.83	17.18-18.94	
38 (*Plectreurys*_fib4)	22.44	13.85	17.18	
39	22.44-27.59	1.97-5.41	17.18-18.94	
40 (*Latrodectus*_PySp1)	45.15	0	7.26	
41	17.69	1.97-5.41	26.6	
42 (*Argiope*_PySp1)	17.69	5.41	27.76	
43 (*Nephila*_PySp1)	13.79	1.97	26.6	
44	22.44-27.59	8.25-12.83	17.18-21.38	
45	23.9-28.39	8.25-12.83	17.18-21.38	
46 (*Antrodiaetus*_fib1)	28.39	29.66	2.97	
47	23.9-28.39	6.23-10.53	22.09-23.07	
48 (*Antrodiaetus*_fib2)	23.9	4.92	31.45	
49	35.17	6.23-10.53	22.09-23.07	
50 (*Megahexura*_fib1)	38.07	10.53	22.09	
51	?	?	?	
52 (*Aliatypus*_fib1)	35.17	6.23	23.07	
53 (*Hexura*_fib1)	?	?	?	
54	20.7-27.59	8.25-8.9	19.37-21.38	
55 (*Poecilotheria*_fib1)	20.7	6.46	23.57	
56	20.7-27.59	8.25-8.9	19.37-21.38	
57 (*Aphonopelma*_fib2)	17.8	8.9	19.37	
58	20.7-33.23	8.25-8.9	21.38	
59	26.03-33.23	8.25-8.9	21.38	
60 (*Poecilotheria*_fib2)	33.23	10.15	21.38	
61	26.03-33.23	8.25	21.38	
62 (*Aphonopelma*_fib3)	26.03	8.25	20.32	
63	30.65-34.51	8.24-8.25	21.74-24.9	
64 (*Avicularia*_fib1a)	34.51	8.24	24.9	
65	30.65-34.51	8.24-8.25	21.74-24.9	
66 (*Avicularia*_fib1b)	30.65	9.57	21.74	
67 (*Avicularia*_fib1c)	36.95	5.99	25.2	
68	20.7-33.23	6.98-8.9	22.58-22.81	
69	20.7-33.23	6.98-8.9	22.81	
70 (*Euagrus*_fib1)	38.57	11.75	24.92	
71 (*Aptostichus*_fib1)	18.02	6.98	22.81	
72	20.7-33.23	6.07-6.19	22.58-22.81	
73 (*Aptostichus*_fib2)	19.97	6.07	23	
74	30.43-38.25	6.07-6.19	22.58-22.81	
75 (*Bothriocyrtum*_fib3)	38.25	6.19	21.04	
76	30.43-38.25	4.66	22.58-22.81	
77 (*Bothriocyrtum*_fib2)	38.28	4.66	22.58	
78 (*Bothriocyrtum*_fib1)	30.43	3.52	23.4	
79	24.23-27.59	40.22-40.91	6.9-18.94	31/73/85
80	24.23-27.59	40.91	6.9-18.94	31/77/89
81 (*Deinopis*_fib2)	16.48	40.91	19.13	31/84/96
82	27.59	46.06	6.9	31/84/96
83 (*Deinopis*_fib1b)	27.59	46.06	6.9	31/91/103
84 (*Deinopis*_fib1a)	31.14	48.23	4.86	31/91/103

85	24.23-27.59	40.22-40.91	6.7-6.73	31/77/89
86	24.23-27.59	40.22-40.91	6.7-6.73	31/81/93
87	24.23-32.25	40.22-40.91	6.12-6.73	31/85/97
88 (*Nephila*_MaSp1)	33.1	44.13	3.56	31/92/104
89	24.23-32.25	35.83-40.91	6.12-6.73	31/92/104
90 (*Nephila*_MaSp2)	22.26	35.09	7.55	
91	32.25	35.83-40.91	6.12	
92 (*Latrodectus*_MaSp2)	32.25	35.83	6.12	
93 (*Latrodectus*_MaSp1)	34.85	44.47	1.86	
94	24.23-26.5	40.22-40.91	6.7-6.73	31/85/97
95	24.23-26.5	43.55-43.6	6.7-6.73	
96 (*Peucetia*_MaSp1)	24.23	44.9	8.42	
97	24.23-26.5	43.55-43.6	6.7-6.73	
98 (*Dolomedes*_fib1)	23.12	43.55	6.51	
99 (*Dolomedes*_fib2)	28.96	43.6	6.73	
100	24.23-26.5	36.22-40.91	6.7-6.73	
101	23.61	36.22	4.46	
102 (*Deinopis*_MaSp2a)	19.41	36.22	4.45	
103 (*Deinopis*_MaSp2b)	23.61	32.97	4.46	
104	26.5	36.22-40.91	6.84	
105 (*Uloborus*_MaSp1)	26.92	44.23	8.65	
106 (*Uloborus*_MaSp2)	26.5	30.77	6.84	
107	24.23-27.59	40.22-40.91	6.7-6.73	31/81/93
108	24.23-35.47	40.22	6.7-6.73	31/86/98
109 (*Nephila*_MiSp1)	36.52	40.22	4.83	
110	24.23-35.47	34.16	10.73	
111 (*Deinopis*_MiSp1)	22.83	24.07	13.4	
112 (*Uloborus*_MiSp)	35.47	34.16	10.73	
113	5.49-12.01	45.58	6.7-6.73	31/86/98
114 (*Deinopis*_Flag)	3.49	45.58	6.7	
115	5.49-12.01	53	6.7-6.73	
116 (*Nephila*_Flag)	5.49	55.24	8.05	
117 (*Argiope*_Flag)	12.01	53	5.3	

Figures

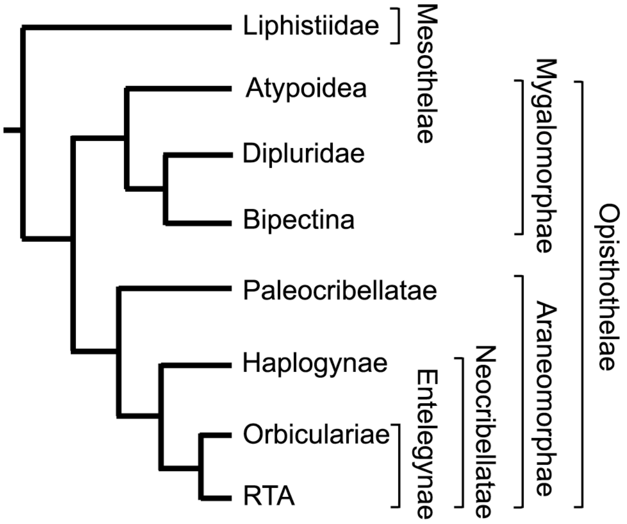


Figure 1.1. Phylogeny for spider groups analyzed in this study. Phylogeny is based on (Coddington and Levi, 1991; Hedin and Bond, 2006).

Figure 1.2. Alignment of Egg Case Proteins (ECPs) and Egg Case Protein-like proteins (ECPLs). A) Schematic of alignment of *Latrodectus hesperus* ECPs and *Liphistius malayanus* ECPLs. B) Alignment of amino acid sequences, abbreviated using single letters. Only partial *Latrodectus* (Latr) ECPs are shown as *Liphistius* (Liph) ECPLs lack the extended repetitive region. Alignment columns were highlighted using GeneDoc (Nicholas and Nicholas, 1997) according to physiochemical properties (Text color/Shade color: Proline Blue/Red; Glycine Green/Red; Tiny Blue/Yellow; Small Green/Yellow; Positive Red/Blue; Negative Green/Blue; Charged White/Blue; Amphoteric Red/Green; Polar Black/Green; Aliphatic Red/Gray; Aromatic Blue/Gray; Hydrophobic White/Black). Upper-case single letters occur above alignment positions showing 100% amino acid conservation, while lower case single letters occur above positions showing >50% conservation.

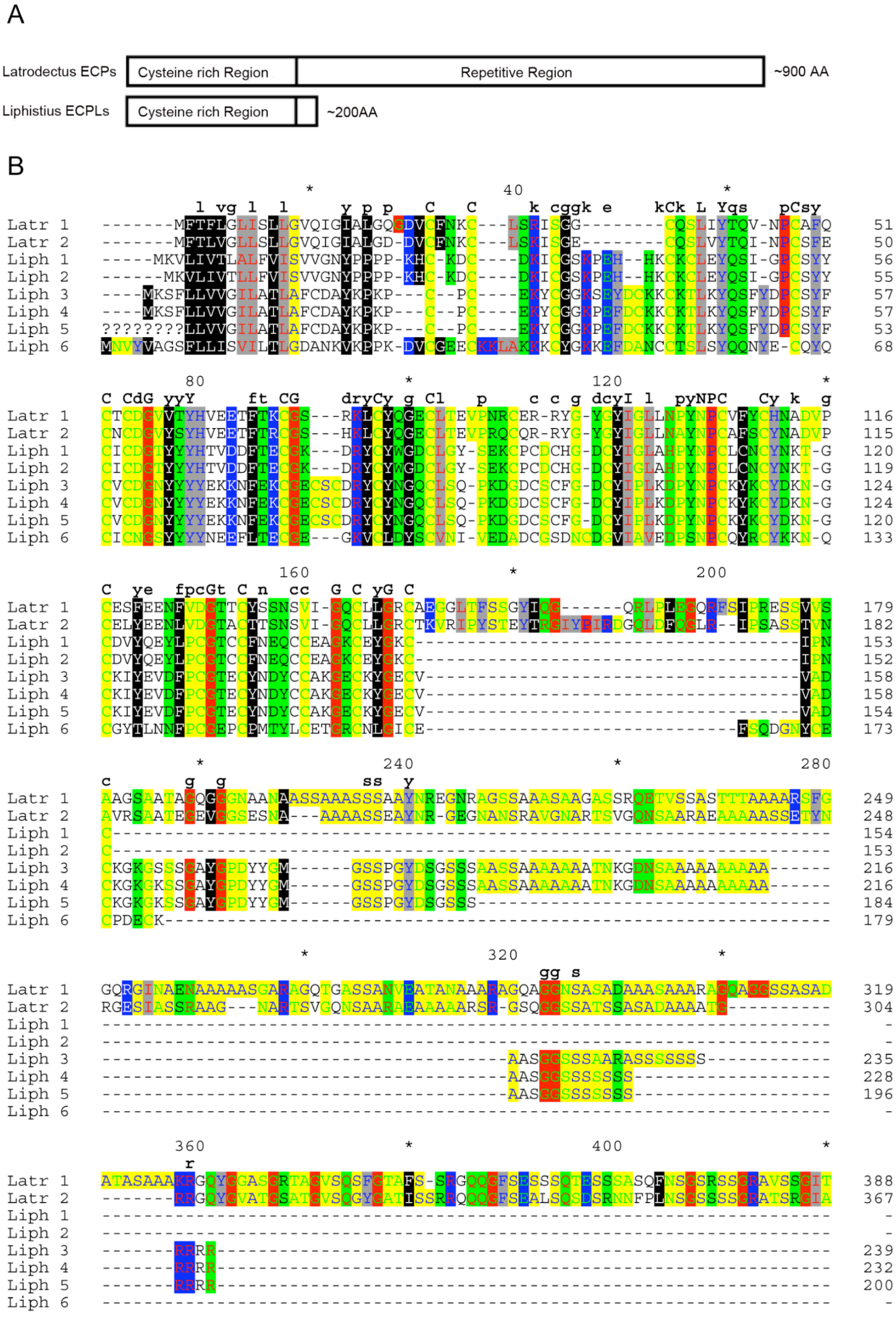

A
Latrodectus ECPs
Cysteine rich Region
Repetitive Region
~900 AA
Liphistius ECPLs
Cysteine rich Region
~200AA
B
* 40 *
l vg l l y p p C C k cggk e kCk L Yqs pCsy
Latr 1 -------MFTFIGLISLLGVQIGIALGQGDVCFNKC---LSRISGG--------CQSLIYTQV-NPCAFQ 51
Latr 2 -------MFTLVGLLSLLGVQIGIALGD-DVCFNKC---LSKISGE--------CQSLVYTQI-NPCSFE 50
Liph 1 -----MKVLIVTLALFVISVVGNYPPPP-KHC-KDC----DKICGSKPEH--HKCKCLEYQSI-GPCSYY 56
Liph 2 ------MKVLIVTLLFVISVVGNYPPPP-KHC-KDC----DKICGSKPEH--HKCKCLEYQSI-GPCSYY 55
Liph 3 ----MKSFLLVVGILATLAFCDAYKPKP---C--PC----EKYCGGKSEYDCKKCKTLKYQSFYDPCSYF 57
Liph 4 ----MKSFLLVVGILATLAFCDAYKPKP---C--PC----EKYCGGKPEFDCKKCKTLKYQSFYDPCSYF 57
Liph 5 ????????LLVVGILATLAFCDAYKPKP---C--PC----EKYCGGKPEFDCKKCKTLKYQSFYDPCSYF 53
Liph 6 MNVYVAGSFLLISVILTLGDANKVKPPK-DVCGEECKKLAKKKCYGKKEFDANCCTSLSYQQNYE-CQYQ 68
80 * 120 *
C CdG yyY ft CG dryCy g Cl p c c g dcyI l pyNPC Cy k g
Latr 1 CTCDGVVTYHVEETFTKCGS---RKLCYQGECLTEVPNRCER-RYG-YGYIGLLNPYNPCVFYCHNADVP 116
Latr 2 CNCDGVYSYHVEETFTRCGS---HKLCYQGECLTEVPRQCQR-RYG-YGYIGLLNAYNPCAFSCYNADVP 115
Liph 1 CICDGTYYYHTVDDFTECGK---DRYCYWGDCLGY-SEKCPCDCHG-DCYIGLAHPYNPCLCNCYNKT-G 120
Liph 2 CICDGTYYYHTVDDFTECGK---DRYCYWGDCLGY-SEKCPCDCHG-DCYIGLAHPYNPCLCNCYNKT-G 119
Liph 3 CVCDGNYYYYEKKNFEKCGECSCDRYCYNGQCLSQ-PKDGDCSCFG-DCYIPLKDPYNPCKYKCYDKN-G 124
Liph 4 CVCDGNYYYYEKKNFEKCGECSCDRYCYNGQCLSQ-PKDGDCSCFG-DCYIPLKDPYNPCKYKCYDKN-G 124
Liph 5 CVCDGNYYYYEKKNFEKCGECSCDRYCYNGQCLSQ-PKDGDCSCFG-DCYIPLKDPYNPCKYKCYDKN-G 120
Liph 6 CICNGSYYYYNEEFLTECGE---GKVCLDYSCVNI-VEDADCGSDNCDGVIAVEDPSNPCQYRCYKKN-Q 133
160 * 200
C ye fpcGt C n cc G C yG C
Latr 1 CESFEENFVDGTTCYSSNSVI-GQCLLGRCAEGGLTFSSGYIQG------QRLPLEGQRFSIPRESSVVS 179
Latr 2 CELYEENLVDGTACTTSNSVI-GQCLLGRCTKVRIPYSTEYTRGIYPIRDGQLDFQGLR--IPSASSTVN 182
Liph 1 CDVYQEYLPCGTCCRNEQCCEAGKCEYGKC-------------------------------------IPN 153
Liph 2 CDVYQEYLPCGTCCRNEQCCEAGKCEYGKC-------------------------------------IPN 152
Liph 3 CKIYEVDFPCGTECYNDYCCAKGECKYGECV------------------------------------VAD 158
Liph 4 CKIYEVDFPCGTECYNDYCCAKGECKYGECV------------------------------------VAD 158
Liph 5 CKIYEVDFPCGTECYNDYCCAKGECKYGECV------------------------------------VAD 154
Liph 6 CGYTLNNFPCGEPCEMTYLCETGRCNLGICE-----------------------------FSQDGNYCE 173
* 240 * 280
c g g ss y
Latr 1 AAGSAATAGQGGGNAANAASSAAASSSAAYNREGNRAGSSAAASAAGASSRQETVSSASTTTAAAARSFG 249
Latr 2 AVRSAATEGEVGGSESNA---AAAASSEAYNR-GEGNANSRAVGNARTSVGQNSAARAEAAAAASSETYN 248
Liph 1 C--- 154
Liph 2 C--- 153
Liph 3 CKGKGSSSGAYGPDYYGM------GSSPGYDSGSSSAASSAAAAAATNKGDNSAAAAAAAAA------ 216
Liph 4 CKGKGKSSGAYGPDYYGM------GSSPGYDSGSSSAASSAAAAAATNKGDNSAAAAAAAAA------ 216
Liph 5 CKGKGKSSGAYGPDYYGM------GSSPGYDSGSSS---------------------------------- 184
Liph 6 CPDECK-- 179
* 320 *
gg s
Latr 1 GQRGINAENAAAAASGARAGQTGASSANVEATANAAARAGQAGGNSASADAAASAAARAGQAGGSSASAD 319
Latr 2 RGESIASSRAAG---NARTSVGQNSAARAEAAAAARSR-GSQGGSSATSSASADAAAATG---------- 304
Liph 1 -- -
Liph 2 -- -
Liph 3 --------------------------------------AASGGSSSAARASSSSSS------------ 235
Liph 4 --------------------------------------AASGGSSSSSS------------------- 228
Liph 5 --------------------------------------AASGGSSSSSS------------------- 196
Liph 6 -- -
360 * 400 *
r
Latr 1 ATASAAAKRGQYGGASGRTAGVSQSFGTAFS-SRGQQGFSESSSQTESSSASQFNSGSRSSGRAVSSGIT 388
Latr 2 -------RRGQYGVATGSATGVSQGYGATISSRRQQQGFSEALSQSDSRNNFPLNSGSSSGRATSRGIA 367
Liph 1 -- -
Liph 2 -- -
Liph 3 -------RRRR--- 239
Liph 4 -------RRRR--- 232
Liph 5 -------RRRR--- 200
Liph 6 -- -

Figure 1.3. Spidroin gene tree based on ML analysis of the carboxy-terminal encoding region. In the analysis, gaps were coded as binary characters and monophyly of some groups was constrained (see Methods). Numbers at nodes correspond to information in supplementary Tables 1.1 and 1.2. Node numbers indicated in red are constrained nodes. Green dots indicate nodes that do not conflict between the analysis with node constraints and the unconstrained ML analysis. Dots at terminal nodes indicate web type constructed by the taxa from which the spidroin sequence was obtained (red=trapdoor, blue=sheetweb, purple=purseweb, teal=turret). Hash marks on branch indicate arbitrary shortening of branch for figure quality purposes. Brackets indicate taxonomic group of spiders from which spidroins were characterized and select spidroin clades using the following abbreviations: Me=Mesothelae, My=Mygalomorphae, Ar=Araneomorphae, AcSp=Aciniform, TuSp=Tubuliform, PySp=Pyriform, MaSp=Major ampullate, MiSp=Minor ampullate, Flag=Flagelliform.

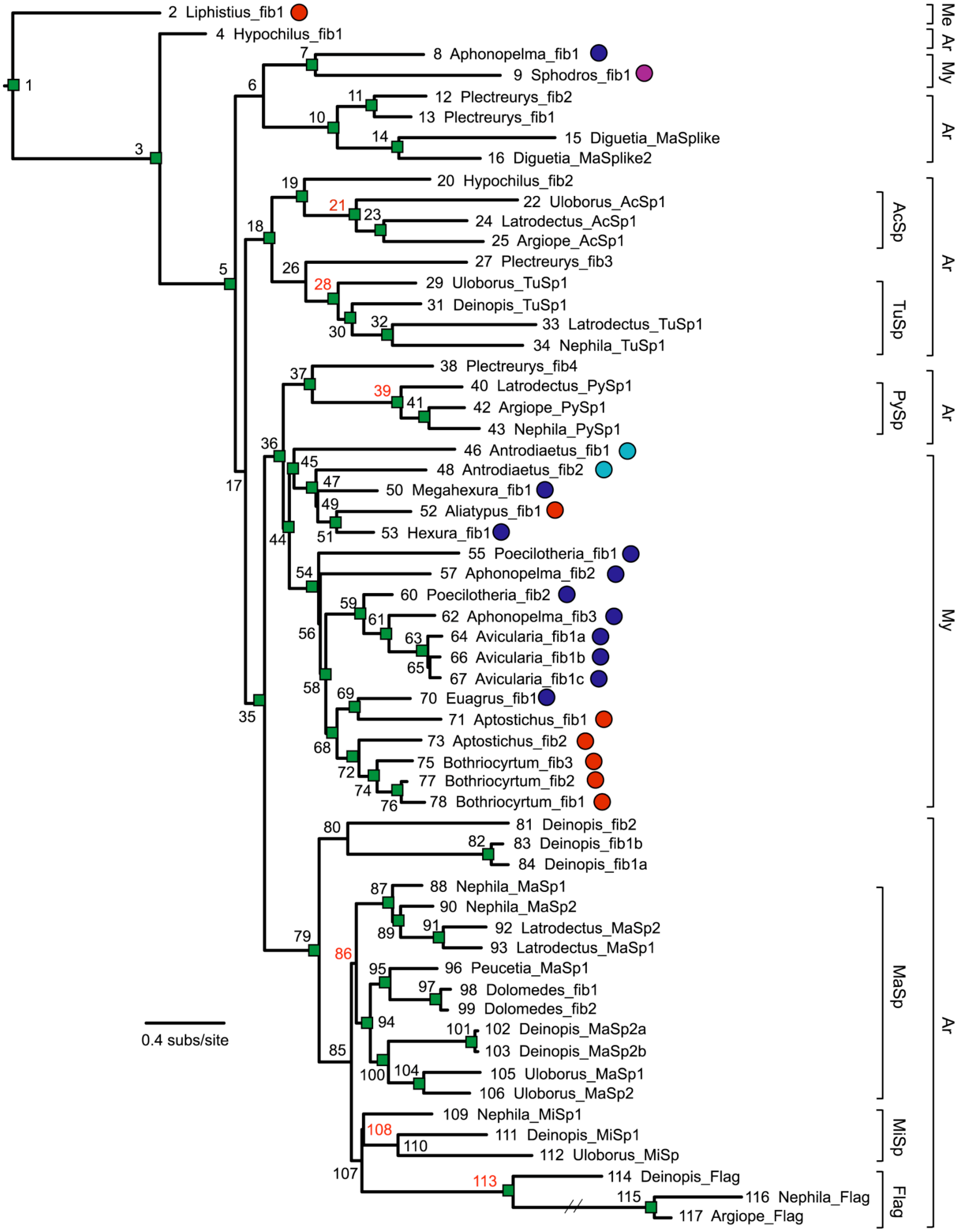
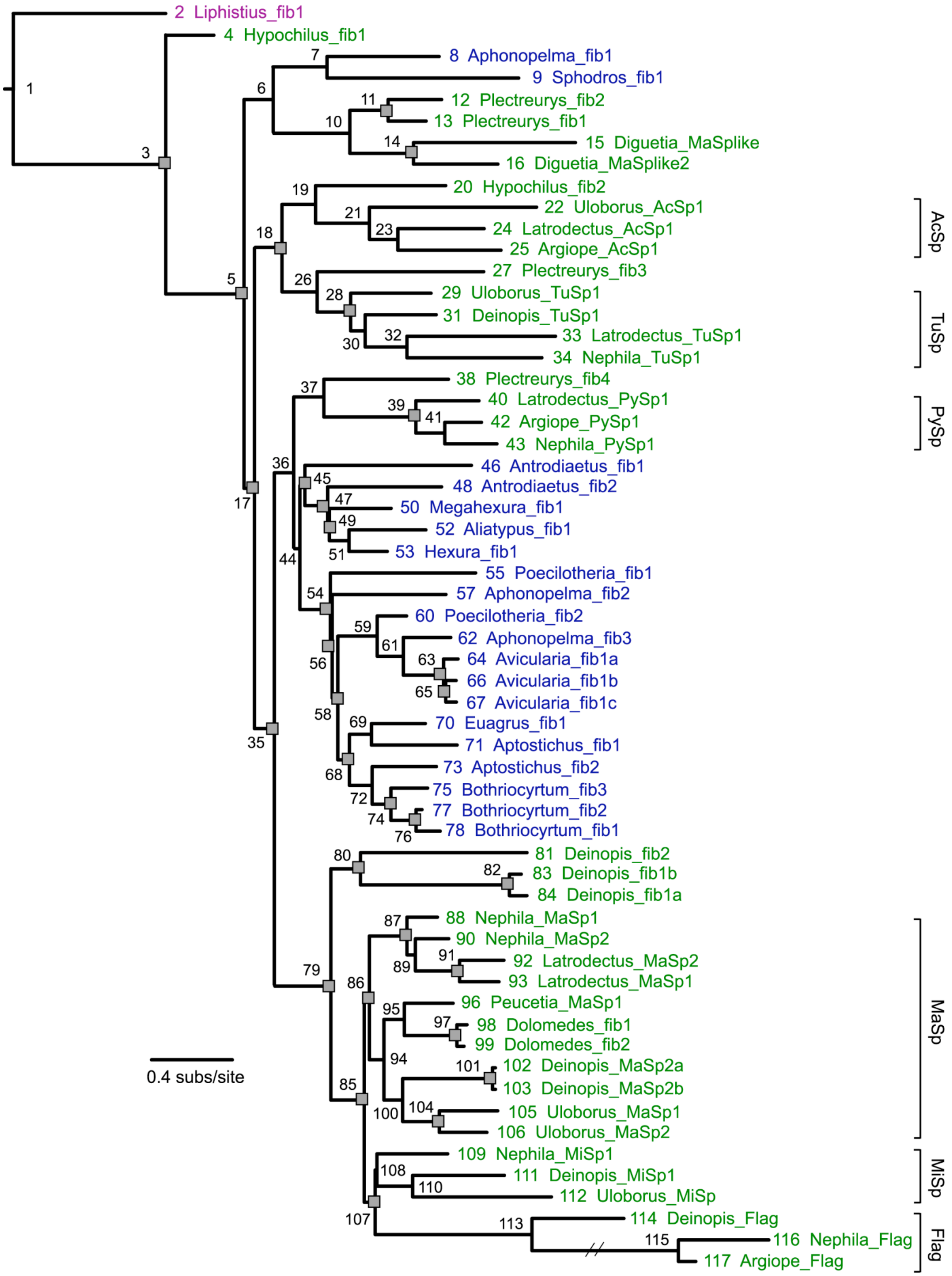

1
2 Liphistius_fib1
3
4 Hypochilus_fib1
5
6
7
8 Aphonopelma_fib1
9 Sphodros_fib1
10
11
12 Plectreurys_fib2
13 Plectreurys_fib1
14
15 Diguetia_MaSplike
16 Diguetia_MaSplike2
17
18
19
20 Hypochilus_fib2
21
22 Uloborus_AcSp1
23
24 Latrodectus_AcSp1
25 Argiope_AcSp1
26
27 Plectreurys_fib3
28
29 Uloborus_TuSp1
30
31 Deinopis_TuSp1
32
33 Latrodectus_TuSp1
34 Nephila_TuSp1
35
36
37
38 Plectreurys_fib4
39
40 Latrodectus_PySp1
41
42 Argiope_PySp1
43 Nephila_PySp1
44
45
46 Antrodiaetus_fib1
47
48 Antrodiaetus_fib2
49
50 Megahexura_fib1
51
52 Aliatypus_fib1
53 Hexura_fib1
54
55 Poecilotheria_fib1
56
57 Aphonopelma_fib2
58
59
60 Poecilotheria_fib2
61
62 Aphonopelma_fib3
63
64 Avicularia_fib1a
65
66 Avicularia_fib1b
67 Avicularia_fib1c
68
69
70 Euagrus_fib1
71 Aptostichus_fib1
72
73 Aptostichus_fib2
74
75 Bothriocyrtum_fib3
76
77 Bothriocyrtum_fib2
78 Bothriocyrtum_fib1
79
80
81 Deinopis_fib2
82
83 Deinopis_fib1b
84 Deinopis_fib1a
85
86
87
88 Nephila_MaSp1
89
90 Nephila_MaSp2
91
92 Latrodectus_MaSp2
93 Latrodectus_MaSp1
94
95
96 Peucetia_MaSp1
97
98 Dolomedes_fib1
99 Dolomedes_fib2
100
101
102 Deinopis_MaSp2a
103 Deinopis_MaSp2b
104
105 Uloborus_MaSp1
106 Uloborus_MaSp2
107
108
109 Nephila_MiSp1
110
111 Deinopis_MiSp1
112 Uloborus_MiSp
113
114 Deinopis_Flag
115
116 Nephila_Flag
117 Argiope_Flag
0.4 subs/site
Me
Ar
My
Ar
AcSp
Ar
TuSp
PySp
Ar
My
MaSp
Ar
MiSp
Flag

Figure 1.4. Spidroin gene tree with inferred duplication events. Spidroin gene tree is based on a ML analysis of the carboxy-terminal encoding region with gaps coded as binary characters and monophyly of some groups constrained (see Methods). Numbers next to nodes and terminals correspond to numbers in supplementary Tables 1.1 and 1.2 showing support values, alternate rootings, and continuous character data. Spidroins are colored according to the taxonomic group from which they were characterized: purple=Mesothelae, blue=Mygalomorphae, green=Araneomorphae. Gray squares indicate duplication events inferred by reconciliation. Hash marks on branch indicate arbitrary shortening of branch for figure quality purposes. Brackets indicate clades with the following abbreviations: AcSp=Aciniform, TuSp=Tubuliform, PySp=Pyriform, MaSp=Major ampullate, MiSp=Minor ampullate, Flag=Flagelliform.


1
2 Liphistius_fib1
3
4 Hypochilus_fib1
5
6
7
8 Aphonopelma_fib1
9 Sphodros_fib1
10
11
12 Plectreurys_fib2
13 Plectreurys_fib1
14
15 Diguetia_MaSplike
16 Diguetia_MaSplike2
17
18
19
20 Hypochilus_fib2
21
22 Uloborus_AcSp1
23
24 Latrodectus_AcSp1
25 Argiope_AcSp1
AcSp
26
27 Plectreurys_fib3
28
29 Uloborus_TuSp1
30
31 Deinopis_TuSp1
32
33 Latrodectus_TuSp1
34 Nephila_TuSp1
TuSp
35
36
37
38 Plectreurys_fib4
39
40 Latrodectus_PySp1
41
42 Argiope_PySp1
43 Nephila_PySp1
PySp
44
45
46 Antrodiaetus_fib1
47
48 Antrodiaetus_fib2
49
50 Megahexura_fib1
51
52 Aliatypus_fib1
53 Hexura_fib1
54
55 Poecilotheria_fib1
56
57 Aphonopelma_fib2
58
59
60 Poecilotheria_fib2
61
62 Aphonopelma_fib3
63
64 Avicularia_fib1a
65
66 Avicularia_fib1b
67 Avicularia_fib1c
68
69
70 Euagrus_fib1
71 Aptostichus_fib1
72
73 Aptostichus_fib2
74
75 Bothriocyrtum_fib3
76
77 Bothriocyrtum_fib2
78 Bothriocyrtum_fib1
79
80
81 Deinopis_fib2
82
83 Deinopis_fib1b
84 Deinopis_fib1a
85
86
87
88 Nephila_MaSp1
89
90 Nephila_MaSp2
91
92 Latrodectus_MaSp2
93 Latrodectus_MaSp1
94
95
96 Peucetia_MaSp1
97
98 Dolomedes_fib1
99 Dolomedes_fib2
100
101
102 Deinopis_MaSp2a
103 Deinopis_MaSp2b
104
105 Uloborus_MaSp1
106 Uloborus_MaSp2
MaSp
107
108
109 Nephila_MiSp1
110
111 Deinopis_MiSp1
112 Uloborus_MiSp
MiSp
113
114 Deinopis_Flag
115
116 Nephila_Flag
117 Argiope_Flag
Flag
0.4 subs/site

Liphistius malayanus fib1 (182 AA)
AEARSASYSGAVAKSFAESFADVIRRNDRYGSSYDSDLVSRYPSAYDDAISEALYQNTYFRGINRVAIAEA
VARARAEAGAGASSSAYAESHASALVRLLQSYGVITRDIVLSVAEATATAFASAVAKATSHARAESSAAAL
AKAAASSEAKSSTVTITTSEARAAAAAEASAESSASSSSE

Hypochilus thorelli fib1 (34 AA)
QQGQGGSSAAAAAAAAAAAAAGASGAQGQGQGYG

Hypochilus thorelli fib2
>type 1 (141 AA, corresponds to repeats in amino acid sequence 1-309)
GNLLAAAGYLATGGNASAIASSFASALSLFNVSSAAASAADAAGFAFGQLGSFVAAGSASSVSFAQAFASS
LVNSADFVSFCNSGVSPASIFPILRATAISLGYDETFASTVASAVAQSTATVGAGASASAYAQAWSTAIA
>type 2 (8 AA, corresponds to repeats in amino acid sequence 350-519)
AASGASAG

Megahexura fulva fib1 (365 AA)
QASYYASASAQAAASAAGFGSSVANAVASASASAAGGVSVGAGAQAYASALGNGLGQALLANGVLNSGNYL
QLANSLAYSFGSSLSQYSSSAAGASAAGAASGAAGAGAGAASSGGSSGSASSSTTTTTTTSTSAAAAAAAA
AAAASAAASTSASASASASASASASAFSQTFVQTVLQSAAFGSYFGGNLSLQSAQAAASAAAQAAAQQIGLGS
YGYALANAVASAFASAGANAGSYAYAQAAAGAFANVLFQAGVLTYANASALASAYASAYASSVASAAASAS
AGASASSGASASASAAAAAAAAAAAASASAGASASAGASASAGASASASASATASAFASAFASAVVQTGYFS
NVFGNVYSAA

Sphodros rufipes fib1 (192 AA)
VYACANASSASAIAYAIAFAFAQALVSSSFATASSFRSISTAFAQALVSAVASSAASATSAAAAASAAASA
SGAAAVAAAGAAGAAAGSSTSTTTTTAAASAAASSASAAAAGSAASSAAAARASASATSVSQTVSSFLVQS
SRFQSAVSSLYALGTDAYSSAYADAAARAISGAGFSSAEASAFASSVYNA

Antrodiaetus riversi fib1 (50 AA)
GAAGAGADGAAGGAGGYGREGARASAEAKAYAGEARAGGYGRERAAGGEA

Antrodiaetus riversi fib2 (194 AA)
ESGSGSSSYARAYAVAKSSAVVLQNSGVLSSSNSRSLSSAFAKAFASSSASSYASYKSRNTGTSRASSAGS
GAAGSATSSTTTTTTVSKSAAAAAAAAAAASSASSASRDSDRSSSTAAEAYASSSAYASSNFENYLASDLS
NSDEFESVYGSLYSSSDAASYARSSAEYASNTLGLSKETSYALASAAAKAVS

Poecilotheria regalis fib1 (157 AA)
HHLGVLTAANGNILVFQLANLIPSAMSSSYSAVSTGSAAAAASSASSATSSTTTTSTSSAAASSSAAASAS
TADYTSSLVSLLVSNTEFRSGVNEITSLSAANAVSYAIAKSTADYLGIANYTSLASALSTSISGVGIGGSA
NSYAFAIAGPTLKFL

Poecilotheria regalis fib2 (166 AA)
AVASAGNNAGAYAYARAYASAISQSLSSLGILNSGNAIALANAFSSGASDSAAAAALSAASASAASAATAA
STTTSTSTSAATAAAAAASAAGAAGAGAAAGATASSSFGQNLLSGLLRSDAVVSALSQAYSASTASALASS
YAQSGADRAGLGNYGSVIASAAAS

Figure 1.5. Majority rule consensus of ensemble repeats within spidroins. Ensemble repeats are tandemly arrayed. Amino acid sequences with single letter abbreviations are shown. Alanine (red), serine (blue), and glycine (green) are highlighted. Single amino acids repeated in tandem are underlined. Repeat lengths are given in parentheses.

Figure 1.6. Heat map of percent compositions of alanine, glycine, and serine from spidroin repetitive regions. Cladogram adjacent to heat map shows relationships as in Figures 1.3 and 1.4. *Hexura* fib1 was omitted since no repetitive region sequence was obtained for that cDNA. Here, red indicates levels furthest below the mean, while white indicates levels furthest above the mean. Histograms on columns also show relative composition levels of the three amino acids across spidroins. Spidroin colors and abbreviations for clade names are as in Figure 1.4. Numbers at nodes correspond to information in supplementary Tables 1.1 and 1.2.

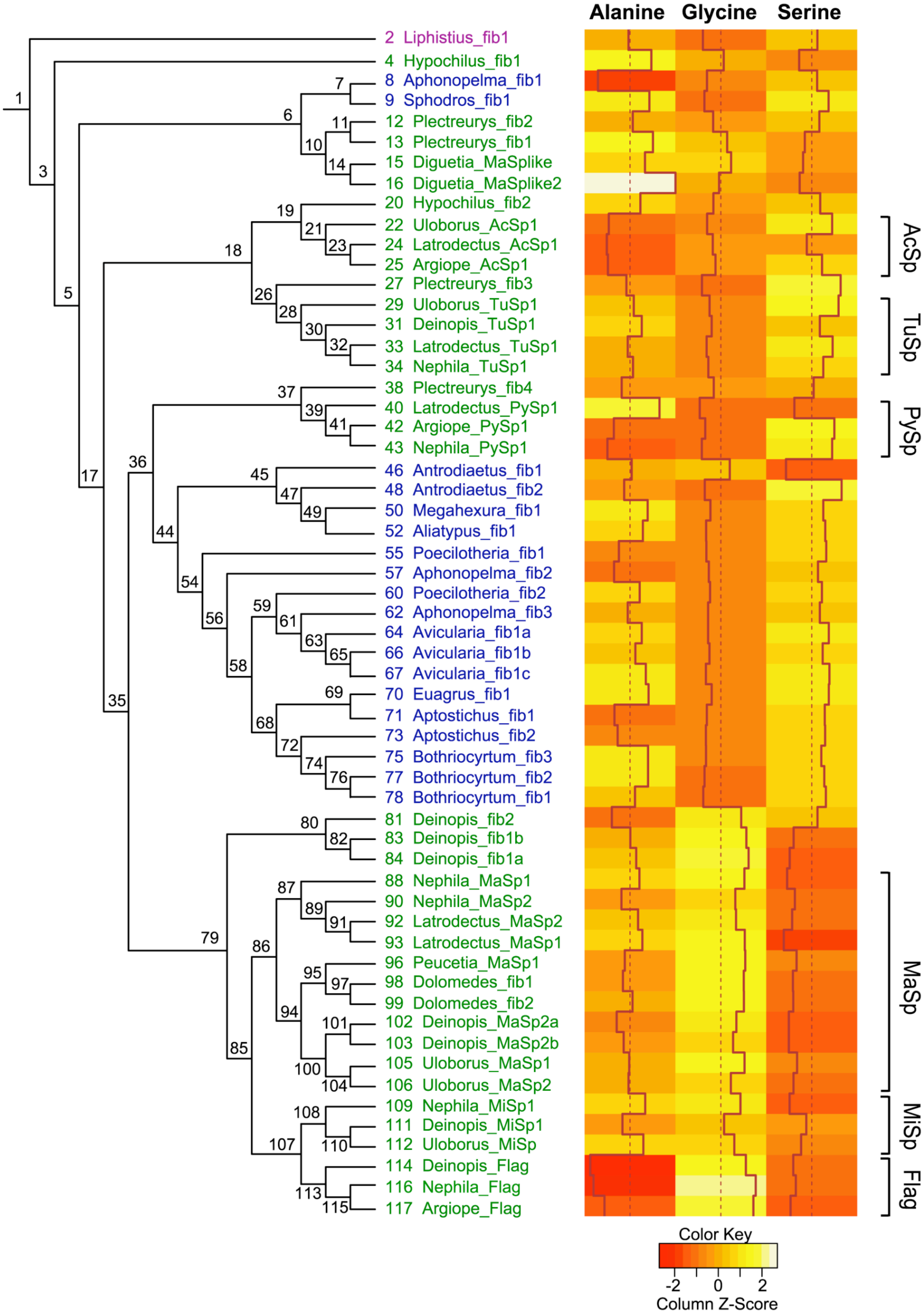

Alanine
Glycine
Serine
2 Liphistius_fib1
4 Hypochilus_fib1
8 Aphonopelma_fib1
9 Sphodros_fib1
12 Plectreurys_fib2
13 Plectreurys_fib1
15 Diguetia_MaSplike
16 Diguetia_MaSplike2
20 Hypochilus_fib2
22 Uloborus_AcSp1
24 Latrodectus_AcSp1
25 Argiope_AcSp1
27 Plectreurys_fib3
29 Uloborus_TuSp1
31 Deinopis_TuSp1
33 Latrodectus_TuSp1
34 Nephila_TuSp1
38 Plectreurys_fib4
40 Latrodectus_PySp1
42 Argiope_PySp1
43 Nephila_PySp1
46 Antrodiaetus_fib1
48 Antrodiaetus_fib2
50 Megahexura_fib1
52 Aliatypus_fib1
55 Poecilotheria_fib1
57 Aphonopelma_fib2
60 Poecilotheria_fib2
62 Aphonopelma_fib3
64 Avicularia_fib1a
66 Avicularia_fib1b
67 Avicularia_fib1c
70 Euagrus_fib1
71 Aptostichus_fib1
73 Aptostichus_fib2
75 Bothriocyrtum_fib3
77 Bothriocyrtum_fib2
78 Bothriocyrtum_fib1
81 Deinopis_fib2
83 Deinopis_fib1b
84 Deinopis_fib1a
88 Nephila_MaSp1
90 Nephila_MaSp2
92 Latrodectus_MaSp2
93 Latrodectus_MaSp1
96 Peucetia_MaSp1
98 Dolomedes_fib1
99 Dolomedes_fib2
102 Deinopis_MaSp2a
103 Deinopis_MaSp2b
105 Uloborus_MaSp1
106 Uloborus_MaSp2
109 Nephila_MiSp1
111 Deinopis_MiSp1
112 Uloborus_MiSp
114 Deinopis_Flag
116 Nephila_Flag
117 Argiope_Flag
1
3
5
6
7
10
11
14
17
18
19
21
23
26
28
30
32
35
36
37
39
41
44
45
47
49
54
56
58
59
61
63
65
68
69
72
74
76
79
80
82
85
86
87
89
91
94
95
97
100
101
104
107
108
110
113
115
AcSp
TuSp
PySp
MaSp
MiSp
Flag
Color Key
-2
0
2
Column Z-Score

Figure 1.7. Alignment of DNA sequences for *Liphistius* fib1 repeats. Amino acid translation and DNA consensus sequences are above repeat sequences. Dots indicate identity to the consensus sequence. Non-synonymous and synonymous differences from the consensus are indicated by upper and lower case letters, respectively.

```
           A  E  A  R  S  A  S  Y  S  G  A  V  A  K  S  F  A  E  S  F  A  D  V  I  R  R  N  D  R  Y  G 
con       GCAGAAGCAAGAAGTGCATCATATTCTGGGGCAGTGGCAAAATCCTTCGCCGAATCTTTCGCCGATGTTATACGCAGGAACGACAGGTATGGT  93
rep1      .............................................................................................  93
rep2      .............................................................................................  93
rep3      ............................A................................................................  93
rep4      ..............................................A..................................T...........  93
rep5      .........................................................................................c...  93

           S  S  Y  D  S  D  L  V  S  R  Y  P  S  A  Y  D  D  A  I  S  E  A  L  Y  Q  N  T  Y  F  R  G 
con       TCGTCCTACGATTCAGACTTGGTTTCGCGATATCCATCTGCATACGACGATGCGATCTCAGAGGCCTTATATCAGAATACCTATTTTAGGGGC 186
rep1      ............................................................................................. 186
rep2      ............................................................................................. 186
rep3      ............................................................................................. 186
rep4      .........................................................A................................... 186
rep5      ............................................................................................. 186

           I  N  R  V  A  I  A  E  A  V  A  R  A  R  A  E  A  G  A  G  A  S  S  S  A  Y  A  E  S  H  A 
con       ATCAATAGAGTTGCCATTGCAGAAGCAGTTGCACGCGCTAGGGCGGAAGCTGGTGCTGGAGCGAGCAGTTCTGCGTATGCAGAGTCACATGCC 279
rep1      ............................................................................................. 279
rep2      ............................................................................................. 279
rep3      ............................................................................................. 279
rep4      ............................................................................................. 279
rep5      ............................................................................................. 279

           S  A  L  V  R  L  L  Q  S  Y  G  V  I  T  R  D  I  V  L  S  V  A  E  A  T  A  T  A  F  A  S 
con       TCGGCATTGGTCCGGCTATTGCAGAGTTACGGAGTGATCACACGCGACATCGTCTTAAGTGTTGCAGAAGCCACTGCCACTGCTTTTGCTAGT 372
rep1      ..............................A.............................................................. 372
rep2      ............................................................................................. 372
rep3      ............................................................................................. 372
rep4      ............................................................................................. 372
rep5      ......A...................................................................................... 372

           A  V  A  K  A  T  S  H  A  R  A  E  S  S  A  A  A  L  A  K  A  A  A  S  S  E  A  K  S  S  T 
con       GCAGTTGCAAAGGCAACGTCACATGCTCGTGCCGAATCCTCAGCAGCTGCTTTGGCGAAAGCCGCTGCGTCATCCGAGGCTAAGTCCTCCACC 465
rep1      ............................................................................................. 465
rep2      ............................................................................................. 465
rep3      .....................................................................A....................... 465
rep4      ............................................................................................. 465
rep5      ............................................................................................. 465

           V  T  I  T  T  S  E  A  R  A  A  A  A  A  E  A  S  A  E  S  S  A  S  S  S  S  E 
con       GTCACCATAACGACCAGTGAAGCACGGGCAGCAGCTGCGGCTGAAGCGTCTGCTGAATCATCTGCGTCGTCTTCGTCAGAG 546
rep1      .....t........................................................................... 546
rep2      ..............................................................g.................. 546
rep3      ..........................................AG.....A............................... 546
rep4      ..........................................AG..................................... 546
rep5      ...........................C..................................................... 546
```

Chapter 2

Mosaic evolution of Silk Genes in *Aliatypus* Trapdoor Spiders (Mygalomorphae, Antrodiaetidae)

Abstract

Spider silk genes are composed mostly of repetitive sequence that is flanked by non-repetitive terminal regions. Inferences about the evolutionary processes that have influenced silk genes have largely been made from analyses using distantly related taxa and ancient silk gene duplicates. These studies have primarily relied on comparisons across the conserved non-repetitive terminal regions to determine orthologous and paralogous relationships, as well as the influence of selection acting on silk genes. While the repetitive region makes up the bulk of a spider silk gene and heavily influences silk fiber mechanical properties, few molecular evolutionary analyses have been conducted on this region as homology is often difficult, sometimes impossible, to determine. Here, we sample internal repetitive and carboxy terminal regions from all described species of the trapdoor spider genus, *Aliatypus*. *Aliatypus* spiders are highly dispersal limited and rely on their silk lined burrow for protection. We are able to determine positional homology across species for the carboxy terminal regions and relative positional homology for the internal repetitive regions. Gene trees based on each of these regions are in good agreement with the *Aliatypus* species tree indicating that a single spidroin ortholog was sampled in each species. Additionally, we test for signatures of selection and find that purifying selection and concerted evolution both have acted to conserve *Aliatypus* spidroin internal repetitive regions. In contrast, selection testing identifies evidence of sites that evolved under positive selection and amino acid replacements that result in radical physicochemical changes in the carboxy terminal region. These findings indicate that comparison of spidroin orthologs across a comprehensive sample of congenerics is useful for assessing the molecular evolutionary forces that have shaped the genes encoding silk fibers.

Key Words: Alpha helical tendency, Concerted evolution, Positive destabilizing selection, Repetitive sequence, Spidroin

Introduction

Spiders generate and utilize silk throughout their lifetime for numerous essential functions. Silk fibers are composed mostly of proteins known as spidroins (Hinman and Lewis, 1992), which are encoded by members of a multi-gene family (Guerette et al., 1996; Gatesy et al., 2001; Hayashi et al., 2004; Garb and Hayashi, 2005; Garb et al., 2007; Blasingame et al., 2009; Garb et al., 2010; Starrett et al., 2012). Spidroins are large proteins (typically 200-350 kiloDaltons; Ayoub et al., 2007), consisting of a highly repetitive region flanked by non-repetitive amino and carboxy terminal regions. The different mechanical properties exhibited by silk fibers are largely attributed to the characteristics (e.g., amino acid sequence motifs and sub-repetitiveness) of the repetitive region of the protein, whereas the terminal regions have an important role in cell transport, protein aggregation, and protein monomer bonding (Hayashi et al., 1999; Motriuk-Smith et al., 2005; Ittah et al., 2006; Hagn et al., 2010; Gnesa et al., 2012).

Spidroin evolutionary studies have largely focused on ancient divergences among distantly related spiders and gene duplicates. These comparisons indicate that the internal repetitive and terminal regions of spidroin encoding genes evolve through very different mechanisms from each other (Gatesy et al., 2001; Garb et al., 2007; Garb et al., 2010; Starrett et al., 2012). Even across distantly related spidroins, terminal regions may be aligned and used to infer orthologous and paralogous relationships. However, internal repetitive regions, while usually similar across orthologs in repeat length and amino acid composition (Gatesy et al., 2001; Starrett et al., 2012), are often difficult to align across orthologs and generally impossible to align with confidence across paralogs. This challenge in determining positional (sequence) homology is hypothesized to be due, at least in part, to homogenization of repeats within a single spidroin gene via concerted evolution (Garb et al., 2007). Concerted evolution of repeats results in rapid divergence of repeats across species, which obscures inferences of positional homology (Elder and Turner, 1995). Because of this inability to confidently identify homologous regions across repeats, the selective factors acting on this region of the protein have been difficult to assess.

Most spidroin evolutionary studies have sampled across distantly related paralogous genes (e.g., Gatesy et al., 2001; Challis et al., 2006; Garb et al., 2007; Blasingame et al., 2009; Starrett et al., 2012), yet, a complete understanding of spidroin gene evolution requires determining the processes driving divergence across spidroin orthologs. Examination of spidroin evolution across orthologs in closely related taxa may allow for inferences of the nature and strength of selection acting not only on terminal regions, but the internal repetitive region as well.

The trapdoor spider genus *Aliatypus* (Mygalomorphae, Antrodiaetidae) provides a unique model system for studying spidroin evolution. *Aliatypus* consists of eleven species that are mostly restricted to California, with a single species in Arizona (Figure 2.1; Coyle 1974; Coyle and Icenogle, 1994; Hedin and Carlson, 2011). The different species of *Aliatypus* vary in the habitat types they occupy, ranging from forest to desert, with some species highly restricted to one habitat type in one region. For example, *A. gnomus* and *A. trophonius* are known only from redwood forests near the San Francisco Bay area

(Coyle, 1974; Satler et al., 2011). All *Aliatypus* are extremely poor dispersers and rely heavily on their silk lined burrows for protection. Hypotheses of phylogenetic relationships of *Aliatypus* species using morphological and molecular data are largely congruent, revealing *A. gulosus* as sister to all other species, two derived species groups (Erebus and Californicus groups), and three other species that fall sister to one of these two groups (Figure 2.2; Satler et al., 2011). Additionally, *A. gulosus* is characterized as expressing a single spidroin gene (Garb et al. 2007). Thus, silk orthologs can be sampled from all extant *Aliatypus* species and silk gene evolution can be compared among closely and distantly related species.

In this study, we obtain spidroin internal repetitive and carboxy terminal regions from all extant species of *Aliatypus*. These sequences are easily aligned without ambiguity, and we generate phylogenetic trees for spidroin repeat and carboxy terminal regions and compare them to hypothesized species relationships. Additionally, we examine the molecular evolutionary dynamics of repetitive and carboxy terminal regions across species. Our results allow for an assessment of the evolutionary forces that have shaped silk genes across different levels of species divergence.

Materials and Methods

Sampling

We sampled one to two individuals for each species of *Aliatypus*. Collection localities for each individual are shown in Figure 2.1 and locality information is given in

Table 2.1. Specimen voucher codes correspond to those used in Satler et al. (2011). Genomic DNA was extracted using a DNeasy Blood and Tissue kit (Qiagen, Valencia, CA), and two regions of the spidroin gene were amplified by PCR from each individual. Primers were designed from the *Aliatypus gulosus* spidroin cDNA sequence (GenBank accession EU117159), which covers part of the internal repetitive region through the carboxy terminal region. Forward primer 5'-GTGGAAAGAGTGACATCGGGAG-3' and reverse primer GTTGGGCAAAGGACTGAGCGAATG or GGACTGAGCGAATGCTCTAG were used to amplify the terminal repeat plus carboxy terminal (TR+C) encoding region. PCR amplification was conducted with recombinant *Taq* DNA polymerase (Invitrogen, Carlsbad, CA) using standard cycling conditions with an annealing temperature of 62˚C for most individuals. Some samples required lower annealing temperatures (50-58˚C). The second PCR reaction used forward primer GTTGCCTACGCCTTTGCTTACGC or GTTGCTTACGCCTTTGCATACGC and reverse primer GAGAATTGGCGTTAGCAGAGTTC to simultaneously amplify multiple internal repeats (IR) within the spidroin gene. This method of repeat sampling is akin to the simultaneous amplification and subsequent sequencing of multi-copy nuclear ribosomal genes (Hillis and Dixon, 1991), rather than a comprehensive sampling of all repeats in a spidroin gene, which is currently not possible to do for *Aliatypus*. An annealing temperature of 52˚C was used for most individuals, but some individuals required higher annealing temperatures (60-64˚C). PCR products were sequenced in both directions.

Sequence chromatograms were edited in Sequencher v. 4.5 (Ann Arbor, Michigan). Sites with multiple peaks in the chromatogram were coded as ambiguous following IUPAC designations. Sequences were translated and the amino acid sequences were aligned using MUSCLE under default parameters (Edgar, 2004). The sequences contained no introns. The DNA sequences were then aligned according to the amino acid alignment with PAL2NAL (Suyama et al., 2006). For the TR+C sequence, the outgroup sequence *Antrodiaetus fib2* was included in the 750 bp alignment; however, the 519 bp repetitive region aligned poorly with the *Aliatypus* IRs and therefore was not included in the final repeat dataset. The Difference of Sums Squares method, as implemented in TOPALi v2.5 (McGuire and Wright, 2000; Milne et al., 2009), was used to test for recombination points within both datasets. Window sizes of 375 and 250 base pairs (bp) were used for the TR+C and IR data sets, respectively. Recombination points were not detected in either dataset. K2P distances were calculated using PAUP* 4.0b8-b10 (Swofford, 2002).

Phylogenetic Analyses

Maximum likelihood (ML) analyses were conducted on TR+C and IR data sets through the CIPRES web server (Miller et al., 2010). Maximum Likelihood (ML) searches for the best tree and bootstrap (BS) were performed simultaneously with 1000 replicates using RAxML v. 7.2.8 (Stamatakis, 2006a; Stamatakis, 2006b; Stamatakis et al., 2008). Analyses with codon position partitioning were performed using the GTR+γ model, following the RAxML manual.

Bayesian analyses were conducted using MrBayes v. 3.1.2 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). For the TR+C and IR data sets, DNA substitution models were determined for each codon position using MrModeltest v. 2.3 (Nylander, 2004). Bayesian analyses consisted of two simultaneous searches run for 10 million generations, with trees and parameters sampled from four MCMC chains every 1000 generations. Substitution rates for codon position partitions were unlinked and the rate multiplier was set to variable. Analyses were considered complete when the standard deviation of split frequencies between the two searches was below 0.01 (Ronquist et al., 2005). The first twenty-five percent of samples were treated as burnin and discarded. Bayesian posterior probabilities (PP) were used to assess clade support.

Amino Acid Composition

We determined the amino acid compositions of the spidroin repetitive region sequences with MacVector 7.2 (Accelrys Inc., San Diego, CA). Compositions based on the translated direct repeat sequence of *Aliatypus gulosus* (AP) generated from PCR with that of the translated cDNA sequence (GenBank accession EU117159; spiders collected from same locality) were very similar, with <1.5% difference for any amino acid.

Selection Tests

Analyses were conducted to measure selection acting on the IR and TR+C encoding *Aliatypus* sequences. Average pairwise rates of non-synonymous (dn) and synonymous (ds) substitutions were determined using the Nei-Gojobori method (Nei-

Gojobori, 1986) under the Jukes-Cantor model in *MEGA* v. 5 (Tamura et al., 2011). A ratio of dn/ds, or ω, <1, =1, >1, indicates evidence of purifying selection, neutrality, and positive selection, respectively. For the TR+C encoding region, tests of selection were also performed for individual sites under models M0 (one ratio), M1a (nearly neutral), M2a (positive selection), M7 (beta), M8a (beta & ω=1), and M8 (beta & ω>1) using the codeml module in PAML v. 4.4c (Yang, 2007). Bayes empirical Bayes was used to identify sites under positive selection for models M2a and M8 (Yang et al., 2005). Likelihood-Ratio-tests were used to determine which model best fit the data, with statistical significance evaluated using the χ^2 distribution.

Selection tests on amino acid properties were performed on the TR+C alignment and phylogenetic tree using TreeSAAP v. 3.2 (Woolley et al., 2003; McClellan and Ellison, 2010). We used 8 magnitude categories for amino acid changes, where magnitude category 1 changes in physicochemical properties were most conservative and magnitude category 8 changes were most radical. Thirty-one amino acid properties for proteins in solution were first evaluated using a window size of the length of the alignment. Amino acid properties identified as category 6 and above with p values ≤ 0.001 were further tested using a sliding window size of 10 sites with the sliding window incremented every single site. Regions identified to be under positive destabilizing selection (categories 6 and above; p≤ 0.001) were graphed and the non-synonymous substitutions were plotted on the TR+C tree.

Results and Discussion

Phylogenetic Analyses

Inferred relationships among orthologous spidroins are often ambiguous or weakly-supported, likely due to problematic alignments across distantly related paralogs and the limited sequence length of non-repetitive regions available for phylogenetic reconstruction (Gatesy et al., 2001; Garb, Ayoub, and Hayashi, 2010; Starrett et al., 2012). In contrast, phylogenetic analyses of the IR and TR+C data sets recovered many definitive relationships for spidroin orthologs that are largely congruent with the hypothesized *Aliatypus* species tree based on a multi-gene analysis (Figures 2.2, 2.3; Satler et al., 2011). This indicates that the spidroin sequences we have obtained are representatives of a single ortholog group, consistent with the expression of a single spidroin in a silk gland cDNA study of *A. gulosus* (Garb et al., 2007). For the TR+C data set, *A. gulosus* was found to be sister to all other *Aliatypus* species (when rooted with *Antrodiaetus fib2*), consistent with trees based on morphology and a multi-gene dataset (Coyle, 1994; Satler et al., 2011). Additionally, both spidroin regions recovered monophyletic Erebus and Californicus species groups (Figures 2.2, 2.3). The phylogenetic placements of *A. thompsoni*, *A. coylei*, and *A. aquilonius* were sister to the Erebus group, Californicus group, or Erebus + Californicus group, as occurs in all morphological and molecular based analyses (Coyle, 1994; Satler et al., 2011). However, beyond these more basal placements, the relationships of *A. thompsoni*, *A. coylei*, and *A. aquilonius* had low support, consistent with the considerable variability in phylogenetic placements based on mitochondrial and non-spidroin nuclear gene data sets (Satler et al., 2011).

The Erebus and Californicus species groups, which are derived members of *Aliatypus*, exhibit contrasting patterns in phylogenetic relationships between spidroin regions. For both the IR and TR+C data sets, relationships within the Erebus group were identical and consistent with hypothesized species relationships (Figures 2.2, 2.3). However, relationships within the Californicus group differed between the IR and TR+C regions. For the TR+C data set, relationships among Californicus group species were identical to those based on coalescent analysis of the multi-gene data set in Satler et al. (2011). In contrast, for the IR region, relationships deviated in that *A. janus* was the most basal Californicus member, and *A. californicus* was found to be polyphyletic with high support, with one *A. californicus* individual sister to *A. gnomus* and the other sister to *A. isolatus*. The basal placement of *A. janus* and polyphyly of *A. californicus* may be a result of the rapid speciation and high population subdivision in the Californicus group, as exhibited by high intraspecific/interspecific ratios for cytochrome oxidase I sequence divergence (Satler et al., 2011). *Aliatypus californicus* individual 1 and individual 2 IR sequences are ambiguous at 1.20% and 2.79% of IR sites, respectively, but the two *A. californicus* IR sequences differ by only 2.85%. The non-monophyly and high intraspecific/interspecific ratios of IR sequence divergence indicate that insufficient time likely has passed for repeats to show a pattern of complete within species homogenization. This pattern is similar to that seen for tubuliform (egg case) spidroins based on an analysis of distantly and closely related species of orbicularian spiders (Garb and Hayashi, 2005). Despite the importance of spidroin repetitive regions to fiber mechanical properties and function, the factors that lead to replacement of one repeat type for another throughout a spidroin in a species has yet to be investigated. IR regions from the two *A. californicus* representatives show that divergent repeats may be retained in geographically distant populations of a species. Thus, it would be worthwhile to sample repeats densely throughout the range of a species. This could allow for observation of intermediate stages in the homogenization process that are missed by sampling spidroins across paralogs and distantly related spider taxa.

Internal Repeat Conservation

Comparison of *Aliatypus* IR sequences suggests that a combination of purifying selection and concerted evolution have acted to conserve this region. Studies focusing on higher-level sampling of spidroins across paralogs indicate that IR regions have changed dramatically (e.g., Gatesy et al., 2001). However, spidroin orthologs generally maintain similar overall amino acid compositions across species (Gatesy et al., 2001; Starrett et al., 2012). Here, we sampled spidroins from all extant species of *Aliatypus* and found that the IR regions exhibit considerable conservation both in amino acid composition and sequence (Figures 2.4, 2.5). Alanine, serine, glycine, threonine, and tyrosine were the five most prevalent amino acids in most (11 of 15) repeat sequences. Combined, alanine and serine make up approximately 60% of each total composition (Figure 2.4). Except for alanine and serine, no other amino acid exceeds 8% of the total composition. In the amino acid alignment of IR regions across species, only six out of 145 sites (5, 50, 53, 57, 84, and 136) show less than 60% conservation (Figure 2.5), consistent with strong selection to maintain amino acid sequence.

Many sites are conserved across *Aliatypus* species not only at the amino acid level, but also at the codon level. The overall conservation in the IR region is in part due to purifying selection, as indicated by the average pairwise ω value of 0.36. Average pairwise K2P distances for the three IR codon positions (position 1: 0.067, position 2: 0.045, position 3: 0.137) were lower than the corresponding values for the TR+C codon positions (position 1: 0.155, position 2: 0.127, position 3: 0.177). K2P values for the IR third codon positions, which are mostly silent substitution sites, are much lower than K2P values of the TR+C third codon positions and nearly as low as those of the TR+C second codon positions, which are all replacement substitution sites. Purifying selection alone is an unlikely explanation for the low K2P values in the IR region in comparison to the TR+C region, particularly for third position sites, where mutations usually do not change the encoded protein. Concerted evolution via unequal crossing over is expected to occur more frequently in internal repetitive regions than near the terminal regions (Hayashi and Lewis, 2000; Hayashi, 2002; Garb and Hayashi, 2005), and can explain why codons are more conserved at both non-synonymous and synonymous sites in the IR region than the TR+C region. Thus, our comparison of homologous regions of the IR region across *Aliatypus* orthologs suggests that purifying selection, in addition to concerted evolution, have had an important role in maintaining spidroin sequence conservation. Concerted evolution likely reduced much of the variation that arose in the repeat sequence, and purifying selection further acted to reinforce this by eliminating most non-synonymous mutations to prevent changes that may have disrupted the protein.

Selection and Divergence in the Terminal Repeat plus Carboxy Terminus Region

In contrast to the IR region, the TR+C spidroin region has evolved rapidly across *Aliatypus* species. The average pairwise ω value for the TR+C region is 0.72. Thus, averaged across all sites, the TR+C area has evolved mostly under weak purifying selection. Likelihood ratio tests comparing individual site models determined that models allowing positive selection fit the data significantly better than models only allowing purifying selection and neutrality (Table 2.2). Bayes empirical Bayes analyses under models M2a and M8 identified 2 and 5 sites, respectively, as having evolved under positive selection with high posterior probabilities (>0.95). All sites estimated to be under positive selection are located in the terminal repeat region, which may indicate that this region has importance for spidroin adaptation to the unique environments of each *Aliatypus* species.

In addition to a high rate of non-synonymous change in the TR+C region, we found evidence that selection has driven amino acid substitutions that result in large physicochemical change. TreeSAAP analysis of the TR+C data set identified three regions that may have evolved under positive destabilizing selection for the physicochemical property, alpha helical tendency (P_α). Sliding window analyses identified two regions as magnitude category 6 changes and one region as a magnitude category 8 change ($p \leq 0.001$; magnitude range is from 1 to 8, with 8 being the greatest physicochemical change). The two magnitude 6 regions are located in the terminal repeat region at AA sites 43-57 and 88-104 and result from non-synonymous substitutions that occurred at 12 and 13 codon sites, respectively (Figure 2.6; Table 2.3). The magnitude 8 region is located in the C-terminal region (AA sites 189-204) and resulted from non-synonymous substitutions at 8 sites. Thus, although no sites were detected as having evolved under positive selection (i.e., ω>1) in the C-terminal region, accounting for physicochemical change suggests that non-synonymous substitutions in this region also may have adaptive significance.

Over half (35 of 61) of the non-synonymous changes that occurred across the tree in the identified regions under positive destabilizing selection were estimated to have resulted in large physicochemical changes for P_α (i.e., magnitude categories 6-8; Figure 2.7; Table 2.3). A majority (41/61) of the amino acid changes that occurred resulted in a negative P_α, indicating that selection may have favored changes that reduced alpha helix formation in these regions for many taxa. A dramatic shift in negative P_α occurs on the branch leading to the common ancestor of the Erebus and Californicus species groups with six negative P_α changes (4 of magnitude 6, 1 of 7, and 1 of 8) to only a single positive P_α change (magnitude 6). However, in some cases where multiple non-synonymous changes occur along a single branch, the numbers of negative and positive P_α amino acid substitutions are roughly equal (e.g., node a to b, node c to d, node f to g; Figure 2.7). One striking example of this occurs from node e to the *A. coylei* terminal node, where ten non-synonymous changes occur. Five of these changes have positive P_α (4 of magnitude 6 and 1 of 8) and the other five have negative P_α (1 of magnitude 3, 2 of

6, and 2 of 8). Alpha helices in the C-terminal region have been implicated in dimerization among spidroin monomers in *Araneus diadematus*, which affects protein aggregation and fiber formation (Hagn et al., 2010). The balance in positive and negative P_α amino acid substitutions suggests that changes at some sites have compensated for changes at other sites that were disruptive to alpha helix formation.

Silk gene research has been overwhelmingly focused on comparisons across distantly related spider taxa and deeply divergent paralogs. Thus, it has been difficult to infer how individual silk genes have been influenced by evolution. Phylogenetic analyses and tests of selection reveal that *Aliatypus* spidroin genes have evolved in a mosaic fashion, with incidents of incongruence between IR and TR+C gene regions (Figure 2.3). This pattern may be particularly true for low vagility species, where local adaptation and absence of gene flow could have contributed to genetic differentiation in IR regions in geographically separated populations. Further sampling of spidroin genes at the interface of population and species divergence should allow for inferences of what factors shaped repeat evolution. The selective regime for spidroin repeat regions has been previously difficult to determine as positional homology can be impossible to establish among spidroins. By comparing IR regions across congeneric species, we were able to show that purifying selection, in addition to concerted evolution, has been responsible for the considerable stasis in sequence evolution. In addition, while the C-terminal region is typically considered the most conserved part of a spidroin, we found high rates of non-synonymous substitutions and amino acid replacements resulting in large physicochemical changes. These results indicate that selection has driven divergence in the TR+C region, and may have been important for adaptation to the range of habitats occupied by the different *Aliatypus* species. By utilizing a comparative approach, we have shown that repetitive and non-repetitive regions of spidroin genes have been shaped by combinations of evolutionary forces. These evolutionary dynamics may also apply to other genes with repetitive architectures and functionally distinct regions, such as genes encoding insect fibroins, collagens, and plaque forming amyloids.

References

Ayoub NA, Garb JE, Tinghitella R, Collin MA, Hayashi CY (2007) Blueprint for a High-Performance Biomaterial: Full-Length Spider Dragline Silk Genes. PLoS ONE. 6: e514.

Blasingame E, Tuton-Blasingame T, Larkin L, Falick AM, Zhao L, Fong J, Vaidyanathan V, Visperas A, Geurts P, Hu X, La Mattina C, Vierra C (2009) Pyriform Spidroin 1, a Novel Member of the Silk Gene Family That Anchors Dragline Silk Fibers in Attachment Discs of the Black Widow Spider, *Latrodectus hesperus*. *Journal of Biological Chemistry*. 284(42): 29097-29108.

Challis RJ, Goodacre SL, Hewitt GM. (2006). Evolution of spider silks: conservation and diversification of the C-terminus. Insect Molecular Biology. 15(1):45-56.

Coyle FA (1974) Systematics of the trapdoor spider genus *Aliatypus* (Araneae: Antrodiaetidae). Psyche 81: 431–500.

Coyle FA (1994) Cladistic analysis of the species of the trapdoor spider genus *Aliatypus* (Araneae, Antrodiaetidae). The Journal of Arachnology 22: 218–224.

Coyle FA, Icenogle WR (1994) Natural history of the Californian trapdoor spider genus *Aliatypus* (Araneae, Antrodiaetidae). The Journal of Arachnology 22: 225–255.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5): 1792-1797.

Elder Jr. JF, Turner BJ (1995) Concerted Evolution of Repetitive DNA Sequences in Eukaryotes. Quarterly Review of Biology. 70(3):297-320.

Garb JE, Ayoub NA, Hayashi CY (2010) Untangling spider silk evolution with spidroin terminal domains. *BMC Evolutionary Biology*. 10: 243.

Garb JE, DiMauro T, Lewis RV, Hayashi CY (2007) Expansion and Intragenic Homogenization of Spider Silk Genes since the Triassic: Evidence from Mygalomorphae (Tarantulas and Their Kin) Spidroins. *Mol. Biol. Evol.* 24(11): 2454-2464.

Garb JE, Hayashi CY (2005) Modular evolution of egg case silk genes across orb-weaving spider superfamilies. Proc. Nat. Acad. Sci. U.S.A. 102(32): 11379-11384.

Gatesy J, Hayashi CY, Motriuk D, Woods J, Lewis RV (2001) Extreme Diversity, Conservation, and Convergence of Spider Silk Fibroin Sequences. Science. 291: 2603-2605.2.

Gnesa E, Hsia Y, Yarger JL, Weber W, Lin-Cereghino J, Lin-Cereghino G, Tang S, Agari K, Vierra C (2012) Conserved C-terminal domain of spider tubuliform spidroin 1 contributes to extensibility in synthetic fibers. Biomacromolecules. 13(2):304-312.

Guerette PA, Ginzinger DG, Weber BH, Gosline JM (1996) Silk properties determined by gland-specific expression of a spider fibroin gene family. Science. 272: 112-115.

Hagn F, Eisoldt L, Hardy JG, Vendrely C, Coles M, Scheibel T, Kessler H (2010) A conserved spider silk domain acts as a molecular switch that controls fiber assembly. Nature. 465:239-242.

Hayashi CY (2002) Evolution of spider silk proteins: insight from phylogenetic analyses. In: DeSalle R, Giribet G, Wheeler W (Eds) Molecular Systematics and Evolution: Theory and Practice. Birkhäuser Verlag, Basel Switzerland. 209-224.

Hayashi CY, Blackledge TA, Lewis RV (2004) Molecular and Mechanical Characterization of Aciniform Silk: Uniformity of Iterated Sequence Modules in a Novel Member of the Spider Silk Fibroin Gene Family. *Mol. Biol. Evol.* 21(10): 1950-1959.

Hayashi CY, Lewis RV (2000) Molecular Architecture and Evolution of a Modular Spider Silk Protein Gene. Science. 287: 1477-1479.

Hayashi CY, Shipley NH, Lewis RV (1999) Hypotheses that correlate the sequence, structure, and mechanical properties of spider silk proteins.International Journal of Biological Macromolecules. 24:271-275.

Hedin H, Carlson D (2011) A new trapdoor spider species from the southern Coast Ranges of California (Mygalomorphae, Antrodiaetidae, Aliatypus coylei, sp. nov,), including consideration of mitochondrial phylogeographic structuring. Zootaxa 2963: 55–68.

Hillis DM, Dixon MT (1991) Ribosomal DNA: Molecular Evolution and Phylogenetic Inference. Quarterly Review of Biology. 66(4):411-453.

Hinman MB, Lewis RV (1992) Isolation of a Clone Encoding a Second Dragline Silk Fibroin. 1992. Journal of Biological Chemistry. 267(27): 19320-19324.

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 17: 754-755.

Ittah S, Cohen S, Garty S, Cohn D, Gat U (2006) An Essential Role for the C-Terminal Domain of A Dragline Spider Silk Protein in Directing Fiber Formation. Biomacromolecules. 7:1790-1795.

McClellan DA, Ellison DA (2010) Assessing and improving the accuracy of detecting protein adaptation with the TreeSAAP analytical software. Int. J. Bioinform. Res. Appl. 6(2):120-133.

McGuire G, Wright F (2000) TOPAL 2.0: improved detection of mosaic sequences within multiple alignments. *Bioinformatics*, 16, 130–134.

Miller MA, Pfeiffer W, Schwartz T (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Proceedings of the Gateway Computing Environments Workshop (GCE), 14 Nov. 2010, New Orleans, LA pp 1 - 8.

Milne I, Lindner D, Bayer M, Husmeier D, McGuire G, et al. (2009) TOPALi v2: a rich graphical interface for evolutionary analyses of multiple alignments on HPC clusters and multi-core desktops. Bioinformatics 25: 126–127.

Motriuk-Smith D, Smith A, Hayashi CY, Lewis RV (2005) Analysis of the Conserved N-Terminal Domains in Major Ampullate Spider Silk Proteins. Biomacromolecules. 6:3152-3159.

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular Biology and Evolution. 3:418-426.

Nylander JAA (2004) MrModeltest v2. Program distributed by the author. Evolutionary Biology Centre, Uppsala University.

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19: 1572-1574.

Ronquist F, Huelsenbeck JP, van der Mark P (2005) mrbayes 3.1 manual, draft 5/26/2005, online at http://mrbayes.csit.fsu.edu/manual.php. Accessed October 14, 2010.

Satler JD, Starrett J, Hayashi CY, Hedin M (2011) Inferring Species Trees from Gene Trees in a Radiation of California Trapdoor Spiders (Araneae, Antrodiaetidae, *Aliatypus*). PLoS ONE 6(9): e25355. doi:10.1371/journal.pone.0025355.

Stamatakis A (2006a) Phylogenetic Models of Rate Heterogeneity: A High Performance Computing Perspective. In Proceedings of 20th IEEE/ACM International Parallel and Distributed Processing Symposium (IPDPS2006), High Performance Computational Biology Workshop, Rhodos, Greece.

Stamatakis A (2006b) RAxML-VI-HPC: Maximum Likelihood-based Phylogenetic Analyses with Thousands of Taxa and Mixed Models. Bioinformatics. 22(21): 2688–2690.

Stamatakis A, Hoover P, Rougemont J (2008) A Rapid Bootstrap Algorithm for the RAxML Web Servers. Syst. Biol. 57(5): 758-771.

Starrett J, Garb JE, Kuelbs A, Azubuike UO, Hayashi CY (2012) Early Events in the Evolution of Spider Silk Genes. PLoS ONE 7(6): e38084. doi:10.1371/journal.pone.0038084.

Suyama M, Torrents D, Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34: W609-W612.

Swofford DL (2002) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates; Sunderland, Massachusetts.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, and Kumar S (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol. Biol. Evol. 28(10):2731-2739.

Woolley S, Johnson J, Smith MJ, Crandall KA, McClellan DA (2003) TreeSAAP: Selection on Amino Acid Properties using phylogenetic trees. Bioinformatics. 19(5):671-672.

Xu M, Lewis RV (1990) Structure of a protein superfiber: Spider dragline silk. Proc. Nat. Acad. Sci. U.S.A. 87: 7120-7124.

Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution 24: 1586-1591.

Yang Z, Wong WSW, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. 22:1107-1118.

Tables

Table 2.1. *Aliatypus* species list with collection locality information and GenBank numbers for spidroin sequences.

Species	Location	Specimen #	TR+C	IR
Antrodiaetus riversi **(O. P.-Cambridge, 1883)**	CA: El Dorado Co., Mosquito Rd. 38.8199, -120.6726	-		
Aliatypus gulosus **Coyle, 1974**	CA: Riverside Co., Riverside 33.9826, -117.3031	AP (i1)		
	CA: Los Angeles Co., Baldwin Hills N34.0094, -118.3696	MY2588 (i2)		
Aliatypus thompsoni **Coyle, 1974**	CA: Kern Co., Lake Isabella 35.6358, -118.4963	MY480		
Aliatypus coylei **Hedin & Carlson, 2011**	CA: Monterey Co., Palo Colorado Rd. 36.3997, -121.8915	MY3058		
Aliatypus plutonis **Coyle, 1974**	CA: Riverside Co., DeLuz Rd. 33.5084, -117.2311	MY469 (i1)		
	CA: San Diego Co., DeLuz Rd., N of Fallbrook 33.4111, -117.2889	MY964 (i2)		
Aliatypus torridus **Coyle, 1974**	CA: Santa Barbara Co., near Gaviota 34.5103, -120.2352	MY992		
Aliatypus erebus **Coyle, 1974**	CA: Santa Clara Co., Bear Creek Rd. 37.1888, -121.9954	MY822		
Aliatypus trophonius **Coyle, 1974**	CA: Santa Cruz Co., N Soquel 37.0383, -121.9446	MY994		
Aliatypus aquilonius **Coyle, 1974**	CA: Humboldt Co., W Redway 40.1223, -123.8419	MY388		
Aliatypus gnomus **Coyle, 1974**	CA: Santa Cruz Co., Henry Colwell 37.0176, -122.0639	MY380		
Aliatypus californicus **(Banks, 1896)**	CA: Lake Co., Bartlett Springs Rd. 39.1671, -122.5070	MY783 (i1)		
	CA: Santa Clara Co., Bear Creek Rd. 37.1888, -121.9954	MY 821 (i2)		
Aliatypus isolatus **Coyle, 1974**	AZ: Coconino Co., Oak Creek 34.9896, -111.7487	MY101		
Aliatypus janus **Coyle, 1974**	CA: Fresno Co., N Pinehurst 36.7061, -118.9981	GMY11		

Table 2.2. Selection test results for the terminal repeat plus carboxy terminal encoding region from codeml analyses (Yang, 2004).

Model	*L*	Average ω	Proportion of sites ω >1	Average ω for proportion of ω>1	BEB sites ω>1 (p>0.95); Site# (ω and SE)
M0	-3083.787985	0.76			
M1a	-3057.572966	0.66			
M2a	-3046.849767	1.02	0.13	3.61	62 (3.47 +- 1.14) 77 (3.44 +- 1.18)
M7	-3057.815033	0.71			
M8	-3046.912634	1.01	0.16	3.33	3 (2.81 +- 0.74) 41 (2.85 +- 0.70) 45 (2.83 +- 0.73) 62 (2.88 +- 0.67) 77 (2.86 +- 0.71)
M8a	-3056.99861	0.67			
Model Comparison	2 *L*	df	P value		
M1a v M2a	21.446	2	<0.001		
M7 v M8	21.805	2	<0.001		
M8a v M8	20.172	1	<0.001		

Table 2.3. TreeSAAP selection test results for the physicochemical property, alpha helical tendency (P_α), in the terminal repeat plus carboxy terminal (TR+C) encoding region. Node labels and codon numbers correspond to those in Figure 2.7.

Codon#	Branch	From Codon	To Codon	From AA	To AA	P_α	Amount
TR							
43	node a --> node c	GCC	GTC	Ala	Val	4	-0.36
43	node d --> node e	GTC	GCC	Val	Ala	4	0.36
43	node e --> coylei	GCC	TCC	Ala	Ser	6	-0.65
43	node i --> torridus	GCC	GTC	Ala	Val	4	-0.36
44	node m --> californicus_i2	GCG	GTG	Ala	Val	4	-0.36
45	node e --> coylei	CAG	TCA	Gln	Ser	3	-0.34
45	node l --> isolatus	CAG	CAT	Gln	His	1	-0.11
45	node h --> trophonius	CAG	CCG	Gln	Pro	5	-0.54
46	node f --> node g	GCC	TCC	Ala	Ser	6	-0.65
46	node i --> torridus	TCC	GCC	Ser	Ala	6	0.65
47	node d --> node e	TCA	GCA	Ser	Ala	6	0.65
47	node i --> torridus	GCA	TCA	Ala	Ser	6	-0.65
48	node c --> thompsoni	GCA	ACA	Ala	Thr	6	-0.59
48	node e --> coylei	GCA	GGA	Ala	Gly	8	-0.85
48	node n --> gnomus	GCT	TCT	Ala	Ser	6	-0.65
48	node g --> node i	GCT	TCT	Ala	Ser	6	-0.65
49	node b --> gulosus_i2	TCT	TTT	Ser	Phe	4	0.36
49	node e --> coylei	TCC	GCT	Ser	Ala	6	0.65
50	node e --> node f	GCA	ACA	Ala	Thr	6	-0.59
50	node f --> node k	ACA	TCA	Thr	Ser	1	-0.06
51	node c --> node d	GCA	TCA	Ala	Ser	6	-0.65
51	node e --> node f	TCA	GCA	Ser	Ala	6	0.65
51	node h --> erebus	GCA	GTA	Ala	Val	4	-0.36
52	node c --> node d	TCC	GCC	Ser	Ala	6	0.65
52	node e --> node f	GCC	TCT	Ala	Ser	6	-0.65
53	node e --> coylei	GCT	ACA	Ala	Thr	6	-0.59
56	node e --> node f	GCA	TCA	Ala	Ser	6	-0.65
88	node f --> node g	TCC	ACC	Ser	Thr	1	0.06
88	node h --> erebus	ACC	AAC	Thr	Asn	2	-0.16
89	node d --> node e	ACG	TCG	Thr	Ser	1	-0.06
89	node e --> coylei	TCG	GCC	Ser	Ala	6	0.65
90	node d --> aquilonius	GCC	ACC	Ala	Thr	6	-0.59
90	node e --> node f	GCC	GGC	Ala	Gly	8	-0.85
90	node f --> node g	GGC	GTC	Gly	Val	5	0.49
91	node e --> node f	GAA	AGC	Glu	Ser	7	-0.74
92	node d --> aquilonius	ACC	AGC	Thr	Ser	1	-0.06
92	node d --> node e	ACC	TCG	Thr	Ser	1	-0.06
92	node e --> coylei	TCG	GCT	Ser	Ala	6	0.65
93	node e --> coylei	ACT	GCT	Thr	Ala	6	0.59
95	node l --> janus	GCG	ACG	Ala	Thr	6	-0.59
95	node m --> californicus_i2	GCG	TCG	Ala	Ser	6	-0.65
95	node f --> node g	GCG	ACG	Ala	Thr	6	-0.59
96	node d --> node e	TCT	AAA	Ser	Lys	4	0.39
97	node d --> node e	GCC	TCA	Ala	Ser	6	-0.65

98	node d --> node e	GCC	AGC	Ala	Ser	6	-0.65
99	node d --> node e	ACT	AGT	Thr	Ser	1	-0.06
100	node e --> node f	GCC	TCC	Ala	Ser	6	-0.65
103	node d --> node e	GCT	ACT	Ala	Thr	6	-0.59
103	node f --> node g	ACT	GCT	Thr	Ala	6	0.59
103	node g --> node h	GCT	GAA	Ala	Glu	1	0.09
C							
189	node b --> gulosus_i1	TCC	CCC	Ser	Pro	2	-0.2
190	node a --> node b	TTA	GAA	Leu	Glu	3	0.3
191	node a --> node b	CTT	TTT	Leu	Phe	1	-0.08
192	node a --> node b	TCC	ATT	Ser	Ile	3	0.31
193	node a --> node b	ACT	TCT	Thr	Ser	1	-0.06
194	node a --> node b	AGT	ATT	Ser	Ile	3	0.31
195	node b --> gulosus_i2	GGA	GAA	Gly	Glu	8	0.94
195	node e --> coylei	GGA	GAA	Gly	Glu	8	0.94
197	node a --> node b	GCC	GAC	Ala	Asp	4	-0.41
197	node c --> thompsoni	GCC	CCC	Ala	Pro	8	-0.85
197	node e --> coylei	GCC	GGC	Ala	Gly	8	-0.85

Figures

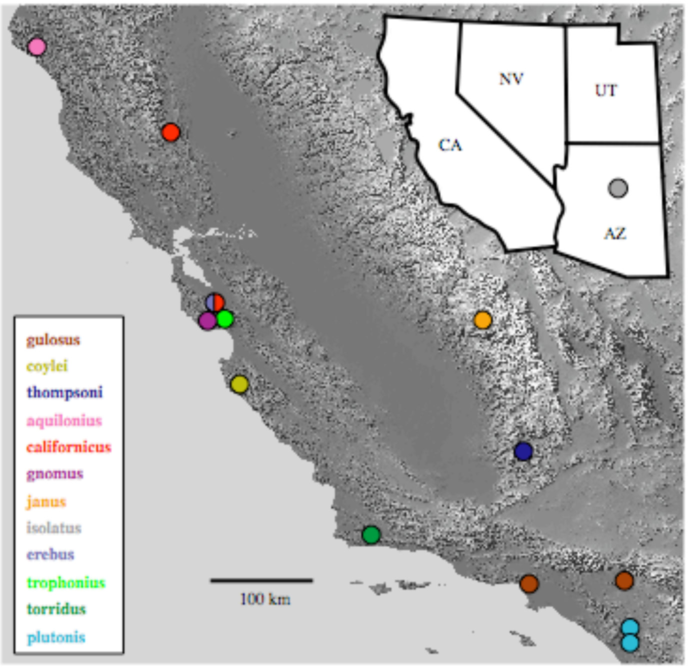


Figure 2.1. Collection localities for *Aliatypus* samples.

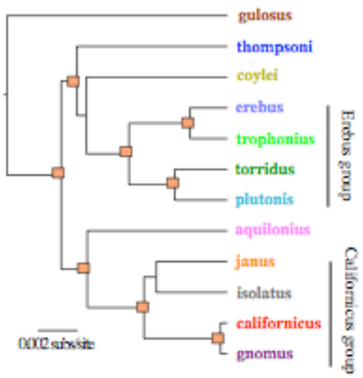


Figure 2.2. *Aliatypus* species tree based on a multi-gene, multi-species coalescent analysis (Satler et al., 2011). Nodes with boxes indicate relationships consistent between phylogenetic analyses using both coalescent and concatenated methods.

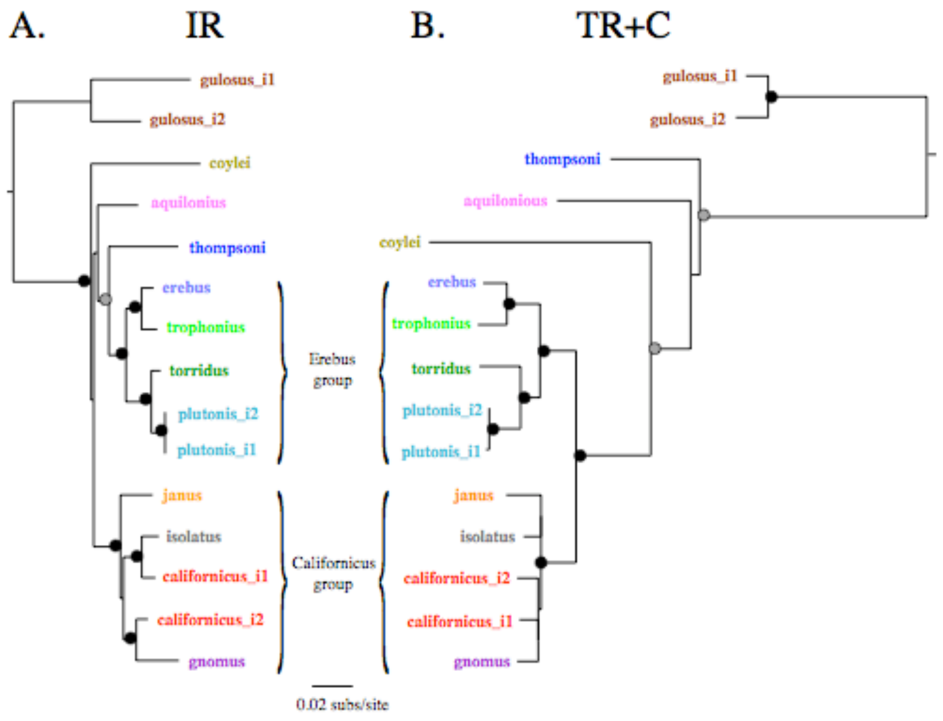


Figure 2.3. Phylogenetic trees based on maximum likelihood analysis of the (A) internal repeat (IR) encoding region, and (B) the terminal repeat plus carboxy terminal (TR+C) encoding region. Black circles at nodes indicate significant support values using both maximum likelihood bootstrap (>70) and Bayesian posterior probabilities (>0.95). Grey circles indicate significant support in only one type of analysis. For both the IR and TR+C data sets, the phylogenetic trees resulting from ML and Bayesian analyses did not conflict, with the ML tree being more resolved. The IR tree was rooted with *Aliatypus gulosus* sequences as no outgroup sequence could be aligned within reason (see methods). The TR+C tree was rooted with *Antrodiaetus fib2* (not shown). For the IR data set, MrModeltest determined that models F81+γ, GTR, and HKY best fit codon positions one, two, and three, respectively. For the TR+C data set, model HKY+ γ was selected for codon positions one and three, and GTR+I was selected for codon position two.

.

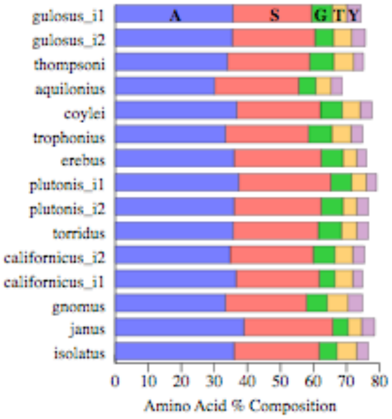


Figure 2.4. Amino acid percent compositions based on conceptual translation of the internal repetitive (IR) region. Percent compositions for the five most common amino acids in most IR sequences are shown: alanine (A), serine (S), glycine (G), threonine (T), tyrosine (Y).

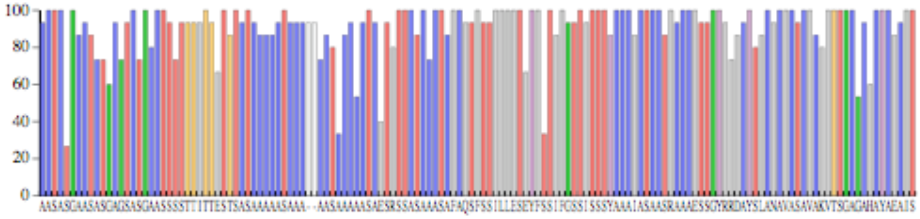


Figure 2.5. Percent identities (shown on y axis) of the internal repetitive (IR) region amino acids by position in a multiple sequence alignment, using. the *Aliatypus plutonis* IR as the reference (shown on x-axis). Color of bars match those in Figure 2.4 for the five most common amino acids.

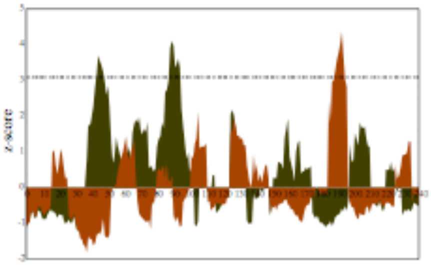


Figure 2.6. TreeSAAP (Woolley et al., 2003) sliding window results for alpha helical tendency in the terminal repeat plus carboxy terminal (TR+C) region. Regions above the z-score of 3.09 (dashed line) were significantly different than neutrality. Green and brown indicate results for magnitude categories 6 and 8, respectively.

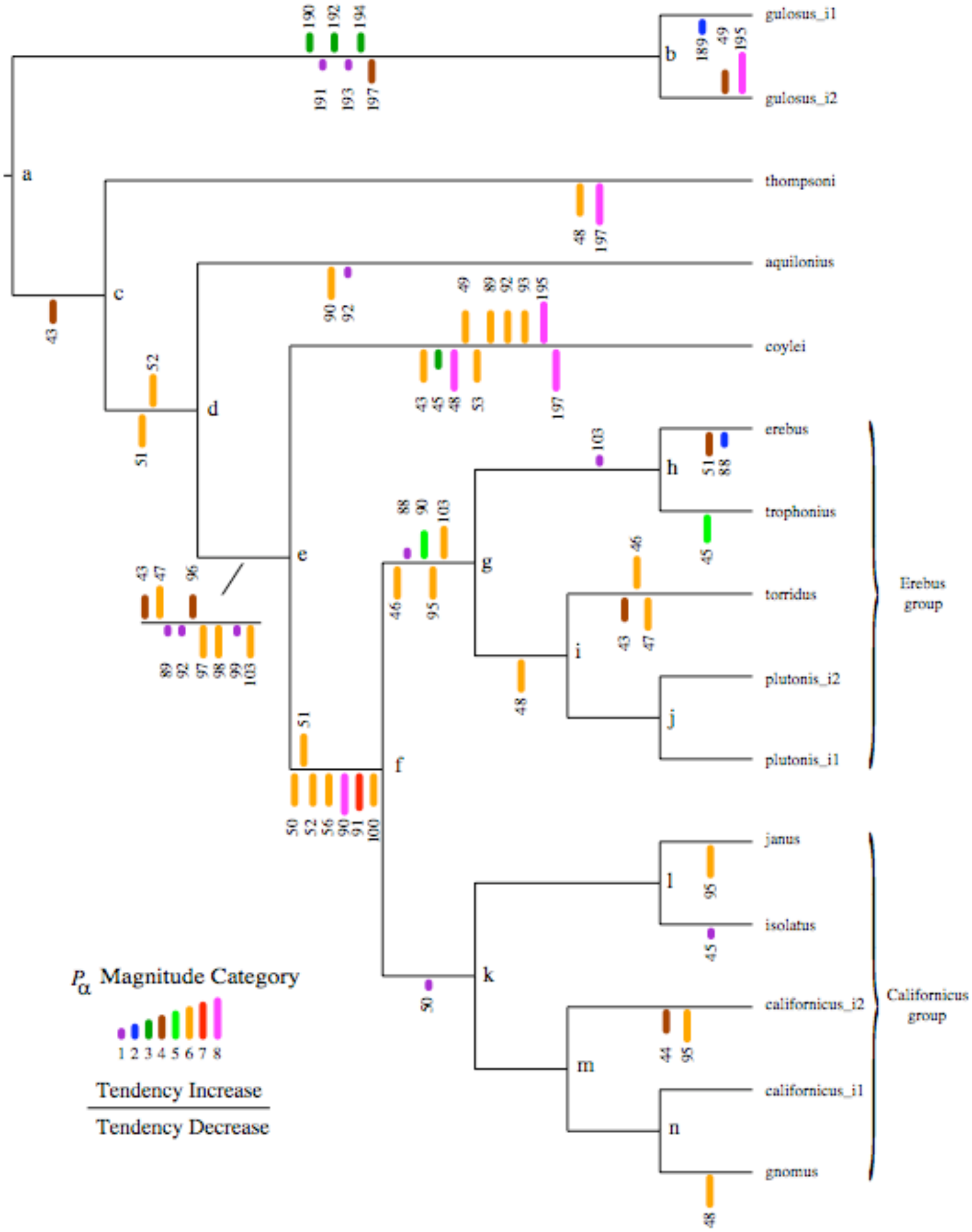


Figure 2.7. Mapping of amino acid changes that occurred in the regions identified as evolving under positive destabilizing selection by TreeSAAP (see Figure 2.6 and Table 2.3; Woolley et al., 2003). The magnitude category change of alpha helical tendency (P_α) and site number are indicated, with positive tendency changes above branches and negative tendency changes below branches. Phylogeny is based on the maximum likelihood analysis of the terminal repeat plus carboxy terminal encoding region. Node labels refer to Table 2.3.

Chapter 3

Characterization of Genes Associated With Abdominal and Tarsal Silk Production in Tarantulas

Abstract

Silk fiber production has evolved independently in many arthropod orders. Although silk proteins in these different taxa evolved independently, many share the characteristics of repetitiveness and richness in few amino acids, namely glycine, alanine, and serine. In spiders, multiple silk protein encoding gene families have been discovered, suggesting independent evolution of silk genes within a taxonomic group. Spiders are well known for their silks produced from abdominal glands, and the genes encoding silk in these glands have been characterized from a variety of spiders. However, the recent discovery of silk-like secretions from the tarsi of tarantulas indicates that spiders posses even greater silk gene diversity than currently understood. Alternatively, known silk genes may be expressed in tissues other than abdominal glands. Here, I sequenced transcripts from the tarsi and abdominal silk glands of two tarantula species and used blastx searches to identify transcripts homologous to silk genes or other structural proteins. Additionally, digital gene expression of multiple tissue types was used to identify candidates for tarsal silk secretions as well as novel silk associated genes from the abdominal glands. I found high expression of members of the spidroin gene family and putative silk-like genes in the abdominal silk glands of both tarantulas. While spidroins were highly expressed in silk glands, significant expression of these genes was not found in the tarsi. Instead, each tarantula species expressed a different transcript with silk-like characteristics in their tarsi. My findings support spidroin expression as the primary role of silk glands but also suggest that additional silk-like genes are important for silk production. In contrast, spidroins genes do not encode the silk-like secretions from tarsi; rather novel silk-like genes putatively encode these secretions.

Key words: Digital gene expression, Mygalomorphae, Secretions, Spidroin, Theraphosidae

Introduction

The ability to produce silk fibers has evolved independently in at least twenty-five arthropod orders, including Araneae (spiders), Lepidoptera (moths and butterflies), Embioptera (webspinners), Neuroptera (lacewings), Hymenoptera (bees and wasps), and Diptera (flies) (Sutherland et al., 2010). Silks are used for a variety of purposes among arthropods. Spiders use silk for draglines, to make complex web shapes, to construct their egg cases, among other purposes. Silkworms have received considerable attention for their silk fibers, which they use in the construction of cocoons and webs. Also, lacewings produce silk for egg cases and webspinners use silk to construct extensive tunnel systems. Silk protein production occurs in different tissue types among these organisms. The most well-known silk producing organs are the abdominal and labial glands of spiders and silkworms, respectively. Silks are also synthesized in other anatomical regions. For example, lacewings produce silk in colleterial glands, which are in the female reproductive organs, and webspinners produce silk from glands in their forelimb tarsi. Thus, silks have evolved in a diversity of arthropod taxa, are produced from a wide variety of tissue types, and are used for many different functions.

In 2006, it was reported that silk production is not limited to abdominal silk glands in spiders, but silk is also excreted from spigots on the ventral surface of the tarsi (Figure 3.1; Gorb et al., 2006; Rind et al, 2011). This discovery adds to the complexity of spider silk biology. Despite the considerable attention spiders have received for their abdominal silks, tarsal silks have likely been overlooked due to their microscopic size.

Thus far, tarsal silk appears restricted to species in the family Theraphosidae (tarantulas; Peattie et al., 2011). Tarantulas are often large-bodied. In fact, the heaviest extant spiders are tarantulas (Coddington and Levi, 1991). Tarantulas do not have dragline silk, as dragline silk is a characteristic restricted to Araneomorphae (true spiders). Hence, lacking draglines, tarantulas may have evolved tarsal silk as a mechanism for added adhesion to surfaces in order to prevent catastrophic falls (Gorb et al., 2006).

The existence of tarantula tarsal silk remains controversial. One study suggested that the silk fibrils attributed to the tarsi were merely artifacts of silk carried over from the abdominal glands (Pérez-Miles et al., 2009). This suggestion was refuted based on the fact that tarsal silk fibrils occur in parallel tracks from the legs rather than being randomly oriented (Gorb et al., 2009). Additionally, tarantula tarsal silks are characterized as having a broad fluid region consistent with initial adhesion to a surface, which would not occur with a dry fiber produced in the abdomen that was secondarily placed by the tarsi. Another group suggested that the spigots that excrete the fluid from the tarsi are actually chemosensory in function rather than silk producing, based on their morphological similarity to chemosensory hairs that occur on all extremities (Foelix et al., 2012). Characterization of gene transcripts from tarantula tarsi would provide more evidence for the burgeoning controversy of tarsal silk secretions.

As with silk spinning ability, silk proteins have evolved convergently in a number of arthropods; yet nearly all silk proteins share a number of characteristics. Silk proteins are typically rich in the amino acids, glycine, alanine, and serine (Sutherland et al., 2010). Furthermore, known silk proteins as well as other structural proteins that form fibrils, such as collagens, are composed of repetitive motifs (e.g., poly-alanine, iterated couplets of glycine-alanine, glycine-proline-glycine; Hu et al., 2006b; Gordon and Hahn, 2010). Similarities in characteristics among silk and other structural proteins can be used as criteria for identifying candidate genes for tarsal silk production.

In this study, we use cDNA libraries to characterize the genes expressed in the tarsi and silk glands of two species of tarantulas. Given their minute size, there is considerable difficulty in collecting tarsal fibrils for amino acid analysis or peptide sequencing. Instead, we assess the tissue specificity and expression level of candidate genes using digital gene expression (counts of RNASeq short read matches to a reference sequence set). The advantage of digital gene expression over traditional methods of gene expression analysis (e.g., Northern blot or reverse-transcriptase PCR) is that numerous genes can be profiled simultaneously. Additionally, digital gene expression provides discrete count data that can be analyzed statistically rather than the continuous response data measured by fluorescence intensity in microarrays (Robinson et al., 2010).

Our goals are to determine whether the same spidroin genes that are expressed in spider abdominal glands are also expressed in tarantula tarsi. Alternatively, spidroin expansion via gene duplication may have resulted in new spidroins being expressed in tarsal glands. A third possibility is that the tarsal silk-like secretions may be encoded by genes unrelated to spidroins.

Materials and Methods

Taxonomic Sampling

We sampled two species of Theraphosidae, *Aphonopelma seemanni* (Costa Rican Zebra Tarantula), which is a ground dwelling tarantula, and *Poecilotheria regalis* (Indian Ornamental Tree Spider), which is an arboreal (tree dwelling) tarantula. Both of these species have been shown to produce silk excretions from their tarsi and have been characterized for their spidroin gene transcripts (Rind et al., 2011; Starrett et al., 2012).

cDNA Library Construction and Screening

Spiders were anesthetized with CO_2. From each spider, the entire set of silk glands was removed intact and the distal half of the tarsi was removed from each leg. Tissues were frozen in liquid nitrogen and stored at -80º C. Total RNA and complementary DNA (cDNA) library construction methods followed those described in Garb et al. (2007). Briefly, total RNA was extracted using TRIzol (Invitrogen, Carlsbad, CA) and the RNeasy Minikit (Qiagen, Valencia, CA). Messenger RNA was isolated from total RNA using Dynal magnetic beads with oligo-(dT) anchors (Invitrogen). Double stranded cDNA was constructed using the Superscript Choice protocol (Invitrogen), and then size selected for large fragments using Chroma Spin 1000 columns (Clontech, Mountain View, CA). The size-selected cDNA was ligated into pZErO 2.0 vectors that had been digested with EcoRV, and then transformed into TOP10 *Escherichia coli*

(Invitrogen). For silk gland and tarsi libraries, we arrayed ~1400 and ~2800 cDNA clones, respectively, into 96-well microtiter plates. Libraries were stored at -80º C.

We screened approximately 480 clones of each library using the method of Beuken, Vink, and Bruggeman (1998) and sequenced clones containing inserts ≥500 base pairs with T7 and Sp6 universal primers. Sequences were compared to the NCBI nr database using blastx (Altschul et al., 1990) to identify genes with potential homology to silk, secretion, or housekeeping functions. For sequences that lacked convincing blastx results, we conducted conceptual translations in all six frames using Sequencher v. 4. 5 (Ann Arbor, MI) and screened for putative gene products that had repeated amino acid sequence motifs or glycine, serine, and/or alanine richness. Libraries were also replicated onto nylon filters and probed with γ^{32} P-labeled oligonucleotides. All libraries were screened with GCDGCDGCDGCDGCDGC and CCWGCWCCWGCWCCWGCWCC, which were designed based on motifs common to spidroins (Gatesy et al., 2001; Garb and Hayashi, 2005; Garb et al., 2006). Additionally, libraries were screened with taxon specific probes designed from the size-selected clones. The *A. seemanni* libraries were screened with probes TTATCACACATCATTTTTCC, GTTTCCRCCTCAGTGCTTGC, and GGAGGACTGGAGCCACCAAGAAG, and the *P. regalis* libraries were screened with probes CATGGCAGAGGGTATCAGGT, AGTGTAATTTGCAATGCC, GCAAGAGCAATGGCGTTTCC, ATAGGCATAAGCACCAGCGTT, GTAAGCATAAGCCTCGGCTCC, TGGCTCAGACGTAGCGGTGGG, MAGGAAATCCTCCAGCGA, AGGATTCACTCCTGGGTATAT, AGCWCCAACDCCAGCTCC, TTCTTCATCACAGAACTCAGT, AGCTTCTTTTTCTTCATCTCC, GAAATTTTGATGCCCATGCTC, CTGATGTTTATGGGTCTGAGC, GGGACCGAAGGGTATAGC, and CTCCTTTGCAAAGTTGAATGC.

Next Generation Sequencing and Digital Gene Expression Analysis

We dissected tissues from an additional individual of each species to be used for next generation sequencing. Prior to dissection, tarantulas were forced to walk or stay stationary on a vertical glass surface for at least 15 minutes to stimulate expression of tarsal secretions. Four tissue types were harvested for each species: silk glands, tarsi, cephalon, and femora. Dissection of silk gland and tarsi tissues was performed as described above. Legs were pulled from the cephalothorax and femora were removed from the remainder of the leg. Total RNA was extracted as described above followed by DNase treatment using TURBO^{TM}-DNase (Ambion).

Total RNA from each tissue type was used to make an RNASeq library with the Illumina TruSeq RNA sample preparation kit v2 by the Johns Hopkins Deep Sequencing and Microarray Core Facility. Each library was assigned a unique index (barcode). Libraries were pooled and run in one 2x100 (paired-end, 100 cycles each end) lane on an Illumina HiSeq2000 at the UC Riverside IIGB Genomics Core. For *A. seemanni*, the cephalothorax, femora, tarsi, and silk gland RNAseq libraries consisted of approximately 18.6, 10.7, 11.8, and 17.8 million (M) paired end reads, respectively. For *P. regalis*, the RNAseq libraries for the cephalothorax, femora, tarsi, and silk gland tissues consisted of approximately 20.7, 24.6, 8.5, and 22.2 M paired end reads, respectively.

RNASeq reads were matched to EST sequences using Bowtie v.2 (Langmead and Salzberg, 2012). Sequence matches with a quality score of 30 or greater and for which both members of a paired end read mapped to the same cDNA were tallied. We did not adjust read counts for differences in transcript length, largely because many of our cDNA transcripts were not full-length. The tallied bowtie matches for each tissue type were analyzed for statistically significant differential expression with DESeq (Anders and Huber, 2010). We treated the read counts of the cephalothorax and femora as control replicates and compared the read counts of the silk glands to the control and the read counts of the tarsi to the control. Differential expression was considered significant if the adjusted p-value was ≤0.1, following Anders (2012).

Results

We obtained 119 and 142 unique EST sequences for *Aphonopelma seemanni* and *Poecilotheria regalis*, respectively. Of the 119 *A. seemanni* ESTs, ten from the silk gland library and four from the tarsi library either had a significant blastx hit (*E* value <1e-4) to a fibroin, secretory protein, or structural protein, or lacked a significant blastx hit but had a conceptual translation that was silk-like. Silk-like was defined as rich (>15% of total amino acid composition) in at least one of the common arthropod silk amino acids, glycine, serine, or alanine, and exhibiting repeating sequence motifs. These fourteen ESTs were designated candidate transcripts of interest. From the 142 *P. regalis* ESTs, sixteen (ten from the silk gland library, five from the tarsi library, one common to both libraries) were candidate transcripts of interest, based on the same criteria used for *A. seemanni*.

The cDNA libraries from *A. seemanni* and *P. regalis* served as sources for transcript discovery, but not of transcript expression level. To determine whether ESTs were upregulated in silk glands and/or tarsi, short cDNA reads were generated from tissue-specific RNASeq libraries. The short reads were mapped to the ESTs (Bowtie2) and the number of reads mapping to each EST was used as an estimate of gene expression level. The 119 *A. seemanni* sequences were compared to paired end reads from the cephalothorax, femora, tarsi, and silk gland tissues. This resulted in a total of ~13.3 M paired end reads mapped to 104 ESTs (Table 3.1). For *P. regalis*, the 142 cDNA sequences were compared to paired end reads from the cephalothorax, femora, tarsi, and silk gland tissues, for a total of ~10.8 M mapped paired end reads to 137 ESTs (Table 3.2).

For both tarantula species, the number of mapped reads to each EST was compared to identify differentially expressed genes (DESeq analyses) between silk glands and cephalothorax/femora, and also between tarsi and cephalothorax/femora. Here, we discuss those genes that were significantly higher in expression in the silk glands and/or tarsi compared to the control tissues. For transcripts with significant blastx hits we provide the name of the top blast hit, the *E*-value, GenBank accession number, and species. With *A. seemanni*, we identified six EST sequences that were upregulated in the silk glands (Figure 3.2A). All six ESTs were among our candidate transcripts of interest and corresponded to two fibroins (fibroin 1_ANS, 0.0 *E*-value hit to JX102557 from *Aphonopelma seemanni*; fibroin 2_ANS, 0.0 *E*-value hit to JX102558 from *A. seemanni*; Figure 3.2B) and four ESTs without significant hits to the GenBank nr database (Figure 3.2C). One of these four ESTs translates into a serine and glycine rich protein (ANS_SerGly_rich1), while the other three predicted proteins are rich in glycine (ANS_Gly_seq1, ANS_GLYrich_var1, ANS_GLYrich_var2). ESTs ANS_GLYrich_var1 and ANS_GLYrich_var2 have high amino acid identity to each other.

Twenty-one EST sequences were significantly upregulated in the *P. regalis* silk glands, fourteen of which had significant blastx hits (Figure 3.3A). Three of the fourteen ESTs were candidate transcripts of interest and corresponded to fibroin 2 (0.0 *E*-value hit to JX102561 from *Poecilotheria regalis*; Figure 3.3B), PRS_GlyHypothetical_var1 (3e-17, XP_001835549; from a *Coprinopsis*; Figure 3.3C), and PRS_GlyHypothetical_var2 (1e-17, also to XP_001835549; Figure 3.3C). The other eleven highly expressed genes with significant blastx hits consist of arylsulfatase A (2e-27; XP_002006611 from a *Drosophila*), fasciclin (3e-10; XP_002409988 from an *Ixodes*), cuticular protein (1e-26; XP_002435673 from an *Ixodes*), phosphoglycerate dehydrogenase (1e-14; NP_955871 from a *Danio*), ADP-ribosylation factor (2e-56; XP_002426585from a *Pediculus*), hemocyanin A (e0.0; P14750 from an *Aphonopelma*), heat shock protein 90 (1e-79; AEF33377 from a *Crassostrea*), elongation factor 1 alpha (1e-44; ACB70375 from an *Ornithodoros*), elongation factor 1 alpha (4e-177; DAA34050 from an *Amblyomma*), protein disulfide isomerase (1e-77; XP_002405832 from an *Ixodes*), and mitochondrial ADP/ATP carrier (e0.0; AEO34408 from an *Amblyomma*). The remaining seven highly expressed ESTs were candidate transcripts of interest that lacked significant blastx hits. These seven ESTs consisted of one sequence (PRF_SerRepetitive1; Figure 3.3C) that had repeats of SSSSEPIPSIGPE in its conceptual translation, and the remaining six had translations that fell into two groups (Figure 3.3D). One of these groups is rich in glycine and serine (PRS_GlySer_rich_var1 and PRS_GlySer_rich_var1; high sequence identity to each other), while the other is rich in glycine and has poly-cysteine (PRS_GlyRich_var1, var2, var3, var4; high sequence identity to each other).

The tarsi of *A. seemanni* had eight ESTs with significantly higher expression than that of the cephalothorax and femora (Figure 3.4A). One of these ESTs (ANF_GLYrich1) was previously identified among our candidate transcripts of interest. This sequence did not have significant blastx hits but its translation was rich in glycine (Figure 3.4B). The other seven of the upregulated ESTs had significant blastx hits to DEAD-box helicase (9e-08; XP_002595060 from a *Branchiostoma*), and hemocyanin subunit D (4e-159, P02241 from an *Aphonopelma*), B (3e-161, Q9NFH9 from an *Aphonopelma*), C (e0.0, Q9NFL6, from an *Aphonopelma*), E (e0.0, CAA34643, from an *Aphonopelma*), or G (e0.0, Q9NFL4, from an *Aphonopelma*).

Poecilotheria regalis had eleven ESTs that were significantly more abundantly expressed in tarsi than the cephalothorax and femora (Figure 3.5). The sequence with the highest fold increase was PRF_SerRepetitive1. This is the EST with repeated SSSSEPIPSIGPE in its translation that was also identified as upregulated in the abdominal silk glands (Figure 3.3C). The remaining eleven ESTs had significant blastx hits and consisted of one candidate transcript of interest that blasted to a putative secreted salivary protein (9e-30; XP_002406260 from an *Ixodes*), and ten others that blasted to retinol dehydrogenase 9 (2e-12; NP_001090337 from a *Xenopus*), hemocyanin subunit A (e0.0; P14750; from *Aphonopelma*), transketolase (1e-141; AEO32930 from an *Amblyomma*), membrane glycoprotein LIG 1 (2e-61; XP_002405264 from an *Ixodes*), protein disulfide isomerase (1e-77; XP_002405832 from an *Ixodes*), heat shock protein 90 (1e-79; AEF33377; from a *Crassostrea*), ribonucleoprotein protein (PRF177; 3e-91; ADV40069 from a *Latrodectus*), or elongation factor 1 alpha (1e-44; ACB70375 from *Ornithodoros*), or elongation factor 1 alpha (4e-177; DAA34050 from an *Amblyomma*).

Discussion

Most silk gland expression research on spiders has focused on characterizing members of the spidroin gene family, which have been shown to be the major constituents of silk fibers (Xu and Lewis, 1990; Hayashi et al., 2004; Hu et al. 2005b; Garb et al., 2007). However, recent studies have identified non-spidroins that are involved in fiber formation and other silk gland secretions. Specifically, Egg Case Protein (ECP) genes have been identified (Hu et al., 2005a; 2006a), as well as genes encoding the glue-like substance produced from aggregate glands (Choresh et al., 2009). ECPs and aggregate glue proteins do not exhibit similarity to each other or to the conserved regions of spidroins, suggesting that there are at least three gene families of structural proteins in spider silk glands. Additionally, the finding of silk-like secretions produced in the tarsi of tarantulas raises questions about the tissue specificity of expression of known silk genes.

Alternatively, novel genes that are not homologous to spidroin genes may encode tarsal silk-like secretions.

We have identified genes that are highly expressed in silk glands and/or tarsi in two tarantula species in order to further elucidate the genes involved in silk production and identify candidates for tarsal silk encoding genes. Using blastx searches, we determined if transcripts had significant similarity to genes that encode structural proteins or proteins involved in housekeeping functions. Additionally, we uncovered numerous genes that did not have significant blastx hits suggesting that these genes are specific to tarantulas or are too divergent in sequence from homologs on GenBank. Our approach using statistical analyses of digital expression data for silk gland, tarsi, and control tissues allowed us to further refine our list of silk-associated transcript candidates of interest.

Characterization of Highly Expressed Genes in Tarantula Silk Glands

For *A. seemanni* and *P. regalis* silk glands, only fibroins were common across both species as ESTs with significant blastx hits that were highly expressed. This commonality in high silk gene expression across species but no other genes indicates there is need for high quantities of silk proteins for fiber formation. While no housekeeping genes were upregulated in *A. seemanni*, *P. regalis* exhibited high expression of multiple genes involved in regular cellular functions. Among these, the most highly expressed were arylsulfatase A precursor, fasciclin, and a cuticular protein encoding gene, which all had log 2 fold increases of ≥11.4 relative to the control sample. The importance of these highly expressed genes in silk production is unclear.

Arylsulfatase A is involved in hydrolysis of sulfates, fasciclin is involved in cell adhesion, and cuticular proteins are essential for cuticular sclerotization. The remaining *P. regalis* ESTs with significant blastx hits showed comparably moderate log 2 fold increases (<4) in comparison to the control sample. These genes function in catalysis of 3-phosphoglycerate into 3-phosphohydroxypyruvate (phosphoglycerate dehydrogenase), vesicular traffic regulation (ADP- ribosylation factor), oxygen transport and storage (hemocyanin A), protein chaperone and folding (heat shock protein 90), translation (elongation factor 1 alpha), disulfide bond breaking during protein folding (protein disulfide isomerase), or ADP/ATP transfer (mitochondrial ADP/ATP carrier).

We identified novel genes with silk gene-like characteristics in the silk glands of both species of tarantulas that were highly expressed. Conceptual translations of all of these novel genes show that they are rich in one or more of the following amino acids: alanine, glycine, serine. Additionally, thirteen of these novel genes exhibit runs of poly-alanine, poly-serine, glycine-alanine, or poly-glycine. These motifs have been determined to create beta sheets or turns in spidroins (Hu et al., 2006b). Additionally, cysteines are present in eleven of the novel genes. Cysteine residues may be important for linking molecules together through disulfide bonds, as has been suggested for *Latrodectus hesperus* Egg case proteins (Hu et al., 2005a; 2006a).

Characterization of Highly Expressed Genes in Tarantula Tarsi

Most of the genes that were highly differentially expressed in the tarsi of both *A. seemanni* and *P. regalis* were genes with significant blastx hits. However, we found no strong evidence that the spidroin genes expressed in the abdominal silk glands are also expressed in tarantula tarsi. Spidroins were not highly expressed in the tarsi for either species and the paired end read counts for any spidroin in the tarsi was ≤6 out of ≥10 million total reads. This means that the fibrils that have been documented from tarantula tarsi (Gorb et al., 2006; Rind et al., 2011) cannot be associated with the same type of spidroins known from spider silk glands. As for the hypothesis that the tarsal secretions are chemosensory molecules, we also did not identify any sequences with blastx hits to chemosensory genes. However, chemosensory genes have not been characterized from spiders to date and thus our results are inconclusive about the potential role of chemosensory genes in tarsal excretions.

While the significant blastx hits did not correspond to spidroin or chemosensory genes, they all appear to be components of regular cellular functions. In both tarantulas, hemocyanin paralogs were highly expressed. Hemocyanins are generated in multiple regions in arachnids, such as the endocuticle of the carapace in a scorpion and the heart in tarantulas (Alliel et al., 1983; Kempter, 1986; Markl et al., 1990). To the best of our knowledge, testing for hemocyanin expression in tarsi has not been done.

Ten highly expressed transcripts in *P. regalis* tarsi had significant blastx searches. These genes putatively function in salivary secretion (secreted salivary protein), retinol oxidation (retinol dehydrogenase 9), oxygen transport and storage (hemocyanin subunit A), catalysis in the pentose phosphate pathway (transketolase), interaction with receptor tyrosine kinases (memebrane glycoprotein LIG 1), disulfide bond breaking during protein folding (protein disulfide isomerase), protein chaperone and folding (heat shock protein 90), RNA binding (ribonucleoprotein protein), or translation (elongation factor alpha). While no obvious connection can be made between most of these genes and tarsal secretions, the secreted salivary protein stands out as a potential candidate. However, blastx searches of this sequence identify homologs in a number of insects and arachnids and this gene may be widely expressed in the body as it occurs in expression libraries from salivary glands in the tick (Ribeiro et al., 2001) and venom glands of the scorpion (Morgenstern et al., 2011). Thus, none of the genes with significant blastx hits are likely candidates for tarsal silk.

The transcripts with the highest fold increase in expression in the tarsi for both species were transcripts that lacked significant blastx hits but had silk-like characteristics. In *A. seemanni*, we identified ANF_GLYrich1, which is composed of ~41% glycine. For *P. regalis*, PRF_SerRepetitive1, which encodes a serine rich protein (~24%) with repeats of SSSSEPIPSIGPE, was identified. PRF_SerRepetitive1 was also highly expressed in *P. regalis* silk glands, which could indicate that this gene is associated with silk production in both tissues. Thus, our best candidates for tarsal silk secretions are consistent with a set of genes unrelated to spidroin genes.

Conclusions

Silk has received considerable attention in spiders, yet, the full complement of silk-associated genes expressed in spiders remains unknown. While much focus has been placed on spidroin genes, other genes important for silk have gone uncharacterized. Using blastx, we identified candidates for tarsal fibril secretions and from these candidates, eighteen were found to be upregulated in tarantula tarsal tissue. Twelve of these candidates are entirely new, with no significant match to any known gene, indicating that they are clade-specific genes. Clade specificity is consistent with tarsal silk secretion as a derived trait that evolved within spiders. Further exploration of these tarsal secretion candidate genes as to their phylogenetic distribution, whether they necessarily co-occur with tarsal silk secretions, and the structural characteristics of the gene products needs to be done. Our expression data shows that spidroins are unlikely the fibril-like product that has been observed emanating from spigots on tarantula tarsi. Instead we uncovered transcripts from the tarsi that are not homologous to any genes in the nr data base but encode proteins that are glycine rich or serine rich and repetitive. These silk-like features make these genes targets for future efforts towards characterizing tarsal secretions.

References

Alliel PM, Dautigny A, Lamy J, Lamy J-N, Jolles P (1983) Eur. J. Biochem. 134, 407-414.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J. Mol. Biol. 215: 403-410.

Anders S (2012) Analysing RNA-seq data with the DESeq package. Available: http://www-huber.embl.de/users/anders/DESeq/. Accessed June 1, 2012.

Anders S, Huber W (2010) Differentila expression analysis for sequence count data. Genome Biology. 11:R106.

Beuken E, Vink C, Bruggeman C (1998) One-step procedure for screening recombinant plasmids by size. BioTechniques. 24: 748–750.

Choresh O, Bayarmagnai, Lewis RV (2009) Spider Web Glue: Two Proteins Expressed from Opposite Strands of the Same DNA Sequence. Biomacromolecules. 10(10):2852-2856.

Coddington JA, Levi HW (1991) Systematics and Evolution of Spiders (Araneae). Annu. Rev. Ecol. Syst. 22: 565-592.

Foelix RF, Rast B, Peattie AM (2012) Silk secretion from tarantula feet revisited: alleged spigots are probably chemoreceptors. J. Exp. Biol. 215: 1084-1089.

Garb JE, DiMauro T, Vo V, Hayashi CY (2006) Silk genes support the single origin of orb webs. Science. 312(5781): 1762.

Garb JE, DiMauro T, Lewis RV, Hayashi CY (2007) Expansion and Intragenic Homogenization of Spider Silk Genes since the Triassic: Evidence from Mygalomorphae (Tarantulas and Their Kin) Spidroins. Mol. Biol. Evol. 24(11): 2454-2464.

Garb JE, Hayashi CY (2005) Modular evolution of egg case silk genes across orb-weaving spider superfamilies. P. Natl. Acad. Sci. USA. 102(32): 11379-11384.

Gatesy J, Hayashi CY, Motriuk D, Woods J, Lewis RV (2001) Extreme Diversity, Conservation, and Convergence of Spider Silk Fibroin Sequences. Science. 291: 2603-2605.

Gorb SN, Niederegger S, Hayashi CY, Summers AP, Votsch W, et al. (2006) Silk-like secretion from tarantula feet. Nature. 443: 407.

Gorb SN, Niederegger S, Hayashi CY, Summers AP, Vötsch W, Walther P (2009) Reply: Silk production from tarantula feet questioned. Nature. 461:E9-E10.

Gordon MK, Hahn RA (2010) Collagens. Cell. Tissue Res. 339:247-257.

Hayashi CY, Blackledge TA, Lewis RV (2004) Molecular and Mechanical Characterization of Aciniform Silk: Uniformity of Iterated Sequence Modules in a Novel Member of the Spider Silk Fibroin Gene Family. Mol. Biol. Evol. 21(10): 1950-1959.

Hu X, Kohler K, Falick AM, Moore AMF, Jones PR, Sparkman OD, Vierra C (2005a) Egg Case Protein-1. J. Bio. Chem. 280(22): 21220-21230.

Hu X, Lawrence B, Kohler K, Falick AM, Moore AMF, McMullen E, Jones PR, Vierra C (2005b) Araneoid Egg Case Silk: A Fibroin with Novel Ensemble Repeat Units from the Black Widow Spider, *Latrodectus hesperus*. Biochemistry-US. 44:10020-10027.

Hu X, Kohler K, Falick AM, Moore AMF, Jones PR, Vierra C (2006a) Spider Egg Case Core Fibers: Trimeric Complexes Assembled from TuSp1, ECP-1, and ECP-2. Biochemistry. 45: 3506-3516.

Hu X, Vasanthavada K, Kohler K, McNary S, Moore AMF, Vierra CA (2006b) Molecular mechanisms of spider silk. Cell. Mol. Life Sci. 63: 1986-1999.

Kempter B (1986) In: Linzen B (ed): "Invertebrate Oxygen Carriers." Berlin, Heidelberg, New York: springer, pp 489-494.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nature Methods. 9(4):357-359.

Markl J, Stump S, Bosch FX, Voit R (1990) in: Invertebrate Dioxygen carriers (eds.) Preaux G, Lontie R Leuven University Press, Leuven, pp. 497-502.

Morgenstern D, Rohde BH, King GF, Tal T, Sher D, Zlotkin E (2011) The tale of a resting gland: Transcriptome of a replete venom gland from the scorpion *Hottentotta judaicus*. Toxicon. 57(5):695-703.

Peattie AM, Dirks J-H, Henriques S, Federle W (2011) Arachnids Secrete a Fluid over Their Adhesive Pads. PLoS ONE. 6(5): e20485.

Pérez-Miles F, Panzera A, Ortiz-Villatoro D, Perdomo C (2009) Silk production from tarantula feet questioned. Nature. 461: E9.

Rind FC, Birkett CL, Duncan B-JA, Ranken AJ (2011) Tarantulas cling to smooth vertical surfaces by secreting silk from their feet. J. Exp. Biol. 214: 1874-1879.

Robinson MD, McCarthy DJ, Smyth GK (2010) edger: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 26(1):139-140.

Starrett J, Garb JE, Kuelbs A, Azubuike UO, Hayashi CY (2012) Early Events in the Evolution of Spider Silk Genes. PLoS ONE 7(6): e38084. doi:10.1371/journal.pone.0038084.

Sutherland TD, Young JH, Weisman S, Hayashi CY, Merritt DJ (2010) Insect Silk: One Name, Many Materials. Annu. Rev. Entomol. 55: 171-188.

Xu M, Lewis RV (1990) Structure of a protein superfiber: Spider dragline silk. P. Natl. Acad. Sci. USA. 87: 7120-7124.

Tables

Table 3.1. Bowtie2 read counts for *Aphonopelma seemanni*. Counts determined to be significantly higher than control counts (cephalothorax and femora) are highlighted.

EST	Cephalothorax	Femora	Tarsi	Silk Glands
Golgi autoantigen (ANS10)	172	108	242	246
Transoldolase 1 (ANF1)	2428	1086	7564	2810
Brahma Protein-like (ANF3)	180	2	348	334
16sND1 (ANS & ANF)	2209240	1226368	1298498	2058122
small Heat Shock Protein 24.1 (ANS1)	33806	5902	754	37822
Orn decarboxylase antienzyme (ANF75)	5012	1660	6638	20488
Gag-like protein (ANS41)	3784	2634	1822	1314
ANS_GLYrich_var1	0	0	2	**67176**
Vasa protein (ANF)	0	2	0	0
Zinc finger protein 470 (ANF41)	2	0	4	2
RNA (guanine-7) methyltransferase (ANF146)	194	104	396	366
Ribosomal protein S7 (ANF84)	3252	1682	6154	10426
Hemocyanin d (ANS55)	1308	2	**17120**	4044
Actin beta (ANS69)	165200	117560	144	40756
WD repeat domain 33 (ANS74)	10	0	14	6
Zeta1-cop (ANS54)	162	98	408	566
Hemocyanin b (ANF145)	28	0	210	16
Te1-like transposase (ANF118)	0	2	20	10
ANF_Hydroxy1	4	10	950	24
Mitochondrial ribsomal protein (ANF81)	268	148	398	482
NADH subunit 5 (ANS108)	25680	16656	10106	22806
Profilin5 (ANF62)	2462	214	6588	4476
UbiE-YGHL1 fusion protein (ANF11)	600	238	3356	2910
Hemocyanin a (ANF)	3034	58	**41688**	22460
heat shock protein hsp20.1 (ANS37)	0	0	0	6
Secreted protein (ANS_SecretedPro)	2626	380	7668	6550
DAZ associated protein 1 (ANS3)	226	0	536	6
cytochrome oxidase subunit III (ANS14)	1160	704	1956	2022
ANS_GLYrich_var2	2	0	4	**27852**
ANF_SecretedSaliv2	11230	970	90544	8062
nd1-16S (ANF137)	4268	116	3334	4020
AN_immune	3400	1710	22798	1974
Notch homolog 1 (ANS127)	1726	54	4574	1884
Hemocyanin d (ANF)	276	0	2618	38
ferritin (ANF35)	14132	9908	13280	17446
sulfide dehydrogenase like (ANF54)	134	70	406	386
thioredoxin-like (ANS87)	394	36	588	926
calcyphosine (ANS95)	7768	2416	16780	17808
RNA binding protein (ANS59)	68	40	64	72
presenilin-like protein 3 (ANF68)	4630	1988	9754	12494
Actin beta (ANS)	400204	76452	286	112046
Laminin (ANS)	8262	2762	13412	36096

CG32018-PE, isoform E (ANS24)	284	8	8	12
myosin 2 light chain (ANS71)	19188	4046	40	10802
metallothionein A (ANS116)	2776	3320	7134	4844
cytochrome b (ANS47)	4	2	0	0
Hemocyanin c (ANF)	1842	30	**24400**	11072
ANS_hydroxy3	63402	25560	112	18672
ubiquitin-conjugating enzyme E2N (ANF53)	222	140	374	460
muscle actin (ANS140)	674	0	0	2
ATPsynthase (ANF)	212	160	124	164
Glutamate receptor (ANS96)	64	32	2	36
Drab2 (ANS11)	136	72	186	912
NADH dehydrogenase subunit 2 (ANS8)	46640	29156	19632	26676
elongation factor 1-alpha (ANS39)	44340	11134	74598	184644
ECHGR Myophilin (ANF28)	13772	4560	22506	30846
ANS_SerGly_rich1	0	0	2	**10706**
ubiquitin C-terminal hydrolase (ANS45)	8	2	4	2
ATP synthase, alpha subunit (ANS111)	1708	10	1200	488
elongation factor 2 (ANF104)	35996	14204	44194	78342
alpha-2-macroglobulin (ANF63)	2602	430	9172	8198
Heat shock cognate 70 (ANS)	47884	9312	91038	81654
G3PDH (ANS)	103286	29962	24808	26450
40S ribosomal protein (ANS30)	3834	2846	4924	5508
alpha tubulin (ANF)	33852	5816	63682	39342
elongation factor 1 delta (ANS16)	5636	2440	7888	12054
elongation factor 1 alpha (ANF74)	17980	10	24718	5604
arginine kinase (ANS131)	38326	278	64	172
Hemocyanin b (ANF145)	1520	34	**20574**	11428
hemocyanin g (ANF)	4546	68	**52230**	21568
translation initiation factor (ANS18)	6118	2494	7848	12778
cytochrome b5 (ANS20)	612	222	1512	1474
hemocyanin e (ANF)	3880	104	**50308**	24704
ANF_SecretedSaliv1	1368	1332	24304	1704
DEAD-box helicase ANF_hydroxy2	0	12	**1378**	40
Syndecan binding protein (ANF140)	1526	8	2760	1184
Fibroin 2 (ANS)	10	2	6	**227114**
progranulin-b (ANS134)	584	10	40	2006
PDGFA associated protein 1 (ANF2)	3906	1220	6002	3970
mariner transposase (ANF10)	2	0	0	0
Muscle protein 20 (ANS)	715554	10804	838	64344
elongation factor-1 gamma (ANF49)	5828	1630	10130	24284
beta tubulin (ANS)	2946	14	2920	1494
translation initiation factor 4A (ANF97)	2214	2	2566	820
ANF_SecretedPro	5040	676	15176	9922
NADH dehydrogenase subunit 1 (ANS156)	16172	7570	6232	11514
Venom protein PN16C3 (ANF29)	1240	462	4548	1526
ANS_Gly_seq1	0	0	2	**11078**
flotillin 1 (ANF56)	24	24	134	24
Tropomyosin (ANS98)	458362	319548	300	26696
LiPoate Ligase (ANF59)	14	2	20	12
NADH dehydrogenase subunit 5 (ANF40)	4	0	0	0

ATP-binding cassette, sub-family F (ANF102)	2648	348	1702	2664
beta-2-tubulin class (ANS44)	8078	3574	13856	19090
RNA binding protein (ANF8)	5462	3382	11638	5692
Solute carrier 25 (ANS)	18742	860	17106	22916
Cytochrome oxidae I (AN)	105954	93694	33322	80664
Fibroin 1 (ANS)	0	0	6	**1506**
ANF_GLYrich1	0	2	**44528**	0
cytochrome c oxidase subunit III (ANS25)	26296	18888	8056	16250
muscle LIM protein (ANS43)	229162	31028	1172	35430
IAP-associated factor (ANS93)	3218	2868	5506	2998
ribosomal protein S10 (ANF66)	4506	2618	7502	9426
chromobox homolog 1 (ANS28)	252	26	410	112
Total Reads	5037788	2119196	2367468	3761444

Table 3.2. Bowtie2 read counts for *Poecilotheria regalis*. Counts determined to be significantly higher than control counts (cephalothorax and femora) are highlighted.

EST	Cephalothorax	Femora	Tarsi	Silk Glands
Chain A, Crystal Structure Of Creatine (PRS120_T7)	492096	795962	27086	26372
beta tubulin (PRF123)	28210	20530	25018	45446
at-rich interactive domain-containing protein (PRF32)	2	14	2	4
matricellular protein osteonectin, Sparc (PRS68)	2950	3440	594	2968
tyrosyl-tRNA synthetase (PRF37)	282	324	294	1022
transcription elongation factor B (PRS171)	12	22	4	84
PRF_CellSurf	254	338	252	472
4SNc-Tudor domain protein (PRS119)	294	368	320	1188
PRS_GlyRich_var1	0	0	0	**1674**
adenosylhomocysteinase (PRF15)	1218	1746	1920	3288
mitochondrial trifunctional protein (PRF83)	208	244	180	432
RAB-18 (PRS124)	1120	1134	908	2942
NADH dehydrogenase subunit 4 (PRF78)	16	58	2	20
actin (PRS183)	10048	10780	8024	35022
heat shock protein 90 (PRF159)	1724	1864	**2836**	**8110**
paramyosin (PRS5)	110	104	2	184
dual specificity phosphatase (PRF17)	454	552	292	804
X-box-binding protein 1 (PRS25)	1842	2486	2866	7206
ATP-dependent clp protease atp-binding (PRS103)	0	6	6	10
folate carrier protein (PRS23)	106	98	76	156
PRF_SecretedSaliv	286	1640	**6910**	1074
Heterogeneous nuclear ribonucleoprotein (PRF127)	0	4	10	4
Short gastrulation (PRS180)	482	888	210	298
aspartate aminotransferase (PRF109)	5782	6092	2228	7106
tetraspanin (PRF)	3882	3526	3072	8430
cytochrome oxidae I (PR)	9990	18580	2168	8946
PRS_GlySer_rich_var2	2	2	0	**48380**
RNA polymerase subunit K (PRS165)	136	154	228	278
ADP-ribosylation factor (PRS126)	1010	1008	990	**12096**
40S ribosomal protein S11 (PRF56)	1486	1666	1564	3620
actin (PRS)	213954	232244	79254	307540
TroponinC (PRS)	280462	362004	15412	25928
similar to SWI/SNF-related matrix (PRS7)	0	2	0	4
Ubiquitin (PRF4)	1882	2894	1792	4032
PRF_SecretedPro	7980	4478	4348	10424
PRS_GlySer_rich_var1	0	0	0	**148**
ferritin heavy chain-1b (PRF129)	10454	14580	9250	22646
heat shock protein 20.6 (PRS134)	75912	111880	12736	23682
PRS_GlyHypothetical_var1	0	0	0	**1684**
retinol dehydrogenase 9 (PRF102)	2	42	**832**	54
ATPase inhibitor (PRF96)	1600	1960	336	778
PRS_GlyRich_var3	0	0	0	**34**

PRS_GlyRich_var2	0	0	4	**38618**
PRS_Collagen	17568	25432	3950	7534
beta tubulin (PRF151)	2588	3536	3716	7824
ferritin (PRF64)	4130	5496	4026	10378
Succinate dehydrogenase iron-sulfur subunit (PRF14)	940	1310	346	984
heat shock protein 70 (PRF147)	16202	19668	12390	36414
NPAC (PR)	4432	5850	4162	9284
hemocyanin a (PRF178)	3854	1862	**11528**	**24824**
internalin A (PRS35)	492	796	558	1318
NADH2 (PR)	1786	2494	442	1566
ribosomal protein 40S (PRS)	4994	6128	4888	9276
myelinprotein expression factor (PRF107)	5168	6874	8584	12414
histone 2B (PRF81)	2846	3520	3628	5456
mitochondrial ADP/ATP carrier (PR)	76822	84508	108110	**296984**
histamine release factor (PRS61)	33982	40160	31046	75578
allergen Aca s 10 (PRS48)	272898	338994	14616	34836
ATP-binding cassette F (PRS80)	3892	5490	2892	6370
E3ubiquitin (PRF)	9662	10178	11142	24208
heat shock 70kDa protein (PRS181)	5678	5372	4424	12040
60S acidic ribosomal protein P0 (PRS69)	6896	8190	5600	16852
2-phosphodiesterase (PRS59)	76	90	62	320
myosin heavy chain (PRS179)	12	14	0	8
TEF2 (PRF)	7164	8742	6254	14038
histone H3 (PRF59)	2	0	2	0
major vault protein (PRS45)	0	2	0	0
SKN-1 Dependent Zygotic transcript family (PRF41)	40	96	58	152
Fibroin 2 (PRS)	4	22	2	**293878**
peptide chain release factor (PRS169)	5564	6276	5378	13220
elongation factor 1 delta (PRF103)	5890	7238	5344	16574
membrane glycoprotein LIG-1 (PRF105)	292	624	**1422**	1234
Fasciclin (PRS)	0	2	4	**98518**
transketolase (PRS182)	4	22	**116**	92
ferritin (PRF95)	558	814	550	1258
proteasome assembly chaperone 1 (PRS58)	106	100	132	264
DnaJ (Hsp40) (PRS107)	230	250	290	900
Putative methyltransferase family protein (PRF1)	0	2	4	2
translation initiation factor eIF3 (PRS137)	7336	8546	6274	11890
ATP synthase F0 subunit 6 (PRS42)	4592	7668	1192	4326
similar to transposase-like (PRF128)	18	12	22	28
karyopherin alpha 2 (PRF75)	3168	4648	5398	6756
cAMP responsive element binding protein (PRF47)	3432	3894	1980	3044
RNA binding motif (PRF149)	2488	3004	3000	5070
elongation factor 1 alpha (PR)	13142	16498	**21028**	**64886**
PRF_GlycoPro	1550	948	1016	2256
enolase-phosphatase 1 (PRS86)	120	102	86	224
Ras-related protein Rab-1A (PRF53)	3710	3618	3538	8746
NADH dehydrogenase subunit 3 (PRS157)	340	602	84	364

catalase (PRF52)	34	56	112	126
ubiquitin C-terminal hydrolase (PRF68)	440	410	314	848
60s ribosomal protein L15 (PRS88)	5478	6128	5752	11932
similar to cathepsin D (PRS114)	3566	4684	2868	7568
apoptosis (PR)	22630	22928	16532	36148
actin (PRS16)	141926	160246	7964	89662
Myotubularin related protein 2 (PRF_13)	8	8	6	16
40S ribosomal protein S7 (PRF153)	4304	4656	3544	10132
ubiquitin ligase Cop1 (PRF60)	78	90	110	314
alpha tubulin (PRF29)	17758	19438	22294	31116
elongation factor-1alpha (PRF_142)	24822	33200	**43076**	**121124**
actin (PRF144)	9920	7074	9496	24680
protein disulfide isomerase (PRS99)	1658	2306	**3630**	**8264**
calponin (PRS63)	49928	29910	2864	9932
helicase (PRF108)	150	204	254	406
60S ribosomal protein l17 (PRS72)	2504	2730	2406	6282
60s ribosomal protein L10 (PRS156)	12732	14656	12844	33920
PRS_Cuticular	6	0	0	**17194**
ATPbind (PR)	8466	9500	3506	9392
PRS_GlyHypothetical_var2	0	4	0	**39068**
leucine aminopeptidase-like protein (PRF97)	3702	4670	2580	7266
arginine kinase (PRS129)	1616	2074	176	2136
transmembrane and ubiquitin-like domain (PRF94)	96	76	108	238
serine/threonine-protein kinase Chk2 (PRF66)	14	24	20	34
glycerophosphoryl diester phosphodiesterase (PRF82)	26	62	92	236
translational elongation factor-2 (PRF58)	2	8	6	20
myosin heavy chain (PRF72)	0	2	2	0
16S (PR)	584630	848982	294078	863630
ferritin (PRS)	25956	34918	10916	32272
arylsulfatase B precursor (PRS50)	0	0	2	**1314**
phosphoglycerate dehydrogenase (PRS101)	1424	1660	2028	**44966**
tubulin alpha-1 chain (PRF124)	25376	16022	2036	5206
PRF_SerRepetitive1	0	8	**568**	**82**
PRS_GlyRich_var4	0	0	2	**5888**
cytochrome C (PRF125)	6528	7646	4922	10512
golgi vesicular membrane trafficking protein (PRF137)	16	12	16	54
60S ribosomal protein (PRS52)	6602	7382	6720	15800
NADP-dependent isocitrate dehydrogenase (PRS147)	1764	2442	2074	4050
troponin T (PRS176)	9944	15892	530	1406
cytochrome b (PRS30)	1168	1958	214	892
thioredoxin-like protein (PRS60)	222	246	134	272
calmodulin (PRF164)	11854	16560	4360	7570
40S ribosomal protein S3a (PRF49)	598	1016	186	296
ubiquitin/40S ribosomal (PRS117)	2050	2188	2244	5558
Muscle Protein 20 (PRS)	162166	104650	9362	45370
calponin (PRF24)	10762	11598	5496	21448

RNA-binding protein (PRF177)	702	1034	**1308**	1728
Total Reads	2854546	3652766	999118	3341790

Figures

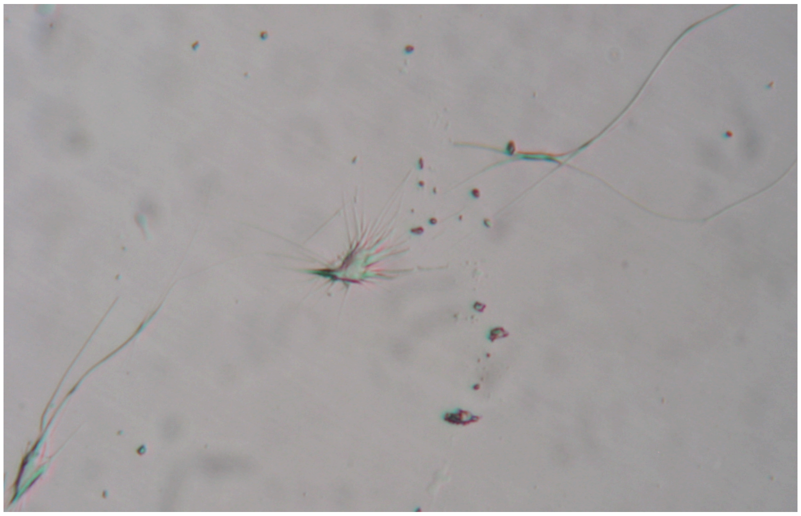


Figure 3.1. Tarsal silk excretion from *Poecilotheria regalis* at magnification of 100X.

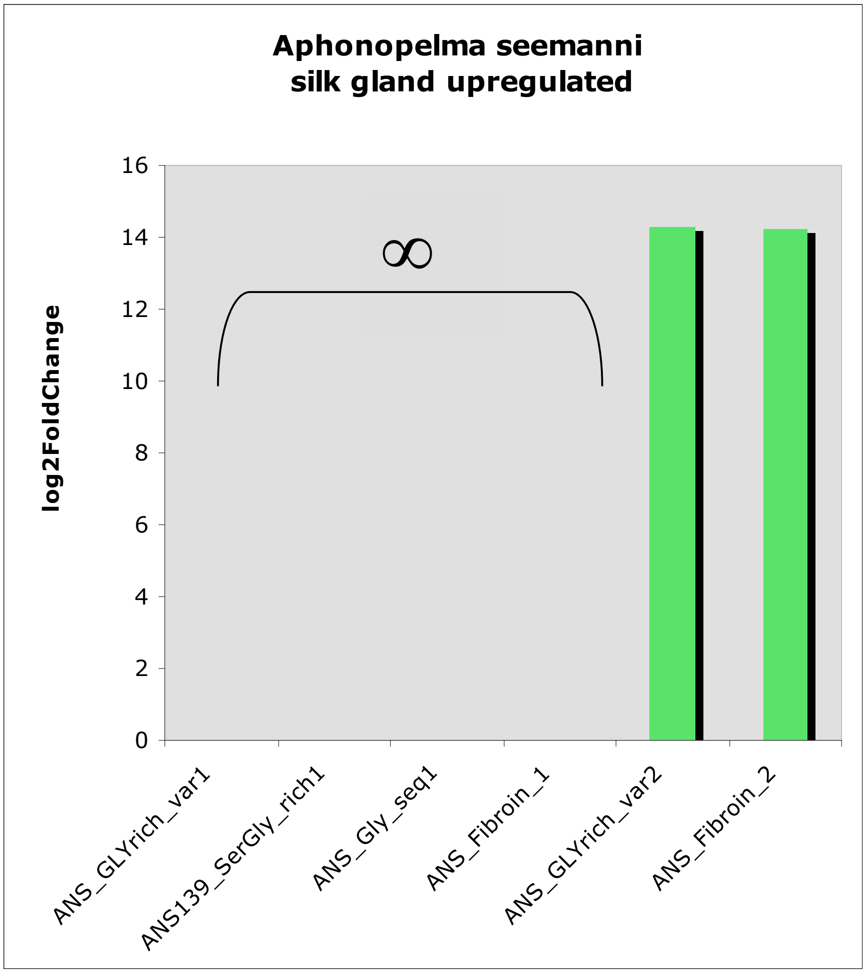


Figure 3.2A. Significantly upregulated transcripts in the silk glands of *Aphonopelma seemanni*. Transcripts with log 2 fold changes that were too high to be calculated are indicated under the bracket with the infinity symbol.

```
ANS_Fibroin_1 (A: 7.88, G: 6.59, S: 19.58)
(…)FQADSQRHDQLPLAT
KKALLSSVTNRGVFLSLTGLSPEAISILSSCAGDTAAQLGIPDLKSIISSEVKRAFTDMSPNASPLTLAEV
LSSILTNIFEGYGSLDHSNAQYFGSEFGKMMCDKFFSSASRDVVSASVTKTVETSSSSRDISSVRSACDLF
SESFVNTLSGQKALYPALGKDILAADIIDCVPDILNSLMSTGVVFTDGSRLQSMLQNTMKSIPRGSGPGVL
CQRIAVPFAEMLSADGKLNSSNASELGRMIGLCVTRVLLRPVYRGGSSASDSTLFTSAGQERTASSTLSST
SRDFRTSSQSTYRAAQTESSPGTSSPSEKF
KKSLLSSLTDRRVFLSLTGFSPEAISILSSCAGDTAAQLGIPDLKSIINSEVKRAFTDMSPNASPLTRAEV
ISSILTNIFEGYGSLDHSNAQYFGSEFGKMMCDKLYFAFGGGTVRSTSRVPASEESYGSFRESTKSYSSVF
SEEASSVFSRLGALAFGVSSSENLPSVAERRISSLESLIFSSITTGNFTSSSFSRILSSVVSEIMKSEPDL
SSREVIMECLLEVIAAMIKIALAGVSSDGIPSVKL*

ANS_Fibroin_2 (A: 15.98, G: 8.30, S: 19.29)
(…)
ARSFAEAFVGNCLKDPIFISIFRKVASSAEASSYTSAGVQSAIRSLGLGLDFVSGTSDLVASKIANVAVGS
SPSEYAEAIASVISSIVSSAGVLNHSNPTTLGMQVAIGFCRGLASSAYPDSGEPLGSPFTGGAAQVGAASS
AISSASATVTESSVAASRAEISSRGADVAAA
ASSFKQVFMSSLCQSEAFNSAFSSPASFSAAVSCILSAVEAATKEVGMGNLASEMAHATSRAAARKSAGSG
STAFAESFASEIGSVLFVRGILNVNNAAYIAARLVKVLLRIFSLTFTVPTAPGGVSSGSATAAAGATSISA
NVPSAGYGNLFPVRLSAPESAGGSFTTGAENIVVVDSNLPAGSLMSGDDQPLFSSLSGSVAQVLSSVEGLL
SPAASRRISALIRSIVSELSTGRLKPSFLKNVLAFVLSQISQTGAGFSASQITIEGLLEILTALLHILLSS
QIGPLNSSTTVSSDVVEAVSSAFLN*
```

Figure 3.2B. *Aphonopelma seemanni* spidroin amino acid sequences. Percent compositions of alanine, glycine, and serine are given in parentheses.

```
ANS_SerGly_rich1 (A: 7.42, G: 15.23, S: 26.95)
MTDFGTHARAVICLTLLCLLPVFIGAAEMKISRLRRSSQPNAKNGSEAESNTETLSGSTSTENANESPKSSSGSSSSASSNGTNGNSSDAAKGAGSESKSG
GTTNGVSASHGENSDGSALESAGDKQTTLPGKANKPSSKSGGNTSGSPVQKDSGEGSRTNSKGTGGSSKSSSSNTSSGSKAKTENDKGTCSSGSSGEGGTR
GSNSKPSGNSKGGNSDSSQVSEPKNKASDGSKPTGSGDSLAGSKSSASNSSGSS*

ANS_Gly_seq1 (A: 3.41, G: 28.29, S: 6.34)
(…)GGAGRGVGRFRGRTRHGANWGFGRSVGGYHGGSGGEEVGGGWPFGFRSNGIRARRKGRRNSFTEGSVSGSGIGVGERTGFGNGNGPGRRYHPRHSPFP
GGRYIIRKGKRHGSDGSGRGHEGGSAGGSGNGIETYLLEKLKKMEETAEMAQMKDMAVDGIHDEDRRHGIFGRDSGDGRWEEMFGRDGGDGRWEEMVGRDG
GDGGDV*


ANS_GLYrich_var1: (A: 4.32, G: 38.13, S: 6.47)
ANS_GLYrich_var2: (A: 3.57, G: 38.57, S: 5.71)
ANS_GLYrich_var1: MRGVLAFLVPVALMFGLQSGLVTGDGGGPGSAGGVGGLLDGSSTPG-GGVGGLLGGVGSLLAGSSPPYVGG
ANS_GLYrich_var2: MRGVLAFLVPVALMFGLQSGLVTGDGGGPGSAGGVGGLLGGSSPPGVGGVPGLLNGVGGLLGGSSPPYVGG


ANS_GLYrich_var1: VPGLLNGVGGLLGGSSPPGGGLGAVTGLLAGGGNFPGNGFGGRGYGGYPGFGGYPFYGGGCCGYGCCCC*
ANS_GLYrich_var2: VPGLLNGVGGLLGGSSPPGGGLGAVTGLLAGGGNFPGNGFGGRGYGGYPGFGGYPFYGGGCCGYGCCCC*
```

Figure 3.2C. *Aphonopelma seemanni* silk gland novel gene amino acid sequences. Percent compositions of alanine, glycine, and serine are given in parentheses.

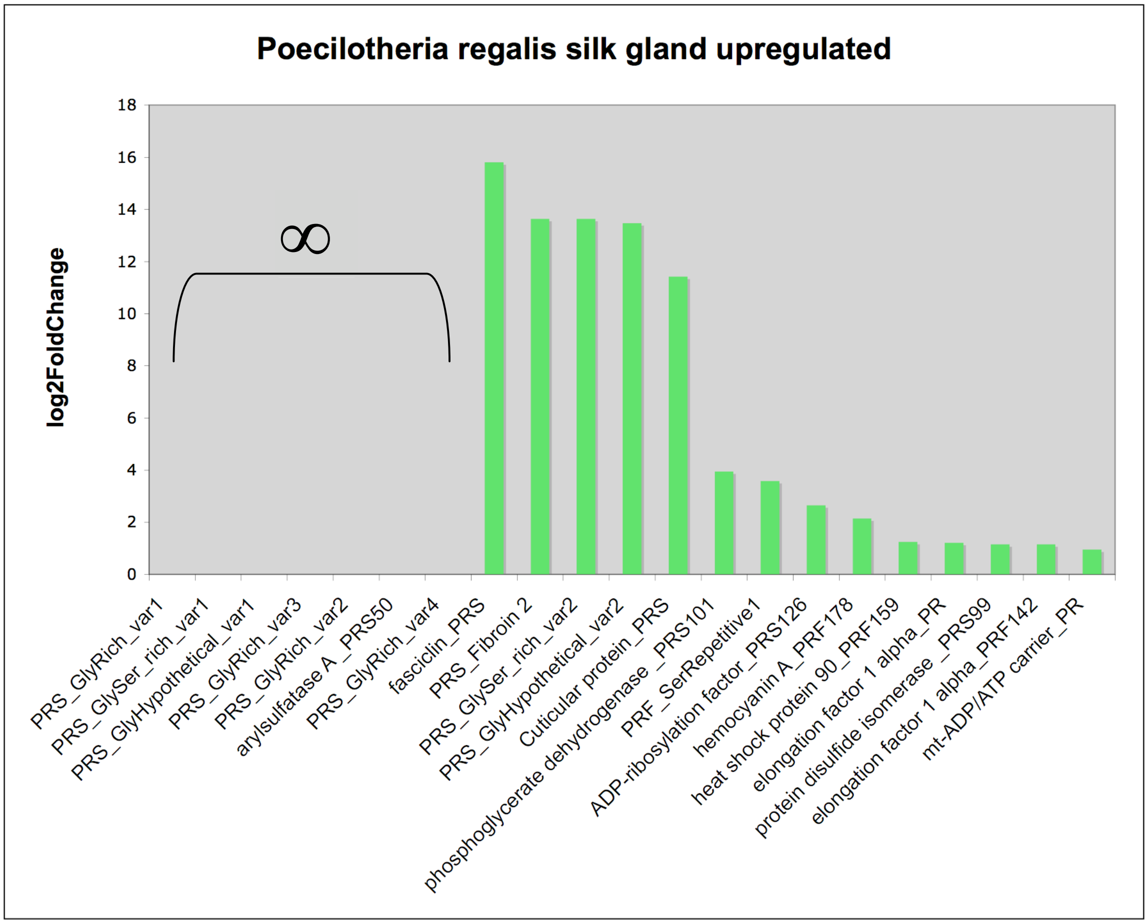


Figure 3.3A. Significantly upregulated transcripts in the silk glands of *Poecilotheria regalis*. Transcripts with log 2 fold changes that were too high to be calculated are indicated under the bracket with the infinity symbol.

```
PRS_Fibroin_2 (A: 30.32, G: 9.71, S: 22.34)
(…)
AVASAGNNAGAYAYARAYASAISQSLSSLGILNSGNAIALANAFSSGASDSAAAAALSAASASAASAATAV
STTTSTSTSAATAAAAAASAAGAAGAGAAAGATASSSFGQNLLSGLLRSDAVVSALSQAYSASTASALASS
YAQSGADRAGFGNYGSVIASAAAS
AVASAGNNAGAYAYARAYASAISQSLSSLGILNSGNAIALANAFSSGASDSAAAAAVSAASASAASAATAS
SRTTSTSTSAATAAAAAASAAGSAGAGAAAGATASSSFGQNLLSGLLRSDAVVSALSQAYSASTASALASS
YAQSGADRAGLGNYGSVIASAAAS
AVASAGNNAGAYAYARAYASAISQSLSSLRILNSGNAIALANAFSSGASDSAAAAALSAASASAASAATAA
STTTSTSTSAATAAAAAASAAGAAGAGAAAGATASSSFGQNLLSGLLRSDAVVSALSQAYSASTASALASS
YAQSGADRAGLGNYGSVIASAAAS
AVASAGNNAGAYAYARAYASAISQSLSSLGILNSGNAIALANAFTSGASSNASSAAAAASTVLAGVAPAGS
SASSSAASASTGAGAAISSVGPAVGFGAGPAPAGGLVSGLPGYSPLNQGFTPYPGVPLPTGSGVSAPVPVS
PLPLGLLPSSLDLSSPSATGRMSSLVRSLLSAVSSGGLNSSLLGSTLTSLVSQISSSRSDLSASQVLVEAV
LEILSAVIQILSSATIGVVSTDSVGATSSAVAQAVSSAFAG*
```

Figure 3.3B. *Poecilotheria regalis* spidroin amino acid sequence. Percent compositions of alanine, glycine, and serine are given in parentheses.

PRS_GlyHypothetical_var1 (A: 6.49, G: 20.45, S: 12.01)
PRS_GlyHypothetical_var2 (A: 6.47, G: 20.71, S: 11.65)
PRS_GlyHypothetical_var1:
MKTLLFFVFLESTVTTILAESNSMSCTMVNGKWSCREMKDSDGTSAFSSSVAFGGPGGGSTFEGVGTGSNSGFSWDQG
PRS_GlyHypothetical_var2:
MKTLLFFVFLGSTVTTILAESNSMSCTMVNGKWSCREMKDSDGTSAFSSSVAFGGPGGGSTFEGVGTGSNSGFSWDQG

PRS_GlyHypothetical_var1:
GTGIFSGADSPGNSGFFPGVNPIGGSGIFSNPYPYGGGVGTFSGLNPSGGVGIYPGVNPSGGGIYPGVNPSDSGIYPG
PRS_GlyHypothetical_var2:
GTGIFSGADSPGNSGFFPGVNPIGGSGIFSNPYPYGGGVGTFSGLNPSGGVGIYPGVNPSGGGIYPGVNPSDSGIYPE

PRS_GlyHypothetical_var1:
VNPSDGGIYPGVNPSDGGIYPGVNPSDGDDIFSGLNPSGGTSESEDSTPSDEADDSAIYPEVTDPPTYNPSVGGFNFG
PRS_GlyHypothetical_var2:
VNPSGGGIYPGVNPSDGGIYPGVNPSDGDDIFSGLNPSGGTNESEDSTPSDEADDSAIYPEVTDPPTYNPSVGGFNFG

PRS_GlyHypothetical_var1:
LLPWG-PRTFAASGSGVFPGVRNPFVGTAAVAGFPFGGARAFAGTPFFNPVVIPGFPLGLPFAGAGAAAYAGKRR*
PRS_GlyHypothetical_var2:
LLPWGVPRTFAASGSGVFPGVRNPFVGTAAVAGFPFGGARAFAGTPFFNPVVIPGFPPGLPFAGAGAAAYAGKRR*

PRF_SerRepetitive1 (A: 1.23, G: 6.79, S: 24.07)
(…)PTTVTETEDQSSTPEVEIGSIEPSTETDESWTTEFIPFGREQTFGTDYGNTVFQ[SSSSEPIPSIGPE]$_5$SNDVELTSGTVNEINSAEPSPSVQPESNS
FEPTFNTEAESSSS*

Figure 3.3C. *Poecilotheria regalis* silk gland novel gene amino acid sequences. Percent compositions of alanine, glycine, and serine are given in parentheses.

```
PRS_GlySer_rich_var1 (A: 8.53, G: 19.91, S: 20.85)
PRS_GlySer_rich_var2 (A: 9.33, G: 20.89, S: 19.11)
PRS_GlySer_rich_var1: MKYFLFVFFFGGAAMTLTAAISESMSCTVVNGVQRCERSRNEIGNVAAGSAAAISGGPSGPSTYQGAGTGNTNYGI
PRS_GlySer_rich_var2: MKYFLFVFFFGGAAMTLTAAISESMSCTVVNGVQRCERSRNEIGNVAAGSAAAISGGPSGPSTYQGAGTGNTNYGI

PRS_GlySer_rich_var1: YDIQQSSSGDYFRSPSGPGGSYSSSSSSDGSGPSILPILSPGWQGSSSYSSSSGSGSGYGFDFSTEPGSSSSSSSYS
PRS_GlySer_rich_var2: YDIQQSSSGDYFRSPSGPGGSYSSSSSSDGSGPSILPILSPGWQGSSSYSSSSGSRSGLG--FSTEPGSSSSSSSYS

PRS_GlySer_rich_var1: GDFSGSIPGVVLGPGVLPV-VPSF---------------GAFTGAIAGGFPGSFAGGFAGGFPFGGAVAIAGRKK*
PRS_GlySer_rich_var2: GDFSGSIPGVVVGPGVLPVGVPSFGFPGAFTGAIAGGFPGAFTGAVAGGLPGAFAGGFAGGFPFGGPVAIAGRKK*

PRS_GlyRich_var1 (A: 16.03, G: 29.77, S: 1.53)
PRS_GlyRich_var2 (A: 13.33, G: 33.33, S: 2.42)
PRS_GlyRich_var3 (A: 14.67, G: 30.67, S: 2.00)
PRS_GlyRich_var4 (A: 13.64, G: 31.17, S: 2.60)
PRS_GlyRich_var1:
MKCAFAILVLVTLTVDVQGYPSGC--GNDDCCCCCP---PIFGPYG--PYAFGGLGSPLGFGAGAGLGAGAAAGA-----------
PRS_GlyRich_var2:
MKFALAFLLLVALAADVKARGGGCNDDDDCCCPCYCCSYPWYPFYG--GYGLGGLGIGAGLGLGAGLGLGAGVGAGVGAGLGLGAG
PRS_GlyRich_var3:
MKFALACLLLVVLTVDVQASHDGCYDDDDDCCCCCCRSYPWYP-WGWPAYG-GGLGLGVGLGAGFGAGVGAGIGA------GVGAG
PRS_GlyRich_var4:
MKFALACLLLVVLTVDVQASHDGCYDDDDDCCCCCCRSYPWYP-WGWPAYG-GGLGLGVGLGAGFGAGVGAGLGAGVGA--GIGAG

PRS_GlyRich_var1:
---GVGVGVGAAIGGGVGVGVGAAAG----AGAGAGAGLGAGGGLGAG--------FPFWLP------PYYQGYGRCCGCDCCC*
PRS_GlyRich_var2:
VGAGLGAGVGAGLGAGVGVGAGVGVGLEADVGVGVGAGAGAGAGLGLGPGLGLGG—LSYGYPGLGFPCSYYGSC--CGYCRCCC*
PRS_GlyRich_var3:
VGAGAGAGAGAGVGVGVGVGAGAGAG----VGAGAGAGVGAGAGLGFG---GFGGISPYFLP----PFPYYGGCGRCCRC-CCC*
PRS_GlyRich_var4:
VGAGVGAGAGAGVGVGVGAGAGAGAG----VGVGAGVGVGAGAGLGFG---GFGGISSYFLP----PFPYYGGCGRCCRC-CCC*
```

Figure 3.3D. *Poecilotheria regalis* silk gland novel gene amino acid sequences. Percent compositions of alanine, glycine, and serine are given in parentheses.

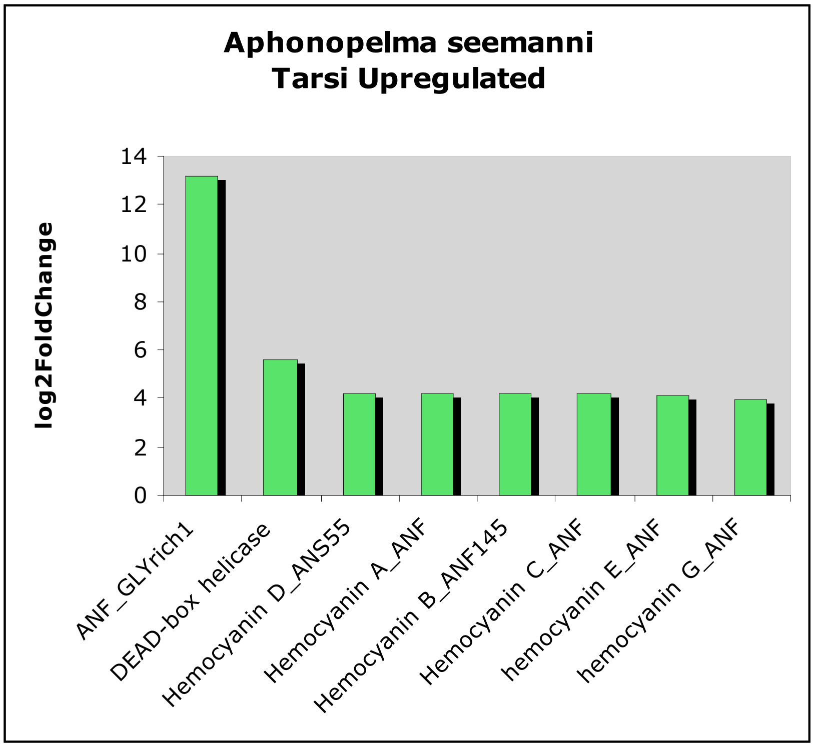


Figure 3.4A. Significantly upregulated transcripts in the tarsi of *Aphonopelma seemanni*.

```
ANF_GLYrich1 (A: 8.20, G: 40.98, S: 3.28)
MAATMIVVLLVGALLAAAPTHGVFPGGGIGYSLGGLGYGGKGLGGLGYGGLAYGGKGGGSGANQYGRGFSY
GNGFDYGNSYGAGGQGFGGNKLGGQGFGGQGHGGNGFYGGGAGGFGGVPVI*
```

Figure 3.4B. *Aphonopelma seemanni* tarsi novel gene amino acid sequence. Percent compositions of alanine, glycine, and serine are given in parentheses.

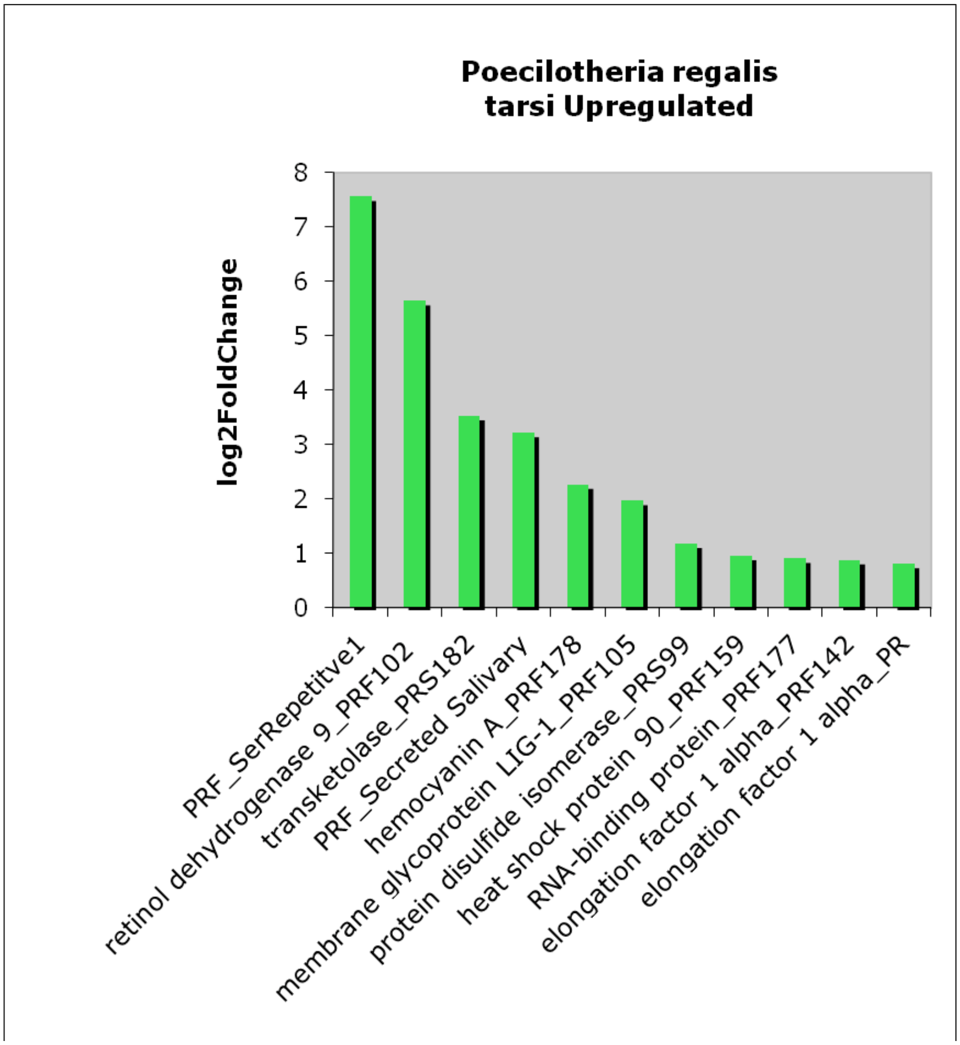


Figure 3.5. Significantly upregulated transcripts in the tarsi of *Poecilotheria regalis*.

Chapter 4

Molecular Evolution and Phylogenetic Utility of the Hemocyanin Gene Family in Spiders (Araneae), Emphasizing Evolution in the Infraorder Mygalomorphae

Abstract

Hemocyanins are multimeric copper-containing hemolymph proteins involved in oxygen binding and transport in all major arthropod lineages. Most arachnids have seven primary subunits (encoded by paralogous genes *a*-*g*), which combine to form a 24-mer (4 X 6) quaternary structure. Within some spider lineages, however, hemocyanin evolution has been a dynamic process with extensive paralog duplication and loss. We have obtained hemocyanin gene sequences from numerous representatives of the spider infraorders Mygalomorphae and Araneomorphae in order to infer the evolution of the hemocyanin gene family and test the phylogenetic utility of these conserved loci. Our hemocyanin gene tree is largely consistent with previous hypotheses of paralog relationships based on immunological studies, but reveals some discrepancies in which paralog types have been lost or duplicated in specific spider lineages. In addition, we use hemocyanin sequences to estimate higher-level spider phylogenetic relationships. Analyses of concatenated hemocyanin sequences resolved deep nodes in the tree and recovered a number of clades that are supported by other molecular studies, particularly for mygalomorph taxa. The concatenated dataset is also used to estimate dates of higher-level spider divergences and suggests that the diversification of extant mygalomorphs preceded that of extant araneomorphs. Spiders are diverse in behavior and respiratory morphology, and our results are beneficial for comparative analyses of spider respiration. Lastly, the hemocyanin gene family provides useful molecular markers for inferring spider relationships and ancient divergence dates.

Key Words: Gene duplication, Gene tree parsimony, Gene tree reconciliation, Molecular clock, Respiration, Supermatrix

Introduction

Hemocyanins are copper-containing hemolymph proteins that bind oxygen and facilitate oxygen transport in all major arthropod lineages, including Chelicerata, Myriapoda, and Pancrustacea (Burmester, 2002; 2004; Kusche and Burmester, 2001; Hagner-Holler et al., 2004; Ertas et al., 2009), as well as the arthropod relative, Onychophora (Kusche et al., 2002). Arthropod hemocyanins are part of a protein superfamily that is divided into two main molecular lineages (Burmester, 2001) – one lineage consists of arthropod phenoloxidases, involved in cuticle sclerotization, whereas the other lineage is composed of hemocyanins and proteins derived from hemocyanins (e.g., crustacean pseudohemocyanins, insect hexamerins, etc.; Figure 4.1A). Hemocyanins are hypothesized to have evolved during the Cambrian, in association with an increase in arthropod body size and a coincident need for more efficient oxygen transport (Burmester, 2001; Kusche et al., 2002).

The primary structure of arthropod hemocyanins includes multiple oxygen-binding subunits. These subunits are approximately 620-660 amino acids in length and are encoded by different paralogous members of the hemocyanin gene family (Burmester, 2001). Depending on species, six related or identical subunits combine to form either hexameric (1 X 6) or multi-hexameric (2-8 X 6) quaternary structures. Beyond these broad generalizations, it is clear that subunit diversification and evolution of quaternary structure have followed different, independent pathways within the major arthropod lineages (reviewed in Burmester, 2002).

In arachnids, immunological studies have been used to identify seven primary subunits (paralogous genes *a-g)*, which combine to form a 24-mer (4 X 6) quaternary structure. This arrangement likely represents the ancestral condition in arachnids (Markl, 1986; Markl et al., 1986). A combination of immunological, electron microscopic, and DNA sequence studies have been used to study hemocyanin evolution in the arachnid order Araneae (spiders), and again, the seven subunit and 24-mer configuration appears ancestral. For example, complementary DNA (cDNA) studies have shown that the distantly related spiders, *Eurypelma* (a tarantula) and *Nephila* (an orb-weaver), share six of the seven related hemocyanin subunits (Figure 4.1B). The subunits are genetically distant, and the estimated divergences among subunits (550-420 MYA) predate the divergence of spiders themselves (Voit et al., 2000; Averdam et al., 2003; Figure 4.1B). Immunological and electron microscopic studies conducted on a broader sample of spider families (40 species from 25 families) confirm that a 4 X 6-mer configuration is most common (Markl, 1986; Markl et al., 1986).

Despite this apparent conservation in paralog representation and quaternary structure, other evidence indicates that hemocyanin evolution has sometimes been a dynamic process within Araneae. For example, the genus *Cupiennius* (wandering spider), a distant relative of *Nephila*, reveals a dramatic shift in protein evolution (Ballweber et al., 2002). All but one subunit appears to have been lost, but this subunit (*g*) has subsequently undergone rounds of duplication to result in at least six copies (Figure 4.1B). Additional evidence suggests that similarly dramatic changes in subunit composition and quaternary structure have independently occurred in other spider lineages (Markl, 1986; Markl et al., 1986; Kuwada and Sugita, 2000).

The combination of both stasis and change seen within spider hemocyanins invites two avenues of research. First, characterization of hemocyanin gene family structure across the spider phylogeny may help reveal the factors that have led to diverse lifestyles, body sizes, and respiratory systems in different spider groups. Second, we investigate the phylogenetic utility of hemocyanin sequences with respect to relationships among major groups of Araneae. Although spiders represent one of the most diverse orders of arthropods, our understanding of spider phylogeny at many hierarchical levels remains uncertain (Coddington and Levi, 1991; Ubick et al., 2005). Additional molecular phylogenetic data can only help in this regard, and available evidence suggests that spider hemocyanins evolve slowly enough to retain deep phylogenetic signal (see Voit et al., 2000; Ballweber et al., 2002; Averdam et al., 2003).

Mygalomorph Spiders

Mygalomorph spiders, which include the tarantulas, trapdoor spiders, and other less well-known groups, represent one of three main spider lineages (Figure 4.2). Although mygalomorphs retain some features that are plesiomorphic in spiders (e.g., two pairs of book lungs), several characters support mygalomorph monophyly, and this monophyly has not been seriously questioned (see Platnick and Gertsch, 1976; Raven, 1985). Current estimates of mygalomorph diversity place roughly 2,700 species into 325 genera and 15 families (Platnick, 2011). Mygalomorphs are essentially worldwide in distribution, with centers of generic diversity in all tropical regions as well as temperate austral areas of South America, southern Africa, and Australasia (Raven, 1985; Platnick, 2011).

Inferring relationships among mygalomorph taxa has been a considerable challenge for systematists. Mygalomorphs are notorious for conservative, and often homoplastic, patterns of morphological evolution making it difficult to determine the placement of many families, as well as larger clades (Raven, 1985; Goloboff, 1993). In addition, the monophyly of many families is based on few synapomorphic morphologic characters. Recent studies based on molecular data (e.g., ribosomal DNA, Bond and Hedin, 2006; Hedin and Bond, 2006; and elongation factor-1 gamma, Ayoub et al., 2007) have contributed to a better understanding of mygalomorph relationships and contradicted monophyly of multiple families, including Cyrtaucheniidae, Dipluridae, Hexathelidae, and Ctenizidae (Figure 4.2). However, pervasive conflict among these gene trees is apparent, highlighting the need for development of additional phylogenetic markers to corroborate hypotheses of higher-level relationships.

In this paper, we sample mygalomorph and araneomorph taxa to explore patterns of hemocyanin molecular evolution and infer phylogenetic relationships in spiders. Our investigation supports hypotheses of a complex history of hemocyanin evolution with both extreme conservation and episodic lineage specific duplication and loss. Additionally, we use supermatrix and gene tree parsimony methods to investigate the phylogenetic utility of hemocyanin. Finally, using relaxed clock methods, we compare higher-level spider divergence dates inferred from hemocyanin sequences to previously hypothesized dates.

Materials and Methods

Taxon Sampling

We collected new hemocyanin gene sequences from seventeen genera representing eight mygalomorph families (Figure 4.2, Table 4.1). Collection information for most specimens is shown in Table 4.2. The families include both members of Atypoidina (a monophyletic group within Atypoidea comprised of Antrodiaetidae and Atypidae), multiple genera representing the hypothesized paraphyletic family, Dipluridae, and five families sampled broadly across the Bipectina clade (Hedin and Bond, 2006). Within Bipectina, we sampled multiple genera representing the likely non-monophyletic families, Cyrtaucheniidae and Ctenizidae. In addition, we downloaded sequences from GenBank for the theraphosids "*Eurypelma californicum*" (synonym of *Aphonopelma californicum*; Voit and Feldmaier-Fuchs, 1990; Voll and Voit, 1990; Voit et al., 2000) and *Acanthoscurria gomesiana* (Lorenzini et al., 2006).

To compare patterns of hemocyanin evolution across araneomorphs and mygalomorphs, we broadly sampled across the araneomorph phylogeny (Figure 4.2, Table 4.1; Coddington and Levi, 1991; Coddington et al., 2004). We sequenced hemocyanin gene fragments from araneomorphs representing ten genera from ten families. Specifically, we sampled the paleocribellate representative *Hypochilus*, two members of the Haplogynae (*Kukulcania* and *Diguetia*), three Araneoidea (a monophyletic group within Orbiculariae; *Gasteracantha*, *Nephila*, and *Nesticus*), and several members of the retrolateral tibial apophysis (RTA) clade (*Habronattus*, *Tengella***,** *Allocosa*, and *Cupiennius*). *Cupiennius* and *Nephila* sequences were obtained from GenBank (Ballweber et al., 2002; Averdam et al., 2003). We rooted the spider hemocyanin trees with the inferred DNA coding sequence (back-translation using MacVector 7.2 Accelrys, Inc., San Diego, CA) for the homologous region of the Aa6 hemocyanin subunit from the scorpion *Androctonus australis* (Buzy et al., 1995). The scorpion sequence does not fall within spiders for any ortholog group.

Hemocyanin Sequences

For amplification of hemocyanin genes from the taxa listed in Table 4.1, genomic DNAs were extracted using the DNeasy tissue kit (Qiagen, Valencia, CA). We amplified these DNAs with primers designed partly on the gene structure of *"Eurypelma"* subunit *e,* which is divided into nine short exons (129 - 476 bp) separated by longer (2.1 – 14.3 kbp) introns (Voll and Voit, 1990). Because of the unpredictability of intron sizes, we targeted our primers to amplify within exon 4. This exon is one of the largest exons (376 bp) and includes coding sequence for one of the two highly-conserved copper-binding sites of the hemocyanin complex (Voll and Voit, 1990; Voit et al., 2000). Sets of forward and reverse primers were designed from published *Cupiennius* (Ballweber et al., 2002) and "*Eurypelma*" sequences (Voit et al., 2000). A cocktail of forward primers (all oligonucleotide sequences are shown 5' to 3') was made by combining equimolar amounts of: TATACGACTGTGAAAGATTGTG, TGTACGACTGCGAGAGATTGTC, TGTACGACTGCGAAAGATTGTC, TATACGACTGTGAGCGATTGTC, TATACGACTCAGAACGTTTATC, TGTACGACTGCGAGAGATTATC, TGTACGACTGCGAGCGTCTCTC, TGTACGACTGCGACCGTCTGTC, TGTATGACTGTGAGAGGTTATC. Similarly, a mixture of reverse primers was made from equimolar amounts of: TTTGGAATCTTGCATCGGGATC, TCTGGAATCTGGTATCGGGATC, TTTGGAATCTGCCGTCGGGATC, TTTGGAATCTGCCATCAGGATC, TCTGGTATCTACCGTCTGGATC, TCCTGTATCTGCCATCAGCATC, TATTGAATCTGTGATCAGGGTC, TCCTGAATCTGCCATCGGGATC, TTTCATATTTGCCATCGGGATC, TCTCAAAGCTTCTGTCAGGATC. PCR amplification was done with recombinant *Taq* DNA polymerase (Invitrogen, Carlsbad, CA) using standard cycling conditions with an annealing temperature of 50˚C.

PCR products were cloned into pCR2.1-TOPO plasmids (Invitrogen, Carlsbad, CA) and electroporated into TOP10 *E. coli* cells (Invitrogen). For each species, recombinant colonies were screened in sets of 24 and ~15 colonies with inserts of the target size were sequenced. Sequencing was done with the T7 universal primer at the U.C. Riverside Genomics Core Instrumentation Facility. DNA sequencing results were edited and translated with MacVector 7.2.

Hemocyanin sequences obtained from two taxa, *Euagrus chisoseus* and *Aphonopelma seemanni*, were found in cDNA libraries that were constructed for the characterization of silk fibroin transcripts. Methods used to generate the cDNA library for

Euagrus were described in Gatesy et al. (2001). cDNA library construction methods for *Aphonopelma seemanni* followed Garb et al. (2007). Hemocyanin homologs from EST sequences were identified using BlastX against the NCBI nr protein database (Altschul et al., 1990).

DNA and amino acid sequences were aligned with Clustal W (Thompson et al., 1994) implemented in MacVector and manually adjusted. All of the sequences were the same length except for a three base insertion in putative orthologs of hemocyanin subunit *f*.

Phylogenetic Analyses

Parsimony and Bayesian phylogenetic analyses were conducted on multiple hemocyanin datasets (datasets are described below). Parsimony analyses were conducted using heuristic searches with 1000 random taxon addition replicates and tree-bisection–reconnection (TBR) branch swapping, as implemented in PAUP* 4.0b8-b10 (Swofford, 2002). Clade support was evaluated using nonparametric bootstrapping (Felsenstein, 1985), based on analyses comprising 1000 pseudoreplicates (heuristic TBR branch-swapping, ten random taxon addition replicates per pseudoreplicate).

A combination of unpartitioned and partitioned Bayesian analyses (Huelsenbeck et al., 2002; Miller et al., 2002; Huelsenbeck and Ronquist, 2001; Brandley et al., 2005) were conducted on hemocyanin matrices using MrBayes v. 3.1.2 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). Model choices were determined by Mr. Modeltest v. 2.1 (Nylander, 2004; all models selected are shown in Table 4.3) using the Akaike Information Criterion (AIC) following Posada and Buckley (2004). Two searches were run simultaneously for at least 5 million generations, with trees and parameters sampled from four MCMC chains every 100th generation. Partitions (described below) were unlinked and the substitution rates of evolution among partitions were allowed to vary. Analyses were considered complete when the standard deviation of split frequencies fell below 0.01 (Ronquist et al., 2005). The first forty percent of samples were treated as burnin and discarded. Bayes factor tests comparing harmonic means were used to determine which partitioning scheme was justified (Kass and Raftery, 1995; Brandley et al., 2005; Brown and Lemmon, 2007).

Identification of Hemocyanin Paralogs

Initially, phylogenetic analyses were conducted on a matrix consisting of all hemocyanin sequences generated. Analyses were then conducted on a reduced hemocyanin dataset that excluded PCR error and allelic variants (see results) to determine the gene tree structure (referred to as the global analysis). In the Bayesian analyses, the global data set was analyzed both as unpartitioned and partitioned by codon position. Sequences were assigned to one of the seven known subunit types by their inclusion in monophyletic groups that contained previously characterized hemocyanin sequences (see results; Figure 4.3).

Inference of Species Relationships

We also investigated the utility of hemocyanin in determining higher-level phylogenetic relationships. First, analyses were conducted on the seven ortholog groups (referred to as individual analyses), each using the scorpion sequence as an outgroup (Table 4.4). In the Bayesian analyses, individual ortholog group data sets were analyzed both as unpartitioned and partitioned by codon position. Second, we conducted analyses on a concatenated dataset, with sequences aligned based on paralog type and the scorpion sequence as the outgroup for every paralog. For both analyses, a single randomly chosen *Cupiennius* sequence (paralog *g*1) was included because all hemocyanins found in members of the RTA clade were restricted to the *g* subunit type in the global analysis (see results). In the Bayesian analyses, the concatenated dataset was analyzed as unpartitioned, partitioned by codon position (3 partitions), partitioned by paralog type (7 partitions), and partitioned by codon position and paralog type (21 partitions).

Gene Tree Parsimony

We used the program Duptree (Wehe et al., 2008) to conduct gene tree parsimony bootstrapping analyses (Page and Charleston, 1997; Cotton and Page, 2002; Slowinski and Page, 1999; Page, 2000). Duptree takes binary gene trees and uses Subtree Pruning and Regrafting with a fast heuristic algorithm to search for the species tree that requires the fewest duplication events (Wehe et al., 2008).

Searches were conducted with two different sets of input trees. For the first analysis, a codon partitioned Bayesian analysis was conducted on the global dataset that included a single representative of the RTA clade (*Cupiennius g*1). One hundred random post-burnin trees from this analysis were used as the binary input trees for Duptree. For the second analysis, one hundred random post-burnin trees from each partitioned by codon position Bayesian analysis of the individual ortholog groups were used as the binary input trees. Again, a single representative of the RTA-clade (*Cupiennius g*1) was used. Duptree analyses were conducted on both datasets to determine the sensitivity of gene tree parsimony to differences in orthologous relationships inferred from the global dataset versus the individual ortholog groups. In both analyses, the search was conducted with 1000 starting trees from a leaf adding heuristic, which were queued to start a full hill climbing heuristic. The maximum number of optimal species trees were retained (100,000 trees for the global, 100,000 for the individual) and used to generate a 50% majority rule consensus tree in PAUP for each analysis. Analyses were conducted twice and had only minor differences in resolution of one or two nodes, each with low bootstrap support.

Divergence Dates

The hemocyanin gene family is highly conserved and evolves in a nearly clocklike manner within Chelicerata (Burmester, 2001; Burmester, 2002), suggesting it may be useful for dating ancient nodes within spiders. We obtained a fully bifurcating input tree for use in molecular clock analyses by conducting a Bayesian analysis (partitioned by codon position) on the concatenated dataset excluding taxa represented by a single paralog (i.e., *Aliatypus*, *Stasimopus*, *Diguetia*, *Nesticus*, all RTA clade representatives).

In addition, the scorpion sequence was not included because outgroups are not necessary for dating with the method we utilized (below). Instead, the resulting fully bifurcating consensus tree (see results) was rooted with the remaining araneomorph spiders. This tree did not conflict with phylogenies that included all taxa.

The Baseml program in PAML v. 4.4c (Yang, 1997; Yang, 2007) was used to determine the likelihood values of a global molecular clock model and an independent rates model for the concatenated dataset. A likelihood ratio test comparing the two models rejected a global molecular clock (df=20, $p<0.001$). Therefore, Bayesian analyses using MCMCtree in the PAML package were conducted. Sequence data from the three codon positions were treated as separate partitions. The analysis was run with a model that assumes correlated rates among branches, as well as with a model that allows independent rates among branches (Yang and Rannala, 2006; Yang, 2006; Rannala and Yang, 2007). We estimated the parameters (shape and scale) of the gamma distribution of the overall substitution rate prior (μ) using Baseml estimates of the substitution rate for each partition, based on five fixed calibration points (shape=0.69, scale=5.38). Two gamma distributed priors were tested for σ^2, which specifies how variable the substitution rate is among branches (shape=1, scale=1 versus shape=1, scale=10; with the former specifying greater rate variation among branches). The HKY sequence model was used and the analysis was run with birth rate, death rate, and species sampling priors of 2, 2, and 0.1, respectively. Gamma priors for κ (the transition/transversion ratio) and α (shape parameter for among site rate variation) were left as default (Yang, 2006). Calibrations (see below) were treated as soft boundaries (i.e., 2.5% chance date falls beyond boundary; Yang and Rannala, 2006; Inoue et al., 2010). The first 20,000 iterations were discarded as burnin, followed by 50,000 iterations sampled every five iterations. Analyses were run twice to ensure MCMC convergence, with negligible differences in posterior date estimates between the two runs of the same analysis.

Five fossil calibrations were used in our analysis, three of which were used in a molecular clock analysis based on the gene encoding elongation factor-1 gamma (*EF-1γ*; Ayoub et al., 2007; Ayoub and Hayashi, 2009). *Rosamygale*, the oldest mygalomorph fossil from the mid Triassic, was used as a minimum age of 240 million years ago (MYA) for the earliest mygalomorph divergence (Selden and Gall, 1992). *Cretamygale*, a bipectin mygalomorph from the early Barremian (Cretaceous), was used as a minimum age of 130 MYA for the deepest bipectin node (Selden, 2002). Two fossil Atypoidina spiders, *Cretacattyma* (Antrodiaetidae) and *Ambiortiphagus* (Atypidae), from the upper half of the Lower Cretaceous, were used to place a minimum age of 96 MYA for the antrodiaetid-atypid split (Eskov and Zonshtein, 1990). The sole areaneomorph fossil representative used in the analysis was the nephilid, *Nephila jurassica*, from the Middle Jurassic, which placed a minimum age of 165 MYA for the nephilid-araneid split (Selden et al., 2011). A maximum age of 392 MYA for the root node in our tree was provided by *Attercopus* from the Middle Devonian (Selden et al., 1991). *Attercopus* is a member of the Uraraneida, the hypothesized sister order to spiders, and is older than all known spider fossils (Selden et al., 2008).

Results

Inference of Ortholog Groups

Our initial global hemocyanin matrix included 110 unique clone sequences, plus 22 GenBank sequences (n = 132). Heuristic parsimony analysis of this dataset (not shown) indicated that within the clearly distinguishable ortholog groups (see arguments below), multiple clones from the same species typically formed clades of closely related or near-identical sequences. We interpreted these sequences as representing either allelic variation or PCR error. For the subsequent phylogenetic analyses, we collapsed closely related sequences into a single consensus sequence using majority rule and ambiguity coding for ties, reducing the global matrix to 84 unique sequences.

Bayes factor analysis revealed that partitioning by codon position fit the data significantly better than no partitioning (Table 4.5). Both partitioned (Figure 4.3) and unpartitioned (not shown) Bayesian analyses of the global matrix yielded seven distinct clades with high posterior probabilities (PP≥0.94; Figure 4.3). These seven clades corresponded to the seven hemocyanin subunits (paralogous genes *a* – *g*) that have been well documented in *Eurypelma* (Voit et al., 2000), six of which have also been found in *Nephila* (Averdam et al., 2003; see Figure 4.1B). Each sequence clade included a subclade of mygalomorph sequences with a *Eurypelma* paralog, and a subclade of araneomorph sequences with a *Nephila* paralog (Figure 4.3). The exception is subunit *c*, which apparently has been lost in *Nephila* (Averdam et al., 2003), and was not found in any of the araneomorphs that we sampled. All hemocyanin sequences from spiders representing the RTA clade (*Habronattus*, *Tengella*, *Allocosa*, and *Cupiennius*) were restricted to the subunit *g* clade.

Relationships among the seven clades were unresolved or poorly supported with the exception of a strongly supported clade consisting of subunits *b* + *c*. The tree topology based on the parsimony analysis (Figure 4.4) was largely consistent with that of the Bayesian analysis but less resolved. Relationships that conflicted under the two different optimality criteria occurred at nodes with low support.

Species Relationships Inferred from Individual Ortholog Groups

Based on the results of the global analyses, we generated seven matrices corresponding to the groups *a*-*g* (Table 4.4). For each dataset, Bayesian analyses (Figure 4.5) with codon partitioning had a significantly better fit to the data than no partitioning, based on Bayes factor analysis (Table 4.5). For most of the ortholog groups, the topology based on the partitioned Bayesian analysis was more resolved than that under the parsimony bootstrap (Figure 4.6). Only in two ortholog groups did the Bayesian and parsimony bootstrap topologies conflict (Figures 4.5, 4.6). However, the placement of the diplurid, *Ischnothele*, within Bipectina in the Bayesian analysis of paralog *d* versus its placement as sister to Bipectina in the parsimony analysis was not well supported (PP=0.65, bootstrap percent (BP)=51). Similarly, the Bayesian analysis of paralog *f* placed araneomorphs within Mygalomorphae with low support (PP=0.76), in contrast to the parsimony analysis, which recovered a well-supported monophyletic Mygalomorphae (BP=79).

Among the Bayesian analyses of the individual ortholog groups, conflict in relationships occurred among *a*, *d*, *e*, and *f*. Conflict in relationships was mostly associated with nodes that have low support. One exception of this conflict between ortholog groups was in the placement of the diplurid, *Euagrus* (Figure 4.5). In the analysis of paralog *f*, *Euagrus* falls within a strongly supported clade of bipectin spiders (PP=0.99), whereas in paralog *e*, *Euagrus* is sister to a strongly supported clade of Bipectina (PP=1).

We also compared orthologous relationships recovered in Bayesian analyses of individual paralogs (Figure 4.5) to those in the Bayesian analysis of the global dataset (Figure 4.3). Most orthologous relationships were identical across the individual and global dataset analyses. Instances of conflict were associated with nodes with low support. One exception to this was that the Bayesian analysis of individual ortholog group *g* recovered a monophyletic Araneomorphae (PP=0.99), whereas araneomorphs were not monophyletic in the corresponding *g* clade based on the global analysis (*Hypochilus g* was sister to a mygalomorph *g* clade; PP=0.99).

Species Relationships Inferred from Concatenated Paralogs

Despite discordance among individual gene trees, Bayesian analysis of the concatenated data set produced a well-resolved tree with strong support at many nodes (Figure 4.7). Parsimony analysis of the concatenated dataset was consistent with the Bayesian analyses but resulted in a largely unresolved topology (Figure 4.8). For the Bayesian analyses, Bayes factor analysis determined that partitioning the data by codon position was most appropriate (Table 4.5). However, topologies from different analyses under unpartitioned and different partitioning schemes were consistent, only differing in degree of resolution. All analyses recovered a monophyletic Mygalomorphae with high support (PP≥0.96). Within mygalomorphs, a monophyletic Atypoidina and a monophyletic Bipectina were both always recovered with high support (PP=1). Diplurids were paraphyletic in each analysis with a clade of *Euagrus* plus *Allothele* being most closely related to Bipectina (PP≥0.91). Within the Bipectina, theraphosids formed a clade that was sister to all others. Also, consistent across analyses was a clade that included "ancylotripines" (*Homostola* and *Ancylotrypa*) allied with nemesiids (*Acanthogonatus* and *Stanwellia).*

A well-supported monophyletic Araneomorphae was recovered in Bayesian analyses for all partitioning schemes (PP≥0.98). In each analysis, the paleocribellate, *Hypochilus*, was sister to all other araneomorphs. Relationships among the remaining araneomorphs were unresolved, with the exception of a weakly supported relationship between the nephilid, *Nephila,* and araneid, *Gasteracantha*.

Species Relationships Inferred from Gene Tree Reconciliation

For the Duptree analyses of the gene trees based on the global and individual ortholog group datasets, the resulting bootstrap trees were largely unresolved (Figures 4.9, 4.10). Both 50% majority-rule trees resulted in a similar number of resolved nodes (global = 12, individual = 11). Mygalomorph monophyly was recovered in the global analysis, but the placement of members of Atypoidina was unresolved in the Duptree analysis of the tree sets from the individual ortholog groups. Both Duptree analyses recovered a non-atypoid clade consisting of members of Bipectina plus Dipluridae, although with weak support (BP≤66). Within this clade, most relationships were unresolved except for a weakly-supported clade consisting of *Ancylotrypa* plus nemesiids (global includes *Homostola*), a theraphosid clade with low support, and a diplurid clade also with low support (BP≤67 for each of the three mentioned clades). It is notable that the bootstrap proportions were similar for these latter two clades given the high support (PP=1) for theraphosid monophyly in the global and individual (*a*-*g*) ortholog analyses. By contrast, a diplurid clade was recovered only in the *a* ortholog group in the global (PP=0.87) and individual (PP=0.98) analyses. In the global analysis and ortholog group *e* individual analysis, diplurids were inferred to be polyphyletic. Furthermore, in the global analysis and ortholog group *f* individual analysis, the placement of *Thelechoris* was unresolved with respect to a clade consisting of the diplurids, *Euagrus* and *Allothele*, and a clade consisting of members of Bipectina.

Relationships among the araneomorph taxa were largely unresolved in the bootstrap trees arising from both Duptree analyses (Figures 4.9, 4.10). In the global analysis, araneomorphs formed a polytomy and only a single relationship was recovered (*Nephila* plus *Gasteracantha*). Araneomorphs were monophyletic in the Duptree analysis of the individual ortholog groups, but relationships among taxa were unresolved except for a clade of *Nephila* and *Gasteracantha*.

Molecular Clock Analysis

Bayesian estimates of mean node ages and 95% credibility intervals were similar between analyses with correlated rates among branches (Figure 4.11, Table 4.6) versus independent rates among branches (not shown; estimates were within ~9 MY difference in mean date estimates and within ~12 MY for 95% CI range size). The posterior estimates of variation in substitution rates among branches, σ^2, were ~2 times higher when the prior was set to shape=1 and scale=1 (Table 4.6) in comparison to shape=1 and scale=10 (not shown). However, this resulted in minor differences in posterior mean date estimates (within ~11 MY) and 95% CI range size (within ~13 MY). Posterior estimates of substitution rates (μ) for each of the codon positions were similar (within 0.01 substitutions/site) across analyses regardless of the priors used for σ^2 or whether rates among branches were correlated or independent (Table 4.6).

Discussion

Gene Family Evolution

The phylogenetic analysis of hemocyanin sequences from our broad taxonomic sample reveals that hemocyanin gene family evolution has been both static and dynamic in spiders. The seven clade structure recovered in our global analysis (Figure 4.3) is consistent with the inferred ancestral arachnid condition of seven subunits, as retained in the tarantula (Burmester, 2001; Markl, 1986; Voit et al., 2000). Lorenzini et al. (2006) identified a possible representative of an eighth paralog in the tarantula, *Acanthoscurria* *gomesiana*, most similar to the *f* paralog based on BLAST (Altschul et al., 1990); however, the partial transcript sequence did not overlap with our dataset and was not included. Although we did not recover all seven paralogs in any single mygalomorph taxon, our data are consistent with relative stasis of the hemocyanin gene family in this group (Markl et al., 1986). For example, all paralogs are found in at least two mygalomorph species, although subunit *b* was restricted to theraphosids. Additionally, no paralog duplicates were detected in any mygalomorph taxa. This is inconsistent with the results of Kuwada and Sugita (2000), who, based on divergent N-terminal protein sequences, reported extensive paralog duplication and loss within various mygalomorph taxa. The limited coverage of paralogs for any particular mygalomorph species in our study could imply that extensive lineage specific subunit loss has occurred. However, our focus was on sampling paralogs from a diversity of taxa rather than demonstrating absence of paralogs.

In contrast to mygalomorph spiders, our results suggest that hemocyanin evolution has been dynamic in araneomorph spiders. Consistent with previous studies based on immunoblotting and cDNA (Markl, 1986; Averdam et al., 2003), the *c* subunit appears to have been lost within Araneomorphae. Immunoblot studies have yet to be conducted on Hypochilidae, and the absence of a *Hypochilus c* subunit in our dataset may be due to incomplete sampling rather than true loss. In *Eurypelma*, the *c* subunit acts as a linker molecule with the *b* subunit and is essential in forming the 4X6 quaternary structure (Van Bruggen et al., 1980; Markl et al., 1981, Markl et al., 1982). Despite this critical role for *c* in *Eurypelma*, many araneomorphs achieve the same 4X6 quaternary structure without *c*. Given the high sequence conservation of hemocyanin paralogs, in some spider lineages it may be possible that different subunits can perform the same oxygen or subunit binding function as other subunits and compensate for paralog loss.

Spiders from the RTA clade exhibit the greatest deviation from the ancestral hemocyanin seven-subunit condition. Hemocyanin sequences from our sample of RTA clade members are consistent with those of *Cupiennius*, where all but subunit *g* appears to have been lost, and this subunit has subsequently duplicated. In the hemolymph of *Cupiennius*, hemocyanins may be found as single hexamers (1X6) or hexamer duplexes (2X6) rather than the larger 4X6 conformation observed in other spiders (Markl et al., 1976; Markl, 1980). The reduced quaternary structures may be a molecular synapomorphy for the RTA clade. Interestingly, Markl et al. (1986) showed that the haplogyne spider, *Dysdera*, exhibited a similar single hexameric hemocyanin composed of the same subunits as *Cupiennius*, and suggested that this is a complex molecular trait uniting haplogynes with the RTA clade. However, the possession of the *d* subunit rather than the *g* subunit in our haplogyne sample, *Diguetia*, indicates that this similarity in single hexameric condition is likely due to convergence.

Evolution of the hemocyanin gene family may correlate with general patterns of morphological and behavioral evolution in spiders. Most mygalomorph species are bulky and live sedentary lifestyles, occupying a burrow for the duration of their lifetime (Foelix, 1996). Mygalomorphs also exhibit considerable conservatism in respiratory morphology; all mygalomorphs retain the ancestral condition of possessing two pairs of book lungs (Raven 1985). The apparent stasis in hemocyanin evolution in mygalomorphs may reflect their conserved respiratory features and sedentary life histories.

In contrast to mygalomorphs, araneomorph spiders show considerable variation in behavior and respiratory morphology. Some araneomorph species live fossorial lifestyles similar to that of mygalomorphs. Others, such as orbicularians, expend energy to construct large aerial webs but then sit and wait until prey entangle in their webs (Foelix, 1996; Blackledge et al., 2009). In contrast, many species of the RTA clade live highly active, cursorial lifestyles (Coddington and Levi, 1991). Respiratory morphological evolution has been considerably dynamic in araneomorphs compared to mygalomorphs. With the exception of paleocribellates (e.g., *Hypochilus*), most araneomorph spiders have one pair of book lungs and the other pair has evolved into tubular tracheae that extend into the body (Foelix, 1996). The degree of tracheal branching varies from short tubes restricted to the opisthosoma to elaborately branched tracheae extending into the prosoma and extremities. The dramatic turnover of hemocyanins in araneomorphs, particularly in RTA clade spiders, may reflect the extensive diversity of behaviors and respiratory morphology in this group. Further investigation of the duplication history of hemocyanin paralogs and differences in oxygen binding and transport abilities should reveal whether there are selective advantages to the different hemocyanin structures.

Species Tree Analyses

Multi-gene families undoubtedly provide important information for inferring phylogenetic relationships, but there is ongoing debate on how best to utilize data from multiple loci to infer species relationships (Bull et al., 1993; Slowinski and Page, 1999; Bininda-Emonds, 2004; de Queiroz and Gatesy, 2006; Edwards, 2008). Few studies have compared results from supermatrix and gene tree parsimony methods applied to single data sets (Simmons and Freudenstein, 2002; Cotton and Page, 2003; McGowen et al., 2008). In our dataset of hemocyanin sequences, there is considerable conflict in relationships among most of the ortholog groups. Gene loss and/or unsampled paralogs make our taxonomic sample sizes for each locus relatively small. In addition, selection of a proper outgroup becomes challenging in this situation where relationships among paralogs are unclear and a distant taxon must be selected for rooting purposes.

Individually, the single ortholog groups used here have limited ability to accurately infer species relationships (Figures 4.5, 4.6). Hence, we generated species trees from the entire hemocyanin gene family using two different methods. Species trees resulting from analyses of the concatenated data (Figures 4.7, 4.8) differ with species trees generated from gene tree parsimony analyses (Figures 4.9, 4.10). Concatenation appears to be less sensitive to missing data and disagreement among paralogs, resulting in a more resolved tree than both gene tree parsimony analyses (i.e., from global hemocyanin trees and individual ortholog group trees). Both gene tree parsimony analyses are unable to recover a monophyletic Araneomorphae, and Mygalomorphae is only recovered in the analysis of the global trees, although with low support. However, we note that many relationships recovered from the concatenated dataset are supported by data only from single paralogs and require further corroboration. Additionally, our concatenated dataset was limited in the sense that relationships among RTA clade spiders could not be assessed due to extensive duplication of the *g* paralog, where ortholog groups could not be determined within this lineage. Given the seemingly more complex history of hemocyanin gene duplication and loss in araneomorphs, gene tree parsimony may prove more advantageous in future studies with a more thorough araneomorph taxon sample.

Based on our results, the hemocyanin gene family appears to have phylogenetic utility, particularly for mygalomorphs. The concatenated analysis supports monophyletic Atypoidina and Bipectina clades, consistent with results based on ribosomal DNA (rDNA) and morphology (Hedin and Bond, 2006; Goloboff, 1993). The family Dipluridae presents an interesting conflict between our concatenated and gene tree parsimony analyses. In concatenated analyses, diplurid representatives form a basal grade with respect to bipectin spiders as hypothesized by studies of morphology and rDNA (Goloboff, 1993; Hedin and Bond, 2006), although considerable conflict in relationships occur among these three datasets. However, diplurid taxa form a monophyletic group in our gene tree parsimony analyses, albeit with low support, and are included in a polytomy with Bipectina. This arrangement is somewhat similar to the findings in the ML and MP (with all spider taxa included) analyses of *EF-1γ* by Ayoub et al. (2007), which placed the two diplurid taxa (*Euagrus* and *Allothele*) as sister to each other and within bipectin spiders. The considerable incongruence in diplurid relationships among these various datasets clearly indicates the need for additional attention to this group. Equally challenging has been the determination of relationships among bipectin taxa. These relationships are largely unresolved based on our hemocyanin dataset, although a theraphosid clade is consistently found sister to remaining bipectins. This bears some similarity to an early diverging theraphosid group in the ML analysis of *EF-1γ* (Ayoub et al., 2007).

Molecular Dating

Molecular clock analyses encompassing both mygalomorphs and araneomorphs have thus far been based on a single nuclear gene, *EF-1γ* (Ayoub et al., 2007; Ayoub and Hayashi, 2009). Our relaxed clock analyses based on hemocyanin sequences are useful in corroborating date estimates for certain nodes, and for indicating which nodal estimates require further investigation. For most comparable nodes, 95% confidence intervals overlap between estimates based on hemocyanin and *EF-1γ*. Our divergence time estimate for the most recent common ancestor (MRCA) of Antrodiaetidae and Atypidae (node 21, Figure 4.11; Table 4.6) at approximately 120 MYA is similar to that based on *EF-1γ* (140 MYA), being older than the minimum date of 96 MYA based on the two fossil Atypoidina spiders. The MRCA of diplurid taxa and more derived mygalomorphs (node 7) is estimated at 210 MYA, which is older than the estimates based on *EF-1γ* (190 MYA from MP, Ayoub and Hayashi, 2009; ~150 MYA based on ML, Ayoub et al., 2007). Divergence date estimates for the split between Araneomorphae and Mygalomorphae (node 1) are similar for hemocyanin (380 MYA) and *EF-1γ* (392 MYA).

A notable discrepancy between hemocyanin and *EF-1γ* occurs in the initial divergence dates among extant members within Mygalomorphae and Araneomorphae. Based on *EF-1γ*, Ayoub et al. (2007) found the age of the MRCA of Mygalomorphae

(~295 MYA) to be much more recent than that of Araneomorphae (~375 MYA). This contrasts with hypotheses based strictly on fossils, which predict that both suborders originated ~240 MYA, coincident with diversification of extant mygalomorphs and diversification of extant araneomorphs occurring later by ~180 MYA (Penney and Selden, 2007; Selden and Penny, 2010). The relative divergence dates of the two suborders based on hemocyanin are more similar to these fossil-based hypotheses in that the mygalomorph MRCA (node 5: ~330 MYA) occurred prior to the araneomorph MRCA (node 2: ~295 MYA), even though the calibration for orbicularians was ~30 million years older in our study compared to the calibration used for *EF-1γ*. However, confidence intervals for the initial mygalomorph and araneomorph divergences do overlap between *EF-1γ* and hemocyanin. Using multiple nuclear and mitochondrial loci, Dimitrov et al. (2011) estimated the initial araneomorph divergence date at ~325 MYA, which was also older than our hemocyanin estimates. Mygalomorph divergences were not estimated in their study. The discrepancy in dates between hemocyanin, *EF-1γ*, and the multigene dataset may be due to our sampling of paralogs, but may also reflect lineage specific rate variation between the different genes used in each study.

Conclusions

Hemocyanins play a crucial role in oxygen storage and transport in spiders, and this is reflected in the high sequence conservation in hemocyanin paralogs across distantly related species. However, hemocyanin gene family evolution has not always been static in spiders as reflected by the apparent loss of all paralog types but *g*, followed by extensive duplication of this paralog in all members of the RTA clade. Although causal mechanisms remain unknown, the diversification history of spider hemocyanins may correspond to the variation in respiratory demand across spider lineages associated with their different body sizes, activity levels, and respiratory structures.

The conservation of hemocyanin sequences makes this gene family very useful for resolving ancient nodes in spider phylogeny, which has been a considerable challenge due to a lack of available molecular markers to date. In spite of incongruence in relationships inferred from the different ortholog groups, phylogenetic analysis of the concatenated paralog data set produced a well-resolved tree with strong support values for many nodes that is largely consistent with other recent molecular studies, particularly for mygalomorphs. Lastly, hemocyanin should be valuable for estimating ancient divergence dates in spiders. Our molecular clock estimates suggest that the initial divergences of mygalomorphs and araneomorphs occurred fairly close together in time (~300 MYA), with the initial mygalomorph divergence occurring only ~40 million years before the initial diversification of araneomorphs.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J. Mol. Biol. 215: 403-410.

Averdam A, Markl J, Burmester T (2003) Subunit sequences of the 4 X 6-mer hemocyanin from the golden orb-web spider, *Nephila inaurata*. Eur. J. Biochem. 270: 3432-3439.

Ayoub NA, Hayashi CY (2009) Spiders (Araneae). in: Hedges SB, Kumar S (Eds), The Timetree of Life. Oxford University Press, New York, pp. 255-259.

Ayoub NA, Garb JE, Hedin MC, Hayashi CY (2007) Utility of the nuclear protein-coding gene, elongation factor-1 gamma (*EF-1γ*), for spider systematics, emphasizing family level relationships of tarantulas and their kin (Araneae: Mygalomorphae). Mol. Phylogenet. Evol. 42: 394-409.

Ballweber P, Markl J, Burmester T (2002) Complete hemocyanin subunit sequences of the hunting spider *Cupiennius salei*. J. Biological Chemistry. 277: 14451-14457.

Bininda-Emonds ORP (2004) The evolution of supertrees. TREE. 19: 315-322.

Blackledge TA, Scharff N, Coddington JA, Szüts T, Wenzel JW, Hayashi CY, Agnarsson I (2009) Reconstructing web evolution and spider diversification in the molecular era. Proc. Nat. Acad. Sci. U.S.A. 106: 5229-5234.

Bond JA, Hedin MC (2006) A total evidence assessment of the phylogeny of North America euctenizine trapdoor spiders (Araneae, Mygalomorphae, Cyrtaucheniidae) using Bayesian inference. Mol. Phylogenet. Evol. 41: 70-85.

Brandley MC, Schmitz A, Reeder TW (2005) Partitioned Bayesian Analyses, Partition Choice, and the Phylogenetic Relationships of Scincid Lizards. Syst. Biol. 54: 373-390.

Brown JM, Lemmon AR (2007) The Importance of Data Partitioning and the Utility of Bayes Factors in Bayesian Phylogenetics. Syst. Biol. 56: 643-655.

Bull JJ, Huelsenbeck JP, Cunningham CW, Swofford DL, Waddell PJ (1993) Partitioning and Combining Data in Phylogenetic Analysis. Syst. Biol. 42: 384-397.

Burmester T (2001) Molecular evolution of the arthropod hemocyanin superfamily. Mol. Biol. Evol. 18: 184-195.

Burmester T (2002) Origin and evolution of arthropod hemocyanins and related proteins. J Comp Physiol B. 172: 95-107.

Burmester T (2004) Evolutionary history and diversity of arthropod hemocyanins. Micron. 35: 121-122.

Buzy A, Gagnon J, Lamy L, Thibault P, Forest E, Hudry-Clergeon G (1995) Complete amino acid sequence of the Aa6 subunit of the scorpion *Androctonus australis* hemocyanin determined by Edman degradation and mass spectrometry. Eur. J. Biochem. 233: 93-101.

Coddington JA, Giribet G, Harvey MS, Prendini L, Walter DE (2004) Arachnida. in: Cracraft J, Donoghue M (Eds.), Assembling the Tree of Life. Oxford University Press, New York, pp. 296-318.

Coddington JA, Levi HW (1991) Systematics and Evolution of Spiders (Araneae). Annu. Rev. Ecol. Syst. 22: 565-592.

Cotton JA, Page RDM (2002) Going nuclear: gene family evolution and vertebrate phylogeny reconciled. Proc. R. Soc. Lond. B. 269: 1555-1561.

Cotton JA, Page RDM (2003) Gene tree parsimony vs. uninode coding for phylogenetic reconstruction. Mol. Phylogenet. Evol. 29: 298-308.

de Queiroz A, Gatesy J (2006) The supermatrix approach to systematics. TREE. 22: 34-41.

Dimitrov D, Lopardo L, Giribet G, Arnedo MA, Álvarez-Padilla F, Horminga G (2011) Tangled in a sparse spider web: single origin of orb weavers and their spinning work unraveled by denser taxonomic sampling. Proc. R. Soc. B. (doi: 10.1098/rspb.2011.2011).

Edwards SV (2008) Is a New and General Theory of Molecular Systematics Emerging? Evolution. 63: 1-19.

Ergas B, von Reumont BM, Wägele J-W, Misof B, Burmester T (2009) Hemocyanin Suggests a Close Relationship of Remipedia and Hexapoda. Mol. Biol. Evol. 26: 2711-2718.

Eskov K, Zonshtein S (1990) First Mesozoic mygalomorph spiders from the Lower Cretaceous of Siberia and Mongolia, with notes on the system and evolution of the infraorder Mygalomorphae (Chelicerate: Araneae). N. Jb. Geol. Palaont. Abh. 178: 325-368.

Felsenstein J (1985) Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution. 39: 783-791.

Foelix RF (1996) Biology of Spiders. 2nd Ed. Oxford University Press, New York.

Garb JE, DiMauro T, Lewis RV, Hayashi CY (2007) Expansion and Intragenic Homogenization of Spider Silk Genes since the Triassic: Evidence from Mygalomorphae (Tarantulas and Their Kin) Spidroins. Mol. Biol. Evol. 24: 2454-2464.

Gatesy J, Hayashi CY, Motriuk D, Woods J, Lewis RV (2001) Extreme Diversity, Conservation, and Convergence of Spider Silk Fibroin Sequences. Science. 291: 2603-2605.

Goloboff PA (1993) A reanalysis of mygalomorph spider families (Araneae). Amer. Mus. Novitates. 3056, 32 pp.

Hagner-Holler S, Schoen A, Erker W, Marden JH, Rupprecht R, Decker H, Burmester T (2004) A respiratory hemocyanin from an insect. Proc. Natl. Acad. Sci. U.S.A. 101: 871-874.

Hedin MC, Bond JE (2006) Molecular phylogenetics of the spider infraorder Mygalomorphae using nuclear rRNA genes (18S and 28S): Conflict and agreement with the current system of classification. Mol. Phylogenet. Evol. 41: 454-471.

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 17: 754-755.

Huelsenbeck JP, Larget B, Miller RE, Ronquist F (2002) Potential applications and pitfalls of Bayesian inference of phylogeny. Syst. Biol. 51: 673-688.

Inoue J, Donoghue PCJ, Yang Z (2010) The Impact of the Representation of Fossil Calibrations on Bayesian Estimation of Species Divergence Times. Syst. Biol. 59: 74-89.

Kass RE, Raftery AE, (1995) Bayes factors. J. Am. Stat. Assoc. 90: 773-795.

Kusche K, Ruhberg H, Burmester T (2002) A hemocyanin from the Onychophora and the emergence of respiratory proteins. Proc. Natl. Acad. Sci. U.S.A. 99: 10545-10548.

Kusche K, Burmester T (2001) Diplopod Hemocyanin Sequence and the Phylogenetic Position of the Myriopoda. Mol. Biol. Evol. 18: 1566-1573.

Kuwada T, Sugita H (2000) Evolution of hemocyanin subunits in mygalomorph spiders: distribution of hemocyanin subunits and higher classification of the Mygalomorphae. Zoological Science. 17: 517-525.

Lorenzini DM, da Silva Jr PI, Soares MB, Arruda P, Setubal J, Daffre S (2006) Discovery of immune-related genes expressed in hemocytes of the tarantula spider *Acanthoscurria gomesiana*. Developmental and Comparative Immunology. 30: 545-556.

Markl J (1980) Hemocyanins in Spiders, XI. The Quaternary Structure of Cupiennius Hemocyanin. J. Comp. Physiol. 140: 199-207.

Markl J (1986) Evolution and function of structurally diverse subunits in the respiratory protein hemocyanin from arthropods. Biol. Bull. (Woods Hole, Mass). 171: 90-115.

Markl J, Kempter B, Linzen B, Biljholt MMC, van Bruggen EFJ (1981) Hemocyanins in Spiders, XVI. Subunit Topography and a Model of the Quaternary Structure of *Eurypelma* Hemocyanin. Hoppe-Seyler's Z. Physiol. Chem. 362: 1631-1641.

Markl J, Decker H, Linzen B, Schutter WG, van Bruggen EFJ (1982) Hemocyanins in Spiders, XV. The Role of Individual Subunits in the Assembly of *Eurypelma* Hemocyanin. Hoppe-Seyler's Z. Physiol. Chem. 363: 73-87.

Markl J, Schmid R, Czichos-Tiedt S, Linzen B (1976) Haemocyanins in Spiders, III. Chemical and Physical Properties of the Proteins in *Dugesiella* and *Cupiennius* Blood. Hoppe-Seyler's Z. Physiol. Chem. 357: 1713-1725.

Markl J, Stöcker W, Runzler R, Precht E (1986) Immunological correspondences between the hemocyanin subunits of 86 arthropods: evolution of a multigene protein family. in: Linzen, B. (Ed), Invertebrate Oxygen Carriers. Springer, Berlin Heidelberg New York, pp 281-292.

McGowen MR, Clark C, Gatesy J (2008) The Vestigial Olfactory Receptor Subgenome of Odontocete Whales: Phylogenetic Congruence between Gene-Tree Reconciliation and Supermatrix Methods. Syst. Biol. 57: 574-590.

Miller RE, Buckley TR, Manos PS (2002) An examination of the monophyly of morning glory taxa using Bayesian phylogenetic inference. Syst. Biol. 51: 740-753.

Nylander JAA (2004) MrModeltest v2. Program distributed by the author. Evolutionary Biology Centre, Uppsala University.

Page RDM (2000) Extracting Species Trees From Complex Gene Trees: Reconciled Trees And Vertebrate Phylogeny. Mol. Phylogenet. Evol. 14: 89-106.

Page RDM, Charleston MA (1997) From Gene to Organismal Phylogeny: Reconciled Trees and the Gene Tree/Species Tree Problem. Mol. Phylogenet. Evol. 7: 231-240.

Penny D, Selden PA (2007) Spinning with the dinosaurs: the fossil record of spiders. Geology Today. 23: 231-237.

Platnick NI (2011) The world spider catalog, version 12.0. American Museum of Natural History, online at http://research.amnh.org/iz/spiders/catalog. DOI: 10.5531/db.iz.0001. Accessed 8 August, 2011.

Platnick NI, Gertsch WJ (1976) The suborders of spiders: a cladistic analysis (Arachnida, Araneae). Amer. Mus. Novitates. No. 2607, pp. 1-15.

Posada D, Buckley TR (2004) Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches over Likelihood Ratio Tests. Syst. Biol. 53: 793-808.

Rannala B, Yang Z (2007) Inferring Speciation Times under an Episodic Molecular Clock. Syst. Biol. 56: 453-466.

Raven RJ (1985) The spider infraorder Mygalomorphae (Araneae): Cladistics and systematics. Bull. Am. Mus. Nat. Hist. V. 182, pp 1-180.

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19: 1572-1574.

Ronquist F, Huelsenbeck JP, van der Mark P (2005) mrbayes 3.1 manual, draft 5/26/2005, online at http://mrbayes.csit.fsu.edu/manual.php.

Selden PA (2002) First British Mesozoic spider, from Cretaceous amber of the Isle of Wight, southern England. Palaeontology. 45: 973-983.

Selden PA, Penny D (2010) Fossil Spiders. Biol. Rev. 85: 171-206.

Selden PA, Shih C, Ren D (2011) A golden orb-weaver spider (Araneae: Nephilidae: *Nephila*) from the Middle Jurassic of China. Biol. Lett. 7: 775-778.

Selden PA, Gall J-C (1992) A Triassic mygalomorph spider from the northern Vosges, France. Palaeontology. 35: 211-235.

Selden PA, Shear WA, Bonamo PM (1991) A spider and other arachnids from the Devonian of New York, and reinterpretation of Devonian Araneae. Palaeontology. 34: 241-281.

Selden PA, Shear WA, Sutton MA (2008) Fossil evidence for the origin of spider spinnerets, and a proposed arachnid order. Proc. Natl. Acad. Sci. U.S.A. 105: 20781-20785.

Simmons MP, Freudenstein JV (2002) Uninode coding vs. gene tree parsimony for phylogenetic reconstruction using duplicate genes. Mol. Phyl. Evol. 23: 481-498.

Slowinski JB, Page RDM (1999) How Should Species Phylogenies Be Inferred from Sequence Data? Syst. Biol. 48: 814-825.

Swofford DL (2002) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates; Sunderland, Massachusetts.

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22: 4673-4680.

Ubick D, Paquin P, Cushing PE, Roth V (Eds), (2005) Spiders of North America: an identification manual. American Arachnological Society. 377 pages.

Voit R, Feldmaier-Fuchs G (1990) Arthropod hemocyanins. Molecular cloning and sequencing of cDNAs encoding the tarantula hemocyanin subunits a and e. J. Biol. Chem. 265: 19447-19452.

Van Bruggen EFJ, Bijlholt MMC, Schutter WG, Wichertjes T (1980) The Role of Structurally Diverse Subunits in the Assembly of Three Cheliceratan Hemocyanins. FEBS Letters. 116: 207-210.

Voit R, Feldmaier-Fuchs G, Schweikardt T, Decker H, Burmester T (2000) Complete sequence of the 24-mer hemocyanin of the tarantula *Eurypelma californicum.* The Journal of Biological Chemistry. 275: 39339-39344.

Voll W, Voit R (1990) Characterization of the gene encoding the hemocyanin subunit e from the tarantula Eurypelma californicum. Proc. Natl. Acad. Sci. U.S.A. 87: 5312-5316.

Wehe A, Bansal MS, Burleigh JG, Eulenstein O (2008) DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. Bioinformatics. 24: 1540-1541.

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13: 555-556.

Yang Z (2006) Computational Molecular Evolution. Oxford University Press, New York.

Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. 24: 1586-1591.

Yang Z, Rannala B (2006) Bayesian Estimation of Species Divergence Times Under a Molecular Clock Using Multiple Fossil Calibrations with Soft Bounds. Mol. Biol. Evol. 23: 212-226.

Tables

Table 4.1. Taxon sample. MY indicates Mygalomorphae, AR indicates Araneomorphae.

Infraorder: Family	Taxon	GenBank #s
MY: Atypidae	*Sphodros abboti* Walckenaer, 1835	
MY: Antrodiaetidae	*Aliatypus californicus* (Banks, 1896)	
MY: Antrodiaetidae	*Antrodiaetus gertschi* (Coyle, 1968)	
MY: Dipluridae	*Allothele australis* (Purcell, 1903)	
MY: Dipluridae	*Thelechoris striatipes* (Simon, 1889)	
MY: Dipluridae	*Euagrus chisoseus* Gertsch, 1939	
MY: Dipluridae	*Ischnothele reggae* Coyle & Meigs, 1990	
MY: Idiopidae	*Ctenolophus oomi* Hewitt, 1913	
MY: Ctenizidae	*Bothriocyrtum californicum* (O. P.-Cambridge, in Moggridge, 1874)	
MY: Ctenizidae	*Stasimopus sp.* Simon, 1892	
MY: Ctenizidae	*Ummidia sp.* Thorell, 1875	
MY: Cyrtaucheniidae	*Apomastus schlingeri* Bond & Opell, 2002	
MY: Cyrtaucheniidae	*Homostola pardalina* (Hewitt, 1913)	
MY: Cyrtaucheniidae	*Ancylotrypa barbertoni* (Hewitt, 1913)	
MY: Nemesiidae	*Acanthogonatus campanae* (Legendre & Calderón, 1984)	
MY: Nemesiidae	*Stanwellia sp.* Rainbow & Pulleine, 1918	
MY: Theraphosidae	*Aphonopelma reversum* Chamberlin, 1940	
MY: Theraphosidae	*Aphonopelma seemanni* (Cambridge, 1897)	
MY: Theraphosidae	*"Eurypelma californicum"*	X16893, AJ290429, AJ277489, AJ290430, X16653, AJ277491, AJ277492
MY: Theraphosidae	*Acanthoscurria gomesiana* Mello-Leitão, 1923	DR444412, DR443840
AR: Hypochilidae	*Hypochilus gertschi* Hoffman, 1963	
AR: Diguetidae	*Diguetia mojavea* Gertsch, 1958	
AR: Filistatidae	*Kukulcania hibernalis* (Hentz, 1842)	
AR: Araneidae	*Gasteracantha cancriformis* (Linnaeus, 1758)	
AR: Nephilidae	*Nephila inaurata* (Walckenaer, 1842)	AJ547807, AJ547808, AJ547809, AJ547810, AJ547811, AJ547812
AR: Nesticidae	*Nesticus tennesseensis* (Petrunkevitch, 1925)	
AR: Salticidae	*Habronattus ustulatus* (Griswold, 1979)	
AR: Tengellidae	*Tengella radiata* (Kulczyn´ski, 1909)	
AR: Lycosidae	*Allocosa sp.* (Moenkhausiana group) Banks, 1900	
AR: Ctenidae	*Cupiennius salei* (Keyserling, 1877)	AJ307903, AJ307904, AJ307905, AJ307906, AJ307907, AJ307909

Table 4.2. Taxon sample with locality data and voucher information.

Infraorder: Family	Taxon	Locality Data	Voucher Info
MY: Atypidae	*Sphodros abboti* Walckenaer, 1835	USA: FL: Alachua Co., Gainesville	MY 26
MY: Antrodiaetidae	*Aliatypus californicus* (Banks, 1896)	USA: CA: Mariposa Co., vic. Mariposa	MY 109
MY: Antrodiaetidae	*Antrodiaetus gertschi* (Coyle, 1968)	USA: CA: Shasta Co., NW of Old Station	MY 432
MY: Dipluridae	*Allothele australis* (Purcell, 1903)	RSA: East Cape, S Alicedale	MY 162
MY: Dipluridae	*Euagrus chisoseus* Gertsch, 1939	USA: AZ: Santa Cruz Co., Santa Rita Mtns	
MY: Dipluridae	*Ischnothele reggae* Coyle & Meigs, 1990	Jamaica: St. Andrew Parish	MY 318
MY: Idiopidae	*Ctenolophus oomi* Hewitt, 1913	RSA: Mpumalanga Province, Songimvelo Nature Preserve	MY 321
MY: Ctenizidae	*Bothriocyrtum californicum* (O. P.-Cambridge, in Moggridge, 1874)	USA: CA: San Diego Co., San Diego	MY 66
MY: Ctenizidae	*Stasimopus sp.* Simon, 1892	RSA: Northern Province, Rust de Winter	MY 161
MY: Ctenizidae	*Ummidia sp.* Thorell, 1875	USA: AZ: Yavapai Co., Ponderosa Park	MY 149
MY: Cyrtaucheniidae	*Apomastus schlingeri* Bond & Opell, 2002	USA: CA: Los Angeles Co., Monrovia Canyon County Park	MY 228
MY: Cyrtaucheniidae	*Homostola pardalina* (Hewitt, 1913)	RSA: Mpumalanga Province, Songimvelo Nature Preserve	MY 314
MY: Cyrtaucheniidae	*Ancylotrypa barbertoni* (Hewitt, 1913)	RSA: Mpumalanga Province, Songimvelo Nature Preserve	MY 316
MY: Nemesiidae	*Acanthogonatus campanae* (Legendre & Calderón, 1984)	Chile	Platnick
MY: Nemesiidae	*Stanwellia sp.* Rainbow & Pulleine, 1918	AUS:Victoria, between Bairnsdale and Orbosi	Cokendo lpher
MY: Theraphosidae	*Aphonopelma reversum* Chamberlin, 1940	USA: CA: San Diego Co., Marron Valley	MY 63
AR: Hypochilidae	*Hypochilus gertschi* Hoffman,1963	USA: KY: Letcher Co., S Whitesburg	H 237
AR: Diguetidae	*Diguetia mojavea* Gertsch, 1958	USA: CA: Imperial Co., Picacho Road	G 126
AR: Filistatidae	*Kukulcania hibernalis* (Hentz, 1842)	USA: NM: Lea Co., 20 mi. WSW of Hobbs	Platnick
AR: Araneidae	*Gasteracantha cancriformis* (Linnaeus, 1758)	USA: FL: Alachua Co., Gainesville	
AR: Nesticidae	*Nesticus tennesseensis* (Petrunkevitch, 1925)	USA: TN; Grainger Co., Indian Cave	N 58

AR: Salticidae	*Habronattus ustulatus* (Griswold, 1979)	USA: OR: Lake Co., Summer Lake	HA 1434
AR: Tengellidae	*Tengella radiata* (Kulczyn´ski, 1909)	Costa Rica, La Selva Biological Station	
AR: Lycosidae	*Allocosa sp.* (Moenkhausiana group) Banks, 1900	Chile	Platnick

Note - MY, H, G, N, HA voucher codes in MH collection.

Table 4.3. Model selection in MrModeltest

Dataset	**1st Pos**	**2nd Pos**	**3rd Pos**	**Unpartitioned**
Global	*GTR +I+γ*	*GTR +I+γ*	*GTR +I+γ*	*GTR +I+γ*
a	*GTR +I*	*GTR +γ*	*GTR +γ*	*SYM +I +γ*
b	*F81 +I*	*F81*	*GTR*	*SYM +I*
c	*HKY +γ*	*GTR*	*HKY*	*SYM +I*
d	*GTR +I*	*GTR +γ*	*GTR +I*	*SYM +γ*
e	*SYM +I*	*GTR +γ*	*GTR +γ*	*SYM +I +γ*
f	*GTR +γ*	*GTR +γ*	*GTR +γ*	*GTR +I +γ*
g[1]	*GTR +I*	*F81 +I*	*GTR +γ*	*SYM +γ*
Concatenated[1]	*GTR +I+γ*	*GTR +I+γ*	*GTR +γ*	*GTR +I +γ*
[1] *Cupiennius 1g* used to represent RTA clade.				

Table 4.4. Paralogs used in individual ortholog group, gene tree parsimony bootstrapping, and concatenated analyses.

Infraorder:Family	Taxon	*a*	*b*	*c*	*d*	*e*	*f*	*g*
MY: Atypidae	*Sphodros abboti*	●		●		●		
MY: Antrodiaetidae	*Aliatypus californicus*	●						
MY: Antrodiaetidae	*Antrodiaetus gertschi*	●					●	
MY: Dipluridae	*Allothele australis*	●					●	
MY: Dipluridae	*Thelechoris striatipes*	●				●	●	
MY: Dipluridae	*Euagrus chisoseus*	●				●	●	●
MY: Dipluridae	*Ischnothele reggae*	●			●			
MY: Idiopidae	*Ctenolophus oomi*	●				●	●	
MY: Ctenizidae	*Bothriocyrtum californicum*	●		●	●		●	
MY: Ctenizidae	*Stasimopus sp.*	●						
MY: Ctenizidae	*Ummidia sp.*	●		●				
MY: Cyrtaucheniidae	*Apomastus schlingeri*	●					●	
MY: Cyrtaucheniidae	*Homostola pardalina*	●					●	
MY: Cyrtaucheniidae	*Ancylotrypa barbertoni*					●	●	
MY: Nemesiidae	*Acanthogonatus campanae*	●					●	
MY: Nemesiidae	*Stanwellia*				●		●	
MY: Theraphosidae	*Aphonopelma reversum*			●	●	●		
MY: Theraphosidae	*Aphonopelma seemanni*		●	●	●	●		●
MY: Theraphosidae	*"Eurypelma californicum"*	●	●	●	●	●	●	●
MY: Theraphosidae	*Acanthoscurria gomesiana*			●			●	
AR: Hypochilidae	*Hypochilus gertschi*	●						●
AR: Diguetidae	*Diguetia mojavea*				●			
AR: Filistatidae	*Kukulcania hibernalis*	●				●		
AR: Araneidae	*Gasteracantha cancriformis*	●						●
AR: Nesticidae	*Nesticus tennessensis*						●	
AR: Nephilidae	*Nephila inaurata*	●	●		●	●	●	●
AR: Ctenidae	*Cupiennius salei*							●

Table 4.5. Comparison of partitioning schemes was tested with Bayes factors (BF), which were based on harmonic means (HM) from Bayesian analyses.

Dataset	HM: Partitioned By Codon	HM: Unpartitioned	HM: Partitioned By Gene & Codon	HM: Partitioned By gene	BF
Global	-14212.28	-14494.00	--	--	563.44
a	-3425.89	-3583.00	--	--	316.22
b	-1062.35	-1097.25	--	--	69.8
c	-1480.28	-1548.81	--	--	137.06
d	-1735.12	-1816.08	--	--	161.92
e	-2323.09	-2423.79	--	--	201.4
f	-3026.33	-3171.54	--	--	290.42
g	-1811.40	-1905.18	--	--	187.56
Concatenated[1]	*-14791.08*	-15532.67	*-14829.95*	-15541.61	77.74

[1] BF for concatenated calculated from HM values in italics.

Table 4.6. Divergence date estimates based on the MCMCtree (σ^2 1, 1) correlated rates analysis of the codon partitioned concatenated dataset. Source tree and node numbers are in Figure 4.11.

Node	Soft Bounds (MYA)	Posterior Mean (95% CI)
1	<392	381.34 (329.95, 423.65)
2		294.30 (232.18, 365.15)
3		255.71 (194.92, 322.95)
4	>165	170.72 (136.75, 211.86)
5	>240	332.42 (270.81, 393.70)
6		237.66 (188.77, 297.90)
7		209.85 (166.55, 264.08)
8	>130	158.33 (126.56, 205.75)
9		146.90 (114.49, 191.86)
10		130.15 (99.33, 171.37)
11		123.91 (92.81, 164.64)
12		109.83 (75.90, 151.90)
13		67.26 (38.98, 104.10)
14		21.25 (8.16, 39.50)
15		111.95 (73.05, 156.98)
16		17.94 (10.31, 31.73)
17		13.70 (7.49, 22.56)
18		6.42 (1.93, 14.13)
19		136.65 (87.92, 190.91)
20		121.94 (68.31, 175.16)
21	>96	117.76 (91.60, 172.38)
Rates		
μ_1		0.0415 (0.0268, 0.0678)
μ_2		0.0204 (0.0097, 0.0434)
μ_3		0.2315 (0.1758, 0.3071)
σ^2_1		0.1267 (0.0027, 0.4722)
σ^2_2		0.5280 (0.0962, 1.4202)
σ^2_3		0.0717 (0.0081, 0.2303)

Figures

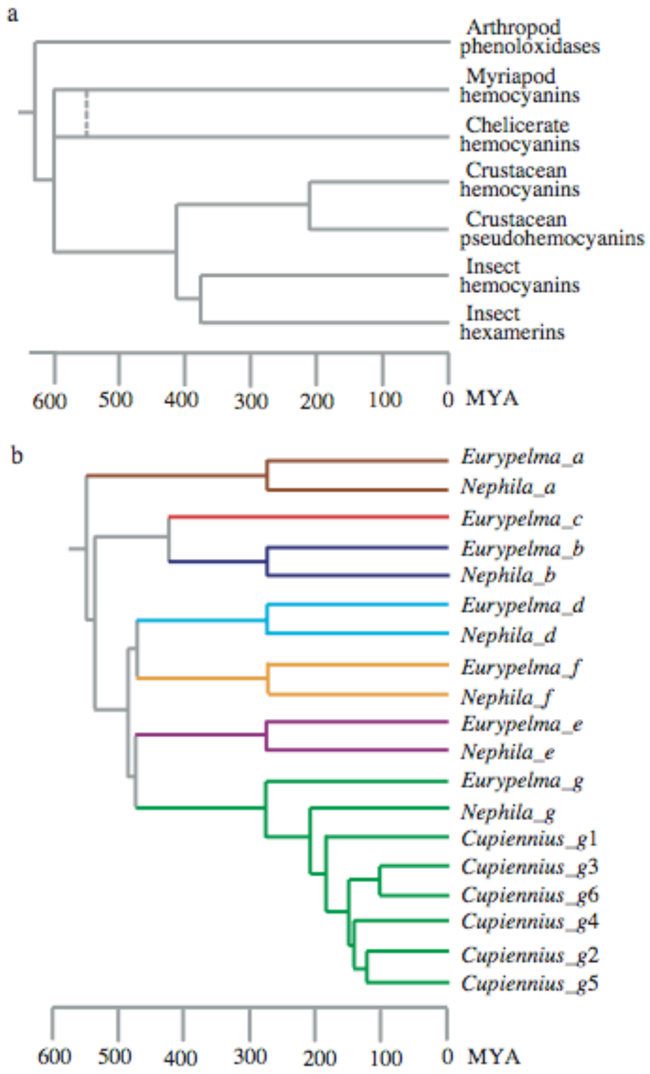


Figure 4.1. (A) Evolutionary relationships of proteins in the arthropod hemocyanin superfamily, including estimated times of molecular divergence (adapted from Figure 4.22 of Burmester, 2002). (B) Evolutionary relationships of hemocyanin subunits in spiders, including estimated times of molecular divergence (adapted from Figure 4.11 of Averdam et al., 2003).

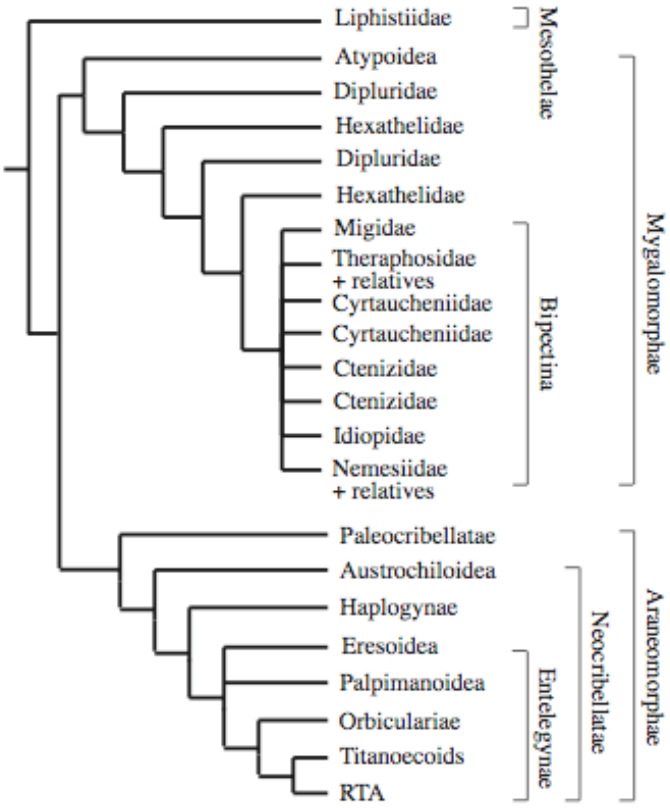


Figure 4.2. Phylogeny of spiders (Araneae) based on the hypotheses of Coddington et al. (2004), Hedin and Bond (2006), and Ayoub et al. (2007).

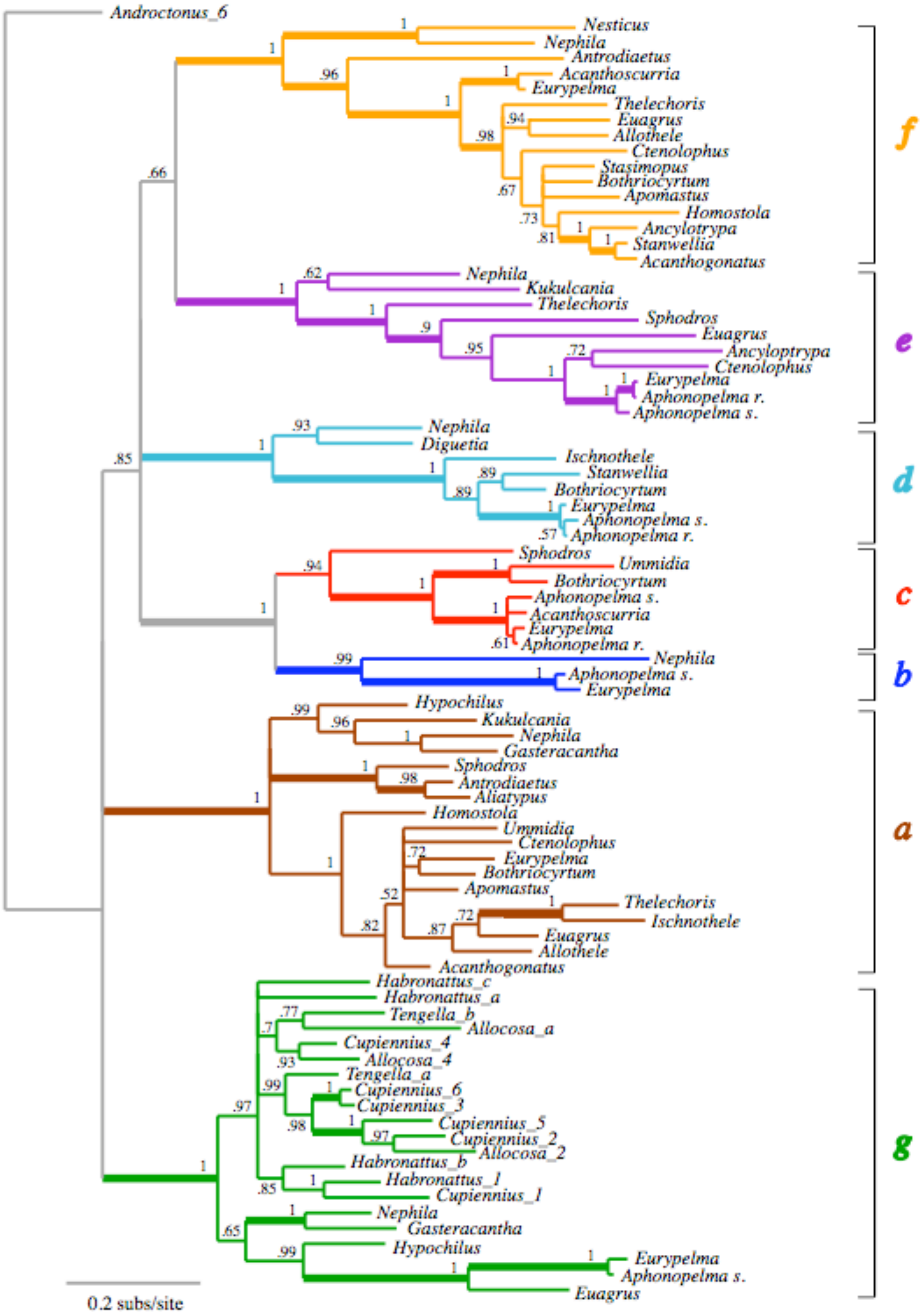


Figure 4.3. Bayesian consensus phylogram based on a codon partitioned analysis of the global hemocyanin data set. Posterior probabilities >0.5 are adjacent to nodes. Branches that are thickened indicate nodes that also have >70% parsimony bootstrap support.

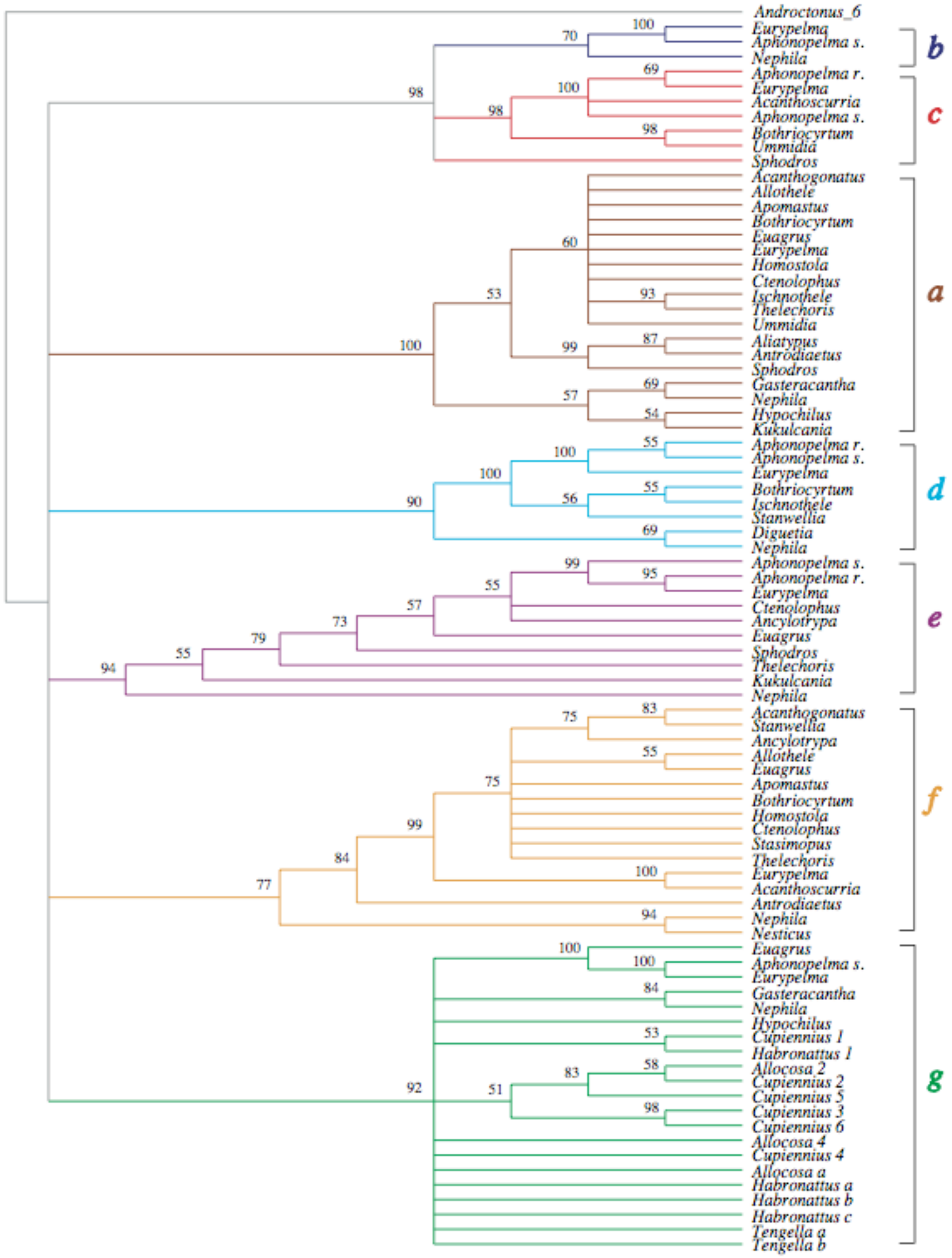


Figure 4.4. 50% majority rule consensus tree from parsimony bootstrap analysis of the global dataset. Nodes with bootstrap percentages >50 are shown.

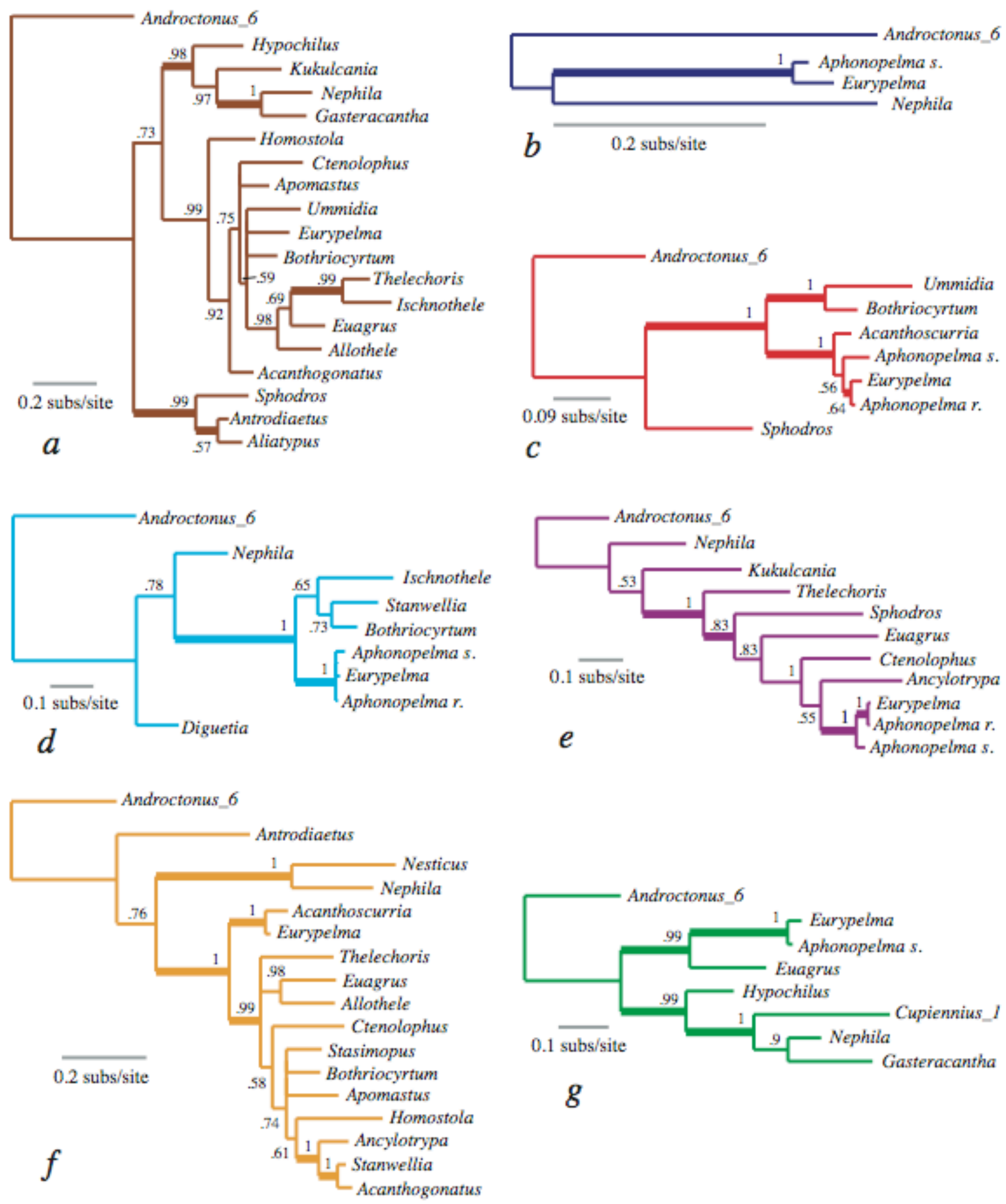


Figure 4.5. Bayesian consensus phylograms based on the partitioned by codon position analyses of the individual ortholog group datasets (*a*-*g*). Posterior probabilities >0.5 are adjacent to nodes. Thickened branches indicate nodes that are supported with parsimony bootstrap percentages >70.

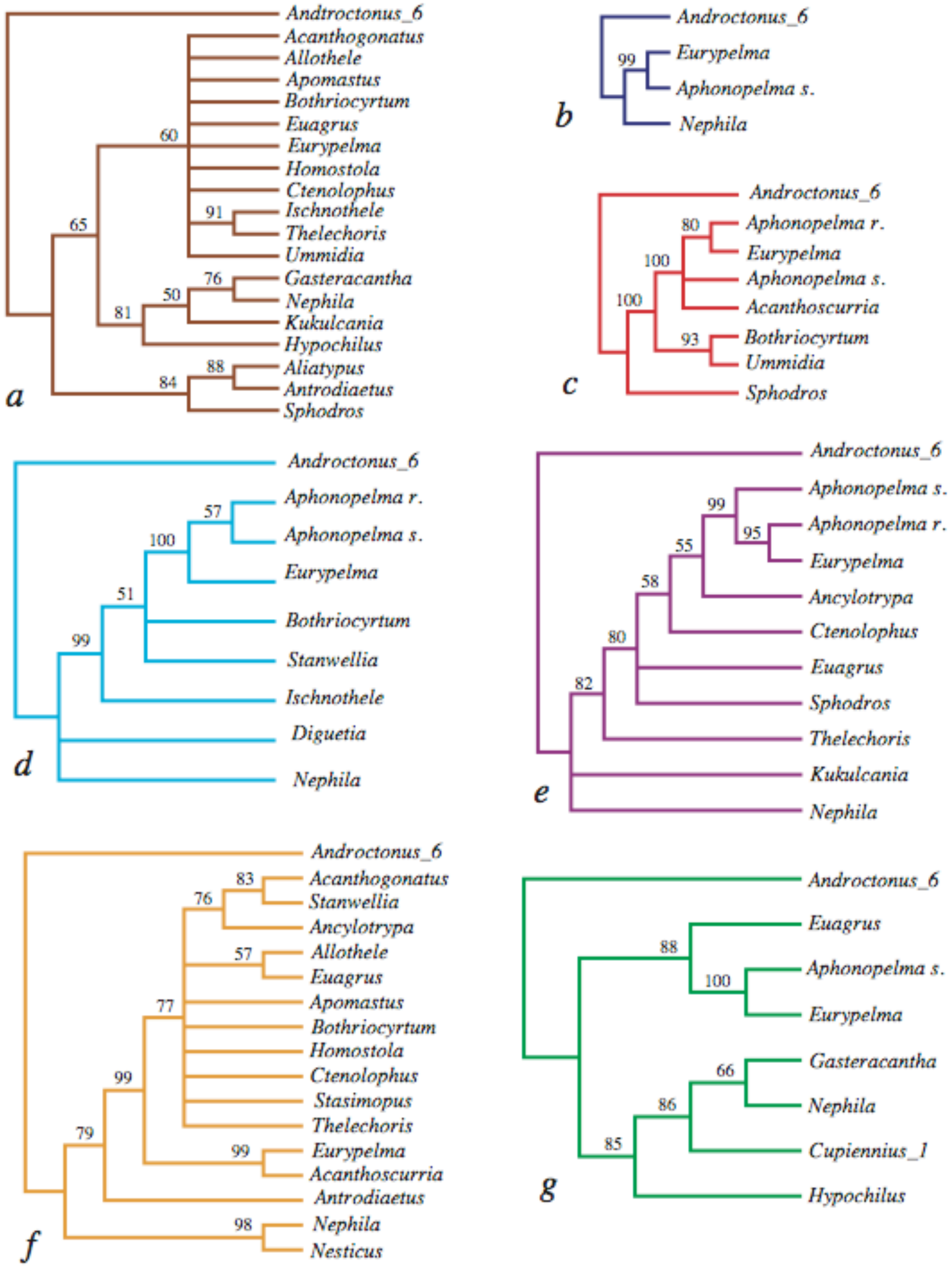


Figure 4.6. 50% majority rule consensus trees from parsimony bootstrap analyses of individual ortholog datasets (*a*-*g*). Nodes with bootstrap percentages >50 are shown.

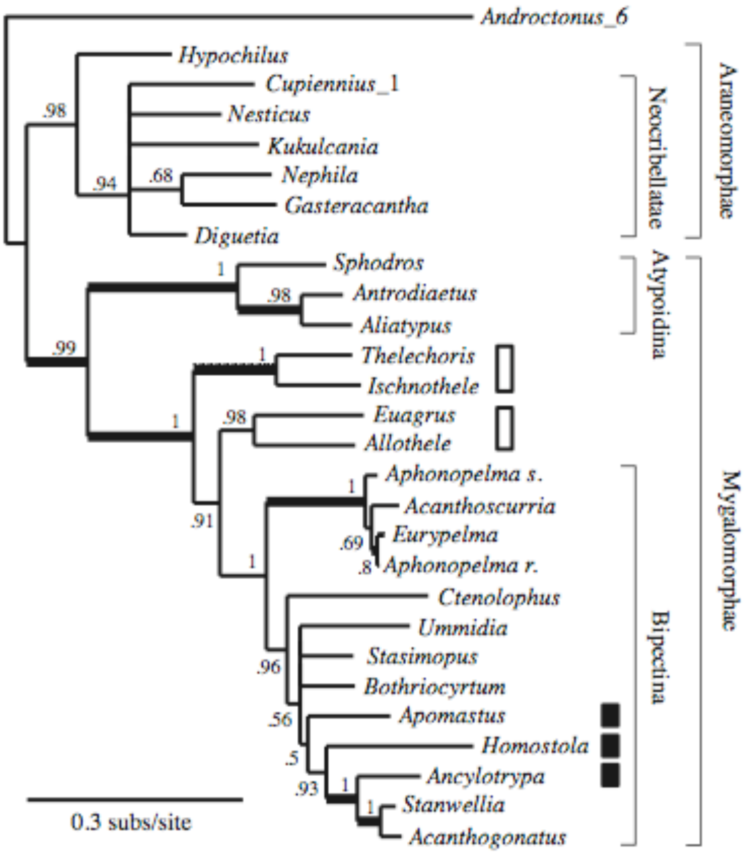


Figure 4.7. Bayesian consensus phylogram based on a partitioned by codon position analysis of the concatenated dataset. Posterior probabilities >0.5 are adjacent to nodes. Thickened branches indicate nodes with >70% parsimony bootstrap support. Boxes indicate non-monophyletic families (open = Dipluridae, closed = Cyrtaucheniidae).

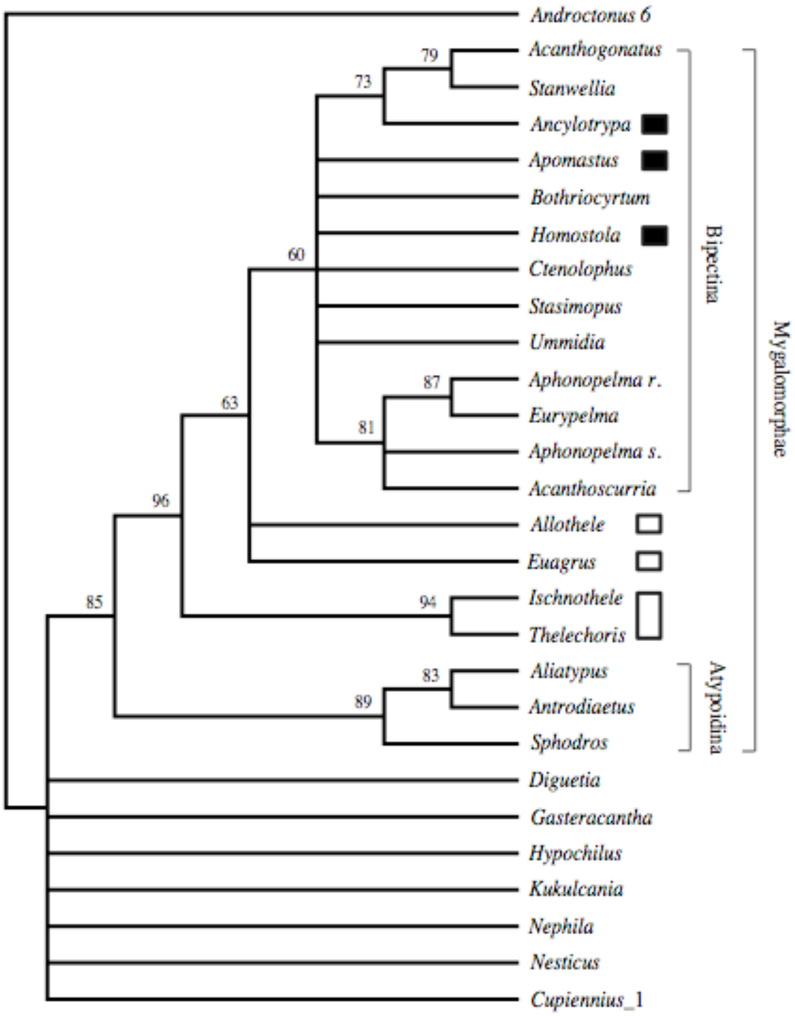


Figure 4.8. 50% majority rule consensus tree from parsimony bootstrap analysis of the concatenated dataset. Nodes with bootstrap percentages >50 are shown. Boxes indicate families not recovered as monophyletic (open = Dipluridae, closed = Cyrtaucheniidae).

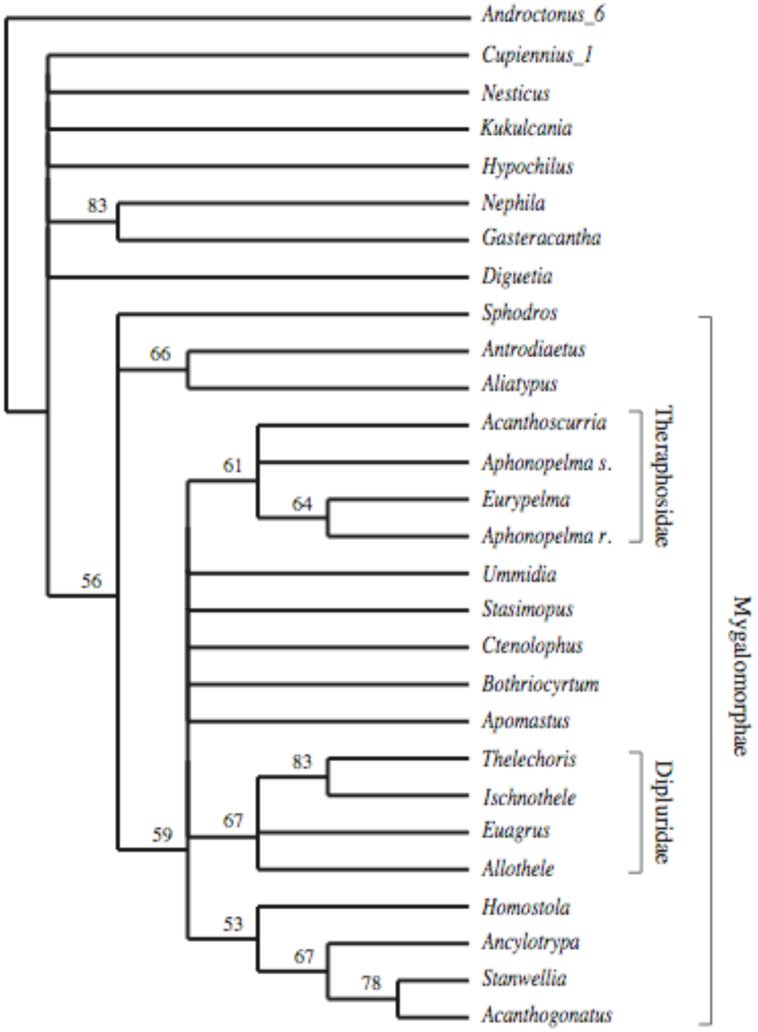


Figure 4.9. 50% majority rule consensus tree from the gene tree parsimony bootstrap analysis based on the global Bayes tree set.

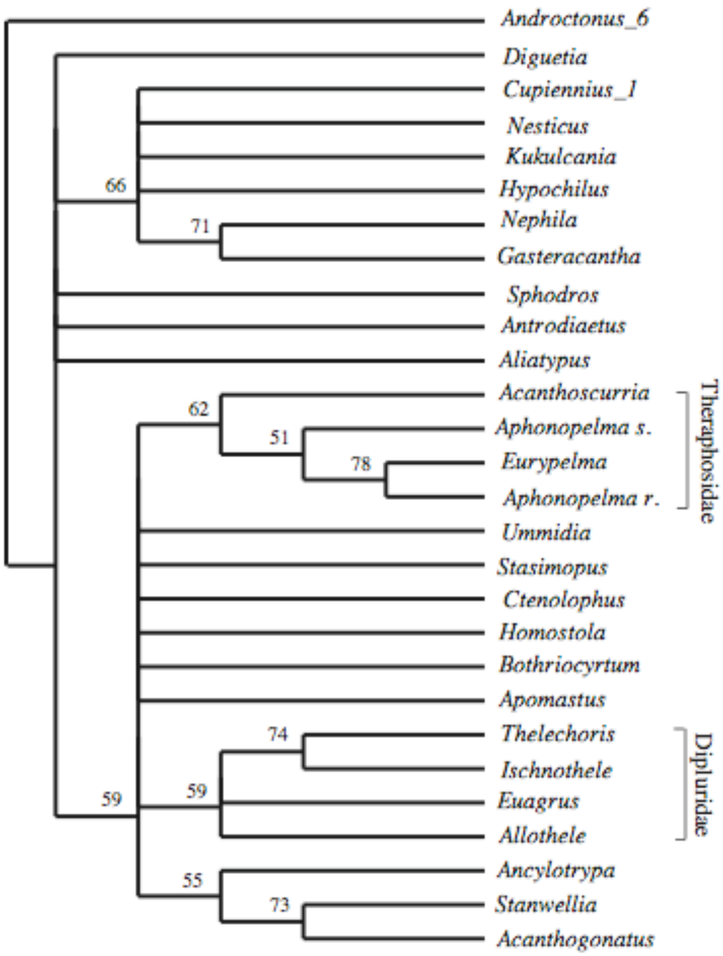


Figure 4.10. 50% majority rule consensus tree from gene tree parsimony bootstrap analysis based on the individual ortholog group Bayes tree sets. Nodes with bootstrap percentages >50 are shown.

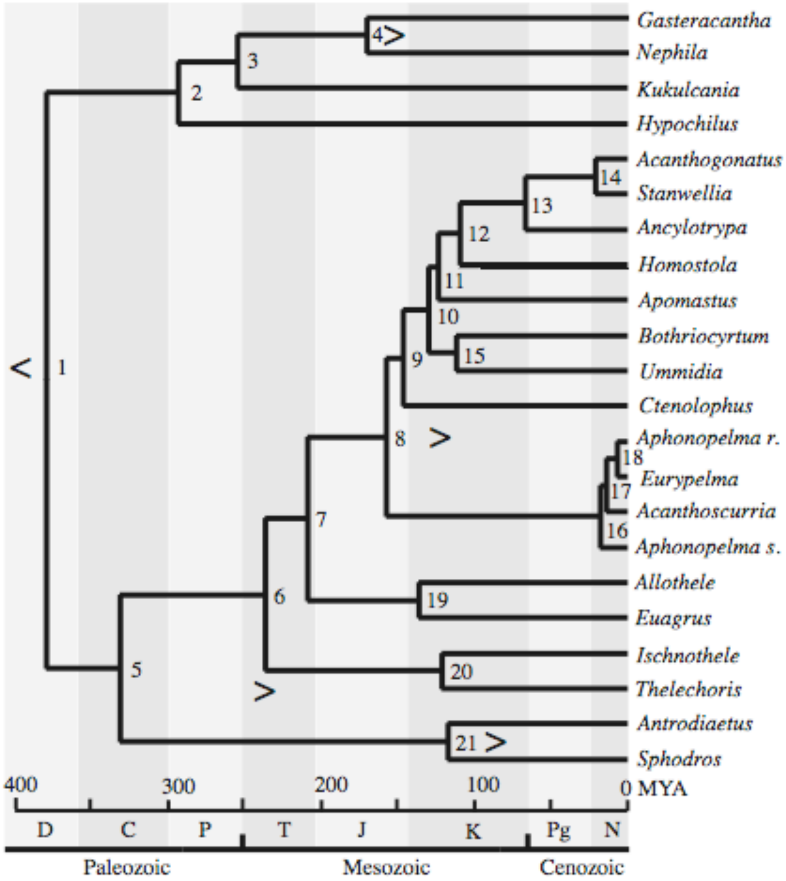


Figure 4.11. Chronogram from Bayesian analysis using correlated rates. Tree topology is based on Bayesian analysis of the trimmed concatenated dataset (species represented by only one paralog removed). Node numbers correspond to Table 4.6. Arrows (> and <) indicate minimum and maximum soft boundaries, respectively.

Conclusion

Mesothele and mygalomorphs are an important component of the phylogenetic diversity of spiders. Yet these groups have often been overlooked from a molecular evolution perspective. Here, I have shown that these spiders are essential for obtaining a full understanding of spider evolution. Sampling silk gene transcripts from mesothele and mygalomorph spiders revealed that the common ancestor of extant spiders possessed a diversity of silk genes, having at least one spidroin gene and one Egg Case Protein. The spidroin gene family later diversified after the split of opisthotheles from mesotheles, which has resulted in the functionally divergent orb-weaver spidroins we see today. Mesothele and mygalomorph spiders may also be beneficial for understanding how selection and concerted evolution act on spidroin orthologs. Mesotheles and mygalomorphs have extremely low dispersal abilities and are thus highly tied to the environment they inhabit. They rely on their silken tools for protection, prey capture, and reproduction. My analysis of spidroin evolution among closely related species of trapdoor spiders reveals that selection has had different influence among the regions of spidroin genes. While purifying selection and concerted evolution have acted to conserve spidroin repetitive regions among orthologs, selection has resulted in radical changes in alpha helical tendency in the carboxy terminal region. The discovery of silk-like secretions from the tarsi of tarantulas indicates that silk gene studies of mygalomorphs and mesotheles may uncover novel silk associated genes. My investigation into transcript expression reveals that known silk genes are not expressed in the tarsi. Instead, novel transcripts with silk-like characteristics were highly expressed in the tarsi, suggesting that the genes encoding tarsal secretions evolved independently of spidroins. Gene families are not only essential for studying adaptation, but also have high phylogenetic utility. The hemocyanin gene family, despite exhibiting lineage specific gene duplication and loss, recovers a well-resolved phylogeny of mygalomorphs and gives insights into ancient spider divergence dates. Further genomic investigation in Mesothelae and Mygalomorphae will give greater insights into gene families with adaptive significance and their contribution to the evolutionary success and incredible species richness of spiders.