Astakines in arthropods – phylogeny and gene structure

Lena Ericsson and Irene Söderhäll

Department of Comparative Physiology, Uppsala University, Norbyvägen 18A, 75326, Uppsala, Sweden

Corresponding author: Irene Söderhäll, irene.soderhall@ebc.uu.se

The editorial process of this article was handled by Mirodrag Belosevic.

Abstract

Astakine1 was isolated as a hematopoietic cytokine in the freshwater crayfish *Pacifastacus leniusculus*. In this study we detect and compare 79 sequences in GenBank, which we consider to be possible astakine orthologs, among which eleven are crustacean, sixteen are chelicerate and 52 are from insect species. Available arthropod genomes are searched for astakines, and in conclusion all astakine sequences in the current study have a similar exon containing CCXX(X), thus potentially indicating that they are homologous genes with the structure of this exon highly conserved. Two motifs, RYS and YP(N), are also conserved among the arthropod astakines. A phylogenetic analysis reveals that astakine1 and astakine2 from *P. leniusculus* and *Procambarus clarkii* are distantly related, and may have been derived from a gene duplication occurring early in crustacean evolution. Moreover, a structural comparison using the Mamba intestinal toxin (MIT1) from *Dendroaspis polylepis* as template indicates that the overall folds are similar in all crustacean astakines investigated.

Keywords: Astakine; Astakine-like; Prokineticin;

1. Introduction

Comparison of protein sequences can provide meaningful insights into how proteins function as well as how they have evolved (Ajawatanawong and Baldauf, 2013). During the past five years, the number of available annotated eukaryotic genomes has increased dramatically, from 40 in 2011 to 359 in 2016 (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/). This increase of genomic data, combined with an even larger increase in reported protein sequences, has made it possible to perform comparisons between sequences from an increasing number of species.

In 2005, we published the first report of a hematopoietic cytokine in freshwater crayfish *Pacifastacus leniusculus* and named the protein astakine (Söderhäll et al., 2005). It was purified from plasma and sequenced by mass spectrometry, and the mRNA sequence was identified and characterized. Astakine was found to contain a prokineticin domain (pfam 06607) with 56% similarity and 31% identity to that of *Bombina variegata* Bv8 (GenBank accession no. AAD45816) and high similarity to other vertebrate prokineticins (Mollay et al., 1999). This astakine is now named as astakine 1 (Ast1). Further, we isolated cDNA for a second astakine from the penaeid shrimp *Penaeus monodon*, which had an insertion of 13 amino acids after amino acid 44 compared with *P. leniusculus* astakine1 (Söderhäll et al., 2005). . Later, we identified a similar, astakine2 in *P. leniusculus,* and this longer types of astakines is now named as astakine 2 (Ast2) (Lin et al., 2010). Since then, several astakine-like proteins have been described from different invertebrate animals, primarily arthropods (Hsiao and Song, 2010; Lin et al., 2010; Li et al., 2016; Shelby et al., 2015).

The prokineticin protein was originally isolated from black mamba venom (Boisbouvier et al., 1998) and then from skin secretions of frogs (Mollay et al., 1999). Prokinectins are 80-90 amino acids in length, and contain 10 cysteines forming 5 bridges. The amino-terminal

sequence in all vertebrate prokineticins is AVIT, and in addition to being present in snake venoms and frog skin secretions, these proteins are expressed in various tissues in mammals (Kaser et al., 2003). Vertebrate prokineticins are involved in not only angiogenesis and cancer (Monnier and Samson, 2010), immunity (Martucci et al., 2006) and hematopoiesis (LeCouter et al., 2004) but also reproduction (Wechselberger et al., 1999) pain regulation (Negri et al., 2009, 2002) and neural repair (Gordon et al., 2016). Moreover, two highly homologous G-protein coupled receptors for prokineticins have been identified (Lin, 2002).

A common trait of all arthropod astakines is that they lack the N-terminal sequence AVIT, which is a signature sequence for vertebrate prokineticins and is important for binding to their G-protein coupled receptors, PROKR1 and PROKR2 (Kaser et al., 2003). To date, no similar receptor has been detected for the invertebrate astakines, but binding studies have shown that *P. leniusculus* astakine binds to the beta subunit of ATP synthase (Lin et al., 2009) a finding later confirmed to also occur in shrimp (Liang et al., 2015).

Several arthropod protein sequences with similarity to that of crayfish astakine can be found in GenBank, but only a few studies about the function of this group of proteins have been published. An important role of *P. leniusculus* Ast1 in hemocyte proliferation and differentiation has been described (Lin et al., 2010) and more reports indicating roles of astakines in immunity and hematopoiesis have been published (Hsiao and Song, 2010; Jiravanichpaisal et al., 2007; Liang et al., 2015; Lin et al., 2008; Li et al., 2016; Shelby et al., 2015; Thomas et al., 2016; Wilson et al., 2015). However, knowledge of arthropod astakine functions remains scarce, and to date, no structure has been experimentally determined for any of these proteins, although we have performed homology modeling for *P. leniusculus* astakine1 and astakine2 by using mamba intestinal toxin 1 as a template (Lin et al., 2010).

Two structures have been determined experimentally for vertebrate prokineticins (Boisbouvier et al., 1998; Morales et al., 2010). A solution structure of the disulfide-bridge topology of mamba intestinal toxin 1 (MIT1), determined by NMR spectroscopy, reveals similarities with colipase (an enzyme secreted from pancreas). Both peptides show resistance to endoproteases, and the authors have suggested that exocrine glands such as the pancreas may have evolved into venom glands, owing to the structural similarities between colipase and Mamba intestinal toxin (MIT1) from *Dendroaspis polylepis* (Boisbouvier et al., 1998). Interestingly, several astakine-like sequences in arthropods show similarities with venom proteins in insects and spiders.

In the present study, we searched 27 arthropod genomes to find genes encoding astakine-like proteins and searched for additional astakine-like protein or cDNA sequences in GenBank. In total, we detected 79 sequences, which we deemed to be possible astakine orthologs. We compared these sequences, focusing on differences in putative indel sequences to identify possible structures that may be of interest for further functional studies. Further, we used the software Phyre2 to compare the putative 3-dimensional structures of crustacean astakines.

2. Materials and methods

2.1 Naming and definition of astakines

The astakine sequences were divided into two groups, astakine 1 (Ast1) and astakine 2 (Ast2), according to the naming in *P. leniusculus* where these molecules were first defined (Lin et al., 2010; Söderhäll et al., 2005). Ast1 contains a prokineticin domain with 10 cysteines, whereas Ast2 in addition has an insert of 10-20 amino acids containing the conserved YP(N/D) motif. Naming of the proteins was done as follows: the protein name begins with an abbreviation of the species followed by Ast1 or Ast2. When multiple copies were found, lower case letters were added in alphabetic order as additional identifier. For example, the two *Stegodyphus*

mimosarum Ast2 were named St-Ast2a and St-Ast2b. If several species had similar initials we named as in the following example: *Procambarus clarkii*= Pcl; *Polistes Canadensis*= Pca, or *Camponatus floridanus* = Ca.f; *Copidosoma floridanus*= Co.f.

2.2. BLAST search and sequence collection

Seventy-nine astakine-like sequences from different arthropods were investigated in this study. The sequence comparison was limited to the prokineticin domain, which was deemed to start with the second amino acid located to the N-terminal side of the first cysteine residue in the N-terminus and to end with the second amino acid after the tenth cysteine in the C-terminus. Sequences were numbered starting with 1 at the second amino acid preceding the first C and extending to the second amino acid after the tenth C (i.e., $X_1X_2C_3 - C_{94}X_{95}X_{96}$).

We searched for astakine-like sequences in several different ways. Astakine and astakine-like were used as keywords to search in GenBank at the NCBI web page (https://www.ncbi.nlm.nih.gov/). The protein and nucleotide sequences from *P. leniusculus* Ast1 and Ast2 (accessions AAX14635.1 and ABQ23256.1 respectively), were used as query sequences in BLAST searches (https://blast.ncbi.nlm.nih.gov/Blast.cgi) by using Protein BLAST, blastx and tblastn. The resulting astakine-like sequences from different arthropods were used for further searching via BLAST. Non-redundant protein sequences or specific arthropod genomes were selected as the BLAST databases. In total, 79 arthropod astakine-like sequences were found and used for further analysis (Supplementary table 1). Some of the astakine-like sequences were found only as nucleotide sequences, and in those cases, we used the ExPASy translate tool (http://web.expasy.org/translate/) to translate them into protein sequences (Artimo et al., 2012). As cutoff value for identity of 30 %, and moreover the conserved ten cysteine pattern was used as criteria for naming the sequence as astakine. In

some species, we found more than one astakine-like sequence, and in those cases, additional identifiers were assigned as described in section 2.1.

2.3. Investigation of the exon structure of arthropod astakines

To explore the exon-intron structure of putative astakine genes, 33 different arthropods found in the GenBank genome assembly database (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/all/) were analyzed with BLASTn for full-length astakine mRNA sequences from these species. BLAST hits, i.e., fragments of the mRNA sequences in the respective genomes, were downloaded. The nucleotide fragments were translated into amino acid sequences to determine the astakine protein sequences and thereby the exon structure and the length of the introns were estimated.

2.4 Multiple sequence alignments and phylogenetic analyses

To compare the different astakine-like sequences, we performed multiple sequence alignment of their protein sequences by using the Muscle (Edgar, 2004) online tool from EMBL-EBI (http://www.ebi.ac.uk/Tools/msa/muscle/).

Phylogenetic trees were constructed for two different datasets of the astakine sequences. Two different methods were used for both datasets; the Bayesian method using MrBayes 3.2, (Ronquist et al., 2012) and the maximum likelihood (ML) method using the IQ-TREE-1.5.5 software (Nguyen et al., 2015). The best-fit evolution model for the ML analysis of the 79 sequences was WAG+I+G4 and for the ML analysis of the 33 sequences VT + I + G4. WAG combines two empirical models of protein evolution, Dayhoff and JTT, using an approximate maximum likelihood method (Whelan and Goldman, 2001). For the Bayesian analysis the best model was estimated for both datasets to be WAG. The rates were set to equal.

IQ-TREE estimates the appropriate evolutionary model using Modelfinder (Kalyaanamoorthy et al., 2017). Ultrafast bootstrap approximation (Minh et al., 2013) was used to assess branch

150 support values. The number of replicates was set to 1000. Bayesian phylogenetic inference
151 uses Markov chain Monte Carlo (MCMC) methods to produce the most likely phylogenetic
152 tree for a given set of data.
153 One of the dataset contained all the 79 astakine protein sequences in this study. The other
154 dataset consisted of 33 astakine sequences based on sequences showing > 35 % identity to Pl-
155 Ast2. All of the four trees were rooted by an astakine-like sequence from the collembolan
156 hexapod *Folsomia candida*.
157
158 *2.5. Detection of signal peptides and calculation of isoelectric points and molecular weights*
159 The presence and locations of putative cleavage sites for signal peptides of the astakine
160 sequences were predicted by the SignalP 4.1 server (Petersen et al., 2011)
161 (http://www.cbs.dtu.dk/services/SignalP/). D-cutoff values were set to default (meaning the
162 score above which the SignalP program will predict a cleavage site for a signal peptide for
163 eukaryotes), and input sequences were allowed to include TM regions. After removal of the
164 signal peptide sequences, the reduced astakine sequences were analyzed with the ExPASy
165 compute pI/Mw tool web site (http://web.expasy.org/compute_pi/), and average resolutions
166 were used to calculate isoelectric points and molecular weights assuming no glycosylation or
167 lipid binding of the proteins (Bjellqvist et al., 1993).
168
169 *2.6. Structure prediction*
170 Three-dimensional structures of full-length mature proteins (without signal peptide) of *L.*
171 *vannamei, P. monodon, M. japonicus, P. clarkii, P. leniusculus* Ast2 and *P. leniusculus* Ast1
172 were predicted using the software Phyre2 (Kelley et al., 2015) at the Phyre2 web page
173 (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index), using normal modeling mode.
174 The three top-scoring models for each crustacean astakine were downloaded in Protein Data

Bank (PDB) format, and Phyton Molecule Viewer (Sanner, 1999) was used to display the models. Charge potentials for the protein models were computed with Adaptive Poisson-Boltzmann Solver (APBS) (Baker et al., 2001) and mapped to the surface with medium quality and a distance of 1.0 Å from the surface.

3. Results and Discussion

3.1. Astakines or astakine-like sequences

In addition to P. leniusculus Ast1 and Ast2, we detected 77 putative astakine protein sequences in GenBank or by BLAST searches, and performed multiple sequence alignment of the conserved prokineticin-like domain of these sequences together with an astakine-like sequence from the primitive hexapod *Folsomia candida* (Springtail) as an outgroup sequence (Figure 1).

The alignment files were then examined manually, and numbering of the amino acids is as described in the method section. Ten highly conserved cysteine residues, two motifs ($R_{22}Y_{23}S_{24}$ and $Y_{57}P_{58}$) and a conserved proline residue (P_{80} between C_{79} and C_{81}) were found in the alignment (numbers according to Figure 1). Seventy-nine sequences were defined as Ast1 or Ast2, among which 11 were crustacean, 16 were chelicerate and 52 were from Insecta (Figure 1, Supplementary table 1). In addition to the highly conserved amino acid motifs mentioned above, some amino acids were identified as being more or less conserved in specific groups of species. For crustaceans, the residues $G_{11}P_{12}$ and $P_{30}L_{31}G_{32}D/E_{33}$ and $T_{95}C_{96}Q_{97}$ were highly conserved among decapods.

Among the 52 insect sequences, including 13 diverse Hemiptera, 15 ants, and 21 other Hymenoptera, the highest similarities between sequences were detected among the ants. The amino acid P_{20} and the motifs $M_{29}P_{30}F/Y_{31}Q_{32}Q_{33}$ and $T_{49}I_{50}T_{51}T_{53}N_{54}L_{55}T_{56}$ are highly conserved among ants but less conserved in other insect species. In bees, we also identified a

conserved Q_{27}, which is also present in shrimp (Figure 1). Unfortunately, there have been few studies about the functions of the insect and chelicerate astakines, and therefore it is premature to draw any functional conclusions regarding this conservation of motifs at the organism level.

In conclusion, we could detect astakine sequences in chelicerats, crustaceans, and some insect orders. So far no astakine sequence was found in Myriapoda, which could be due to lack of sequence data for this group. Interestingly, most of the insect sequences belonged to insects of the orders Hemiptera and Hymenoptera, whereas no astakine-like sequences have been reported to date from Diptera, Coleoptera or Lepidoptera. According to several recent phylogenomic studies of insects, it seems clear that Hymenoptera is a basal order within the holometabolous group and is a sister group to Diptera, Coleoptera or Lepidoptera (Behura, 2015; Peters et al., 2017, 2014). Thus, the data presented in our study suggest that astakine genes may have been lost at the root of the clade Aparaglossata, which includes all holometabolous insects except Hymenoptera (Peters et al., 2014).

3.2 Signal peptides, isoelectric points and molecular weights in arthropod astakines

We could find cleavage sites for signal peptides in all astakine protein sequences. However, when we compared the *Atta cephalotes* (GenBank Accession number XP_012063524.1**)** sequence with that of the close relative *Atta colombica* (GenBank Accession number KYM75707.1), these sequences were nearly identical, except for a longer N-terminal reported in *A. cephalotes*. This *A. cephalotes* sequence was predicted by an automatic analysis from genomic data and submitted as such to GenBank, whereas the *A. colombica* sequence was experimentally identified from transcriptomic sequences. This result indicates that in order to verify protein sequences, experimental confirmation is needed. Thus, the long deduced N-

terminal sequence in *A. cephalotes* reported as predicted astakine-like protein with Accession number XP_012063524.1.

In the crustacean group, the isoelectric point for most of the investigated crustacean astakines in the current study varied between pI = 4.54 and pI = 5.13. However, two exceptions were Pl-Ast2 (pI = 7.04) and Pcl-Ast2 (pI = 7.69) (Supplementary table 1).

The isoelectric points for chelicerate astakines varied between pI = 3.92 and pI = 8.54. Most of the identified astakines had pI values higher than 4.50 and lower than 6.80, and hence they are negatively charged at neutral or physiological pH.

In summary, all astakines detected in our study have a predicted signal peptide and thus may be secreted proteins. That is similar to the prokineticins in vertebrates, which all are secreted. All astakines are small molecules with molecular mass between eight and fourteen kDa, and the prokineticin domain constitutes the main part of the mature protein. However, in contrast to vertebrate prokineticins, most of the arthropod astakines, with some exceptions have a pI below 7, meaning a negative charge at physiological pH. However, there are no studies published so far about experimentally determined structure for any arthropod astakine, and therefore it can not be concluded whether their surface charges are negative or positive with certainty.

3.3. Putative exon-intron structures of arthropod astakines

We identified putative arthropod astakine sequences by searching in the arthropod genomes that are annotated at NCBI. Our deduced gene structures for the identified astakine genes among these arthropods showed some general similarities as well as some differences among the major classes. In all astakine genes investigated in the current study, the exon containing the 3' end of the prokineticin domain ends with CCXX or CCXXX (Figure 2 and

249 Supplementary table 2). Most of the genes in insects consist of two exons in total, with the
250 first one ending as mentioned above (Figure 2). Four of the astakine genes from insects (*A.*
251 *cephalotes, Linepithema humile, Solenopsis invicta* and *Vollenhovia emeryi)* have an
252 additional exon in the N-terminal region, and in the *H. halys* astakine gene, there is an
253 additional exon in the C-terminal region (Supplementary table 2). However, this conclusion
254 must be considered carefully, because these sequences do not seem to have been confirmed
255 experimentally; as mentioned above for *A. cephalotes,* the predicted first exon may not be
256 expressed or may be inaccurate (Suen et al., 2011; Wurm et al., 2011).
257 Among the chelicerate astakine genes analyzed, most had an extra exon at the N-terminus
258 encoding the signal peptide, and the second exon encoding the structure ending with CCXX
259 or CCXXX (Figure 2). In *M. occidentalis*, an exception among the chelicerates, the PROK-
260 domain is encoded by three exons, and the second exon ends with nucleotides encoding
261 **CPC**EG (Supplementary table 2). No decapod genome is available to date, but the N-terminal
262 exon has been found to share the same structure (ending with the nucleotides encoding CCXX
263 or CCXXX) in the unpublished genome of the decapod marble crayfish, *Procambarus fallax*
264 forma *virginalis* (Phattarunda Jaree, Frank Lyko and Julian Gutekunst, personal
265 communication).
266 In summary, all astakine sequences in the current study have a similar exon encoding the
267 structure ending with CCXX(X), thus potentially indicating that they are homologous genes
268 with the structure of this exon highly conserved. In most of the astakines the PROK-domain is
269 encoded by two exons, the first one ending as above and the second encoding the rest of the
270 prokineticin domain. This result may indicate that these astakine sequences are more closely
271 related to each other than the ones with the prokineticin domain encoded for by three exons.
272 Some of the astakines appear to have an additional exon located at the N-terminus of the gene.

Because this part of the protein contains the signal peptide, it may be less conserved than other regions.

3.4 Multiple sequence alignment of arthropod astakines

The alignment of all astakine amino acid sequences showed that some of the residues are highly conserved. Ten cysteine residues are conserved in all sequences (numbering as in figure 1):

C_3 - (X_5) – C_{10} - (X_4) – C_{15} – C_{16} - (X_{11}) – C_{28} - (X_9) – C_{38} - (X_n) – C_{79}(P)C_{81} - (X_5) – C_{89} - (X_n) – C_{96}

Deviating from the structure above are two insertions of a P in Pl-Ast1 and Pcl-Ast1 between C_3 and C_{10}. In Pcl-Ast1, there are also four deletions between C_{16} and C_{28}. Between C_{81} and C_{89} the two Daphnia sequences, Dm-Ast2 and Dp-Ast2, have two insertions consisting of an alanine and an asparagine (Figure 1).

Two motifs, RYS and YP(N/D), are conserved among the arthropod astakines. Only in PclAst1, the YS part of the RYS is missing in the alignment, and in the RYS motif, arginine is in some sequences substituted by the similar amino acid lysine and in one sequence, that of *Cimex lectularius* Cm-Ast2c, it is replaced by leucine. In 14 of the sequences, tyrosine is replaced by phenylalanine. The serine residue of the RYS is in seven of the insects replaced by an alanine, in Ast2 from *Polistes canadensis* and *V. emeryi* by valine and in four other insects by a threonine. In the YP(N/D) motif, the tyrosine is replaced by phenylalanine in *M. occidentalis* by tryptophan in *Copidosoma floridanum*, and by glutamine in *Diachasma alloeum*. It has previously been shown by mutant recombinant protein experiments that this motif is important for the function of Pl-Ast2 (Lin et al., 2010). In Pl-Ast1 and Pcl-Ast1, there

297 are 24 gaps between C_{38} and C_{79}. These two sequences also lack the YP(N/D) motif, which
298 indicates different functions of Ast1 and Ast2.
299 There is also a proline residue in the position between C_{79} and C_{81}. This residue is conserved
300 in all sequences except in Lhu-Ast2 in (replaced by serine), Tz-Ast2 (alanine), Lh-Ast2b and
301 Zn-Ast2 (aspartic acid), Bg-Ast2 (glycine) and Hl-Ast2 (leucine), and if this finding is not due
302 to sequencing errors, it may have implications for the function of these putative astakines
303 (Figure 1).
304 Taken together, all the astakines have a conserved cysteine pattern with ten cysteines, and
305 between the sixth and seventh cysteine there is an insertion of variable length in all Ast2
306 containing an YP(N/D) motif which for Pl-Ast2 is shown to be of importance for the function
307 (Lin et al., 2010).
308
309 3.5. Phylogenetic analyses of some arthropod astakines
310
311 All of the arthropod astakine sequences were subjected to two different phylogenetic analyses,
312 a maximum likelihood (ML) analysis by the IQ-Tree software (Figure 3a), and a Bayesian
313 analysis by MrBayes software (Figure 3b). The study was restricted to these methods since
314 distance based methods are less reliable when analyzing the high number of diverse
315 sequences as in this study. An astakine-like sequence from the springtail *Folsomia candida*
316 was used to root the tree. In the resulting IQ-TREE-file from the phylogenetic analysis there
317 is a warning that deduction of the phylogeny should be done with caution. This was due to the
318 larger number of parameters (branch lengths and model parameters) in relation to the sample
319 size i.e. the length of the alignment. In order to improve the robustness in the phylogenetic
320 estimation and avoid warnings, a second dataset was constructed of 33 sequences (Figure 4a-
321 b). The sequences included were chosen by their percentage of identity to Pl-Ast2. Seven

sequences were crustacean, eight from chelicerates and 18 were insect astakines, and *F. candida* was used as root sequence. Since the number of parameters depends on the number of sequences there was no warning for the second analysis in IQ-Tree (Figure 4a).

Ast2 sequences from chelicerata and insecta are clustered in two distinct clades in all four analyses. The placement of the crustacean astakines seems to be more uncertain. In the smaller dataset, all the crustacean astakines are gathered in one clade with high support. However, in the analysis of the large dataset the topologies of the crustacean sequences are different depending on the phylogenetic method. Using the maximum likelihood method, the crustacean astakines Dm-Ast2 and Dp-Ast2 belonging to the Cladocera branches of early in a minor clade and differ from all the other astakines, which belongs to Decapoda. In contrast, in the Bayesian method, the crustacean sequences can not be fully resolved, and the decapod Cm-Ast2 is found outside all the other crustaceans in a polytomy. A comparison between the sequence structure of the crustacean astakines shows that Cm-Ast2 contains deviating amino acids in 21 positions. Nine of these positions contain amino acid residues not found in any other astakine sequence in this study. For the rest of the 21 positions identical residues have been found in some of the astakines from chelicerates and insects, and thus Cm-Ast2 is different from the other astakines of Decapoda.

In all phylogenetic analyses, Ast1 and Ast2 from *P. clarkii* and *P. leniusculus* were separated into different groups in the trees, indicating that these sequences are distantly related to each other. Gene duplication may have occurred in the crustacean astakines before the divergence of the species included in this analysis. Ast1 has to date been detected in only *P. leniusculus*, *P. clarkii* (Beltz and Brenneis, personal communication) and *P. fallax* forma *virginalis* (Jaree, Lyko and Jutekunst, personal communication), and it is possible that one of the variants has been lost during evolution in some groups (or has not yet been found).

All other sequences analyzed in this study belong to the Ast2 type. For some species more than one sequence was found. For example, three duplicates of Ast2 from *L. polyphemus* were grouped together, and thus are more closely related to one another than Ast1 and Ast2 in *P. leniusculus*.

The dataset also contains two astakine sequences for the American house spider, *P. tepidariorum,* and the African social velvet spider, *S. mimosarum*. Even if their placement is somewhat uncertain, they are related to each other in the same order in both the ML and the bayesian phylogenetic analyses (Figure 3a-b). Pt-Ast2a groups with Sm-Ast2a*,* and Pt-Ast2b groups with Sm-Ast2b, thus indicating earlier gene duplication before speciation, compared with the evolution of astakine duplicates in *L. polyphemus.*

The chelicerate astakines are clustered in almost the same way in all four analyses. However, the phylogeny of De-Ast2 (*D. erythrina)*, Sm-Ast2a and Pt-Ast2a could not be fully resolved in this analysis, since there is a polytomy in both the ML and Bayesian analysis.

A large number of sequences in the current study belong to the insects. Several of the species have more than one Ast2 sequence namely, *A pisum, L Hesperus, C lectularius, T pretiosum, N vitripennis* and *A echinatior.* In the insect clade, all astakines from hymenoptera except Da-Ast2 are clustered in one large clade. Da-Ast2 is found in another clade together with astakines from Phthiraptera, Hempitera, Blattodea and Isoptera (Figure 3a-b). The topology of this clade is similar but not identical in both trees, but the support values in this area of the trees are lower in the Bayesian tree (Figure 3a). A comparison of the sequence structures of the insect astakines shows that Da-Ast2 also contains several different amino acids, compared to other sequences. Some of the residues are identical to the ones in Lhe-Ast2b. Therefore, it is possible that Da-ast2 is another astakine variant than the other hymenoptera astakines.

Taken together, our phylogenetic trees give a hypothetic indication about the evolutionary relationship between astakine sequences, but it has to be taken into account that such tree analysis are limited by the number of sequences available. When all detected astakine sequences were used in one analysis the different trees were similar but several branches showed low values of support (Figure 3a-b). In contrast, the limited analysis in which sequences of high identity were used showed more robust trees with higher support values (Figure 4a-b).

3.6. ***Structure prediction***

Six crustacean astakine sequences from Pm-Ast2, Lv-Ast2, Mj-Ast2, Pcl-Ast2, Pl-Ast2 and Pl-Ast2 were analyzed with Phyre2 for alignment (Kelley et al., 2015). These sequences were used for comparison in order to get an idea about what parts of the structure that is most likely to be of importance for functional difference between some marine and freshwater crustacean species. Phyre2 determines an evolutionary profile for the query sequence by heuristic searches in protein sequence databases. To search for the best templates, this profile, together with the secondary structure predicted in Phyre2, is scanned against a folding library containing proteins of known, experimentally determined structures. The best-scoring alignments between the query sequences and the library sequences are then used to build three-dimensional models of the query protein. The three top-scoring models from the Phyre2 results were identical for all analyzed sequences: mamba intestinal toxin 1 from *Dendroaspis polyepis,* PDB 1MIT (Boisbouvier et al., 1998); prokineticin Bv8 from *Bombina variegata* PDB 2KRA (Morales et al., 2010); and Dickkopf-related protein 1 (DKK1) from *Homo sapiens* PDB 3S8V (Cheng et al., 2011). These three templates gave different alignment with the astakines as shown in Supplementary figure 1. The confidence for the matches between the submitted astakine sequences and the models was between 98.5 and 100, thus indicating a

high percentage probability that the astakines and the models are homologous. The percentage identity between the astakines and the models was between 28% and 37%. In Phyre2, the proportion of disorder in secondary structures was predicted for all astakine sequences as reported in supplementary table 3. Thus, fairly large portions of these proteins probably lack fixed three-dimensional structures and are unstructured with conformational flexibility, owing to random coil structures. Structures with high proportions of disorder are more difficult to predict. The three models with high confidence for the matches, as mentioned above, in Phyre2, were then used as templates for the prediction of astakine structure, and the predicted structures were displayed in Phyton Molecule Viewer (Sanner, 1999). The overall folding and the core of the astakine structures for these models are shown in Figure 5a and Supplementary figures 2 and 3, with the ten cysteine residues, the RYS and YP(N/D) motifs and the indel regions marked (shown by yellow crosses, Figure 5a, Supplementary figures 2 and 3). Figure 5a shows the structure of the astakines predicted with intestinal toxin 1 as template. The extension of the modeled astakines is between GXC_3 and $C_{89}XRXX$, in the Pcl-Ast2 model between $HC_3 - C_{89}SRTS$ (numbering according to Figure 1). The overall folding of the structures appears to be quite similar among the species. Most of the models contain four cysteine bridges, $C_3 - C_{16}$, $C_{10} - C_{28}$, $C_{15} - C_{79}$ and $C_{38} - C_{89}$, although the two cysteine residues C_3 and C_{16} from Pcl-Ast2 and Pl-Ast1 models appear to be too distant from each other to form a bridge. In contrast, the secondary structures are less similar, especially in the region partly consisting of the indel region (Figure 5a). In this region, an alpha helix is found in Ast2 from *L. vannamei* and *P monodon,* whereas Pl-Ast2 contains one helix and one beta sheet, and in Ast2 from *M. japonicus* and *P. clarkii,* the structures in this region consist only of coil structure. The predicted structure of Pl-Ast1, which has a deletion of 13 amino acid residues in this part of the structure, has a beta sheet. For the astakine structures with Bv8 as templates (Supplementary figure 2), the extension of the modeled residues is almost the same

as in the previous prediction. The only difference is one additional residue in the C-terminus. All structures contain four cysteine bridges. The overall folds are similar in all crustacean astakines, and the secondary structures are more similar within the two groups of astakines compared with the structures with the intestinal toxin as template. Among the shrimp, an alpha helix is found in the indel region, whereas only coil structures are found in the two Ast2 structures from *P. clarkii* and *P. leniusculus* (Supplementary figure 3).

In comparison, the predicted structures using the Dickkopf-related protein 1 (DKK1, 3S8V) as template include all ten cysteine residues forming five cysteine bridges, including C_{81} – C_{96} (Supplementary figure 3). Most of the modeled structures contain XC_3 in the N-terminus, but the Pl-Ast1 structure contains GSC_3 and Pcl-Ast2 only C_3. All of the structures end with $C_{10}Q$, except that of Pcl-Ast2, which ends with $C_{96}QL$. Another difference is the overall folding of the structures predicted from DKK1. These differences include, for example, the RYS motif being located closer to the YP(N/D) residues. *L. vannamei* and *P. monodon* Ast2 have similar secondary structures containing an alpha helix in the indel region, as does *P. clarkii*. In *M. japonicus,* only a small helix is found, and in the two astakines from *P. leniusculus,* the indel region consists only of coiled structure (Supplementary figure 3).

Two regions, the RYS motif and the YP(N/D) motif with an additional asparagine residue, are conserved among the Ast2 and were investigated further. The charge potentials were computed and mapped to the surface with 1MIT as a template (Figure 5b). The RYS and YP(N/D) structures of the surfaces together with the structures of the residues in the overall folds (Figure 5a and Supplementary figures 2 and 3) were compared across all the modeled astakine structures. The structures of the residues in the RYS motif seem to be similar in the models predicted with the intestinal toxin and prokineticin templates, although none of the

446 template contains this motif. In contrast, in the models predicted with DKK1 as template, the
447 structure of the RYS motif is different in some astakines, whereas no model for these residues
448 in Pcl-Ast2 could be predicted using this template.
449
450 All six crustacean structures modeled by Phyre2, except Pl-Ast1, contain the YP(N/D) motif,
451 and this part of the molecule has previously been found to be important for function in the
452 granular hemocyte lineage in *P. leniusculus* (Lin et al., 2010). This motif is near the variable
453 indel region, and therefore it is possible that the structure around this motif might vary in the
454 different models (Figure 5a, Supplementary figure 2-3).
455 In conclusion, the overall structure of the models predicted with intestinal toxin and
456 prokineticin as templates seems to be similar for most of the predicted astakine structures.
457 The secondary structure and the backbone of the proteins are also similar in most of the
458 regions, although not in the region containing indels. The structure of the RYS motif also
459 looks similar among the astakines modeled by using these two templates. These similarities
460 may be because the two templates are similar, and the same four cysteine bridges are
461 predicted in the modeled structures.
462 In contrast, in the models predicted with DKK1 as template, the overall folding and number
463 of modeled cysteine residues and resulting number of bridges differ from the others. The
464 structure of the RYS motif is more variable when this template is used, but DKK1 is the only
465 template yielding structures with more similar YP(N/D) structures, possibly be because these
466 models contain five cysteine bridges, which may stabilize the structure, and especially the
467 indel region and YP(N/D).
468
469 4. Conclusions

Since the first reported astakine sequence in 2005, we could find 77 other arthropod astakines or astakine-like sequences in GenBank. A search in available genomes revealed a similar exon-intron structure among the arthropod astakines. Although all sequences are similar and contain the core astakine structure with ten cysteines, the RYS and YP(N/D) motifs, a phylogenetic analysis combining all arthropods were not fully resolved and gave trees with some polytomies. However, both the ML and the Bayesian method showed clearly separate crustacean, chelicerate and insect clades. In addition, the hymenopteran sequences all grouped together with one exception, and the hemiptera formed a common clade also with only one exception.

In crustaceans, the distance between Ast1 and Ast2 from *P. leniusculus* and *P. clarkii* indicates a gene duplication occurring early in crustacean evolution.

A structural comparison using the Phyre2 software gave some indication of a similar overall core structure, but since the available templates are fairly distant, such predictions has to be evaluated with care. Nevertheless, our structural comparison of five crustacean sequences could still show that the indel sequences following the preserved YP(N/D) motif is likely to give a specific surface structure that varies among species, and can be of specific interest to experimentally manipulate in order to reveal possible function.

Acknowledgements

We thank Professor B. S. Beltz and Dr. Georg Brenneis for information about *Procambarus clarkii* astakine 1 sequence, and Professor Frank Lyko and Dr. Julian Gutekunst for insight in *Procambarus clarkii* genome

References

Ajawatanawong, P., Baldauf, S.L., 2013. Evolution of protein indels in plants, animals and

496 fungi. BMC Evol. Biol. 13, 140. doi:10.1186/1471-2148-13-140
497 Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., Duvaud, S.,
498 Flegel, V., Fortier, A., Gasteiger, E., Grosdidier, A., Hernandez, C., Ioannidis, V., Kuznetsov,
499 D., Liechti, R., Moretti, S., Mostaguir, K., Redaschi, N., Rossier, G., Xenarios, I., Stockinger,
500 H., 2012. ExPASy: SIB bioinformatics resource portal. Nucleic Acids Res. 40, W597–603.
501 doi:10.1093/nar/gks400
502 Baker, N.A., Sept, D., Joseph, S., Holst, M.J., McCammon, J.A., 2001. Electrostatics of
503 nanosystems: application to microtubules and the ribosome. Proc. Natl. Acad. Sci. U. S. A. 98,
504 10037–10041. doi:10.1073/pnas.181342398
505 Behura, S.K., 2015. Insect phylogenomics. Insect Mol. Biol. 24, 403–411.
506 doi:10.1111/imb.12174
507 Bjellqvist, B., Hughes, G.J., Pasquali, C., Paquet, N., Ravier, F., Sanchez, J.C., Frutiger, S.,
508 Hochstrasser, D., 1993. The focusing positions of polypeptides in immobilized pH gradients
509 can be predicted from their amino acid sequences. Electrophoresis 14, 1023–1031.
510 Boisbouvier, J., Albrand, J.P., Blackledge, M., Jaquinod, M., Schweitz, H., Lazdunski, M.,
511 Marion, D., 1998. A structural homologue of colipase in black mamba venom revealed by
512 NMR floating disulphide bridge analysis. J. Mol. Biol. 283, 205–219.
513 doi:10.1006/jmbi.1998.2057
514 Cheng, Z., Biechele, T., Wei, Z., Morrone, S., Moon, R.T., Wang, L., Xu, W., 2011. Crystal
515 structures of the extracellular domain of LRP6 and its complex with DKK1. Nat. Struct. Mol.
516 Biol. 18, 1204–1210. doi:10.1038/nsmb.2139
517 Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high
518 throughput. Nucleic Acids Res. 32, 1792–1797. doi:10.1093/nar/gkh340
519 Gordon, R., Neal, M.L., Luo, J., Langley, M.R., Harischandra, D.S., Panicker, N., Charli, A.,
520 Jin, H., Anantharam, V., Woodruff, T.M., Zhou, Q.-Y., Kanthasamy, A.G., Kanthasamy, A.,

2016. Prokineticin-2 upregulation during neuronal injury mediates a compensatory protective response against dopaminergic neuronal degeneration. Nat. Commun. 7, 12932. doi:10.1038/ncomms12932

Hsiao, C.-Y., Song, Y.-L., 2010. A long form of shrimp astakine transcript: molecular cloning, characterization and functional elucidation in promoting hematopoiesis. Fish Shellfish Immunol. 28, 77–86. doi:10.1016/j.fsi.2009.10.016

Jiravanichpaisal, P., Puanglarp, N., Petkon, S., Donnuea, S., Söderhäll, I., Söderhäll, K., 2007. Expression of immune-related genes in larval stages of the giant tiger shrimp, Penaeus monodon. Fish Shellfish Immunol. 23, 815–824. doi:10.1016/j.fsi.2007.03.003

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., Jermiin, L.S., 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14, 587–589. doi:10.1038/nmeth.4285

Kaser, A., Winklmayr, M., Lepperdinger, G., Kreil, G., 2003. The AVIT protein family. Secreted cysteine-rich vertebrate proteins with diverse functions. EMBO Rep. 4, 469–473. doi:10.1038/sj.embor.embor830

Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., Sternberg, M.J.E., 2015. The Phyre2 web portal for protein modeling, prediction and analysis. Nat. Protoc. 10, 845–858. doi:10.1038/nprot.2015.053

LeCouter, J., Zlot, C., Tejada, M., Peale, F., Ferrara, N., 2004. Bv8 and endocrine gland-derived vascular endothelial growth factor stimulate hematopoiesis and hematopoietic cell mobilization. Proc. Natl. Acad. Sci. U. S. A. 101, 16813–16818. doi:10.1073/pnas.0407697101

Liang, G.-F., Liang, Y., Xue, Q., Lu, J.-F., Cheng, J.-J., Huang, J., 2015. Astakine LvAST binds to the β subunit of F1-ATP synthase and likely plays a role in white shrimp Litopeneaus vannamei defense against white spot syndrome virus. Fish Shellfish Immunol. 43,

546 75–81. doi:10.1016/j.fsi.2014.12.015
547 Lin, D.C.-H., 2002. Identification and Molecular Characterization of Two Closely Related G
548 Protein-coupled Receptors Activated by Prokineticins/Endocrine Gland Vascular Endothelial
549 Growth Factor. J. Biol. Chem. 277, 19276–19280. doi:10.1074/jbc.M202139200
550 Lin, X., Kim, Y.-A., Lee, B.L., Söderhäll, K., Söderhäll, I., 2009. Identification and properties
551 of a receptor for the invertebrate cytokine astakine, involved in hematopoiesis. Exp. Cell Res.
552 315, 1171–1180.
553 Lin, X., Novotny, M., Söderhäll, K., Söderhäll, I., 2010. Ancient cytokines, the role of
554 astakines as hematopoietic growth factors. J. Biol. Chem. 285, 28577–28586.
555 doi:10.1074/jbc.M110.138560
556 Lin, X., Söderhäll, K., Söderhäll, I., 2008. Transglutaminase activity in the hematopoietic
557 tissue of a crustacean, Pacifastacus leniusculus, importance in hemocyte homeostasis. BMC
558 Immunol. 9, 58. doi:10.1186/1471-2172-9-58
559 Li, Y., Jiang, S., Li, M., Xin, L., Wang, L., Wang, H., Qiu, L., Song, L., 2016. A cytokine-
560 like factor astakine accelerates the hemocyte production in Pacific oyster Crassostrea gigas.
561 Dev. Comp. Immunol. 55, 179–187. doi:10.1016/j.dci.2015.10.025
562 Martucci, C., Franchi, S., Giannini, E., Tian, H., Melchiorri, P., Negri, L., Sacerdote, P., 2006.
563 Bv8, the amphibian homologue of the mammalian prokineticins, induces a proinflammatory
564 phenotype of mouse macrophages. Br. J. Pharmacol. 147, 225–234.
565 doi:10.1038/sj.bjp.0706467
566 Minh, B.Q., Nguyen, M.A.T., von Haeseler, A., 2013. Ultrafast approximation for
567 phylogenetic bootstrap. Mol. Biol. Evol. 30, 1188–1195. doi:10.1093/molbev/mst024
568 Mollay, C., Wechselberger, C., Mignogna, G., Negri, L., Melchiorri, P., Barra, D., Kreil, G.,
569 1999. Bv8, a small protein from frog skin and its homologue from snake venom induce
570 hyperalgesia in rats. Eur. J. Pharmacol. 374, 189–196.

571 Monnier, J., Samson, M., 2010. Prokineticins in angiogenesis and cancer. Cancer Lett. 296,
572 144–149. doi:10.1016/j.canlet.2010.06.011
573 Morales, R.A.V., Daly, N.L., Vetter, I., Mobli, M., Napier, I.A., Craik, D.J., Lewis, R.J.,
574 Christie, M.J., King, G.F., Alewood, P.F., Durek, T., 2010. Chemical synthesis and structure
575 of the prokineticin Bv8. Chembiochem Eur. J. Chem. Biol. 11, 1882–1888.
576 doi:10.1002/cbic.201000330
577 Negri, L., Lattanzi, R., Giannini, E., Canestrelli, M., Nicotra, A., Melchiorri, P., 2009.
578 Bv8/Prokineticins and their Receptors A New Pronociceptive System. Int. Rev. Neurobiol. 85,
579 145–157. doi:10.1016/S0074-7742(09)85011-3
580 Negri, L., Lattanzi, R., Giannini, E., Metere, A., Colucci, M., Barra, D., Kreil, G., Melchiorri,
581 P., 2002. Nociceptive sensitization by the secretory protein Bv8. Br. J. Pharmacol. 137, 1147–
582 1154. doi:10.1038/sj.bjp.0704995
583 Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: a fast and
584 effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol.
585 Evol. 32, 268–274. doi:10.1093/molbev/msu300
586 Petersen, T.N., Brunak, S., von Heijne, G., Nielsen, H., 2011. SignalP 4.0: discriminating
587 signal peptides from transmembrane regions. Nat. Methods 8, 785–786.
588 doi:10.1038/nmeth.1701
589 Peters, R.S., Krogmann, L., Mayer, C., Donath, A., Gunkel, S., Meusemann, K., Kozlov, A.,
590 Podsiadlowski, L., Petersen, M., Lanfear, R., Diez, P.A., Heraty, J., Kjer, K.M., Klopfstein, S.,
591 Meier, R., Polidori, C., Schmitt, T., Liu, S., Zhou, X., Wappler, T., Rust, J., Misof, B.,
592 Niehuis, O., 2017. Evolutionary History of the Hymenoptera. Curr. Biol. 27, 1013–1018.
593 doi:10.1016/j.cub.2017.01.027
594 Peters, R.S., Meusemann, K., Petersen, M., Mayer, C., Wilbrandt, J., Ziesmann, T., Donath,
595 A., Kjer, K.M., Aspöck, U., Aspöck, H., Aberer, A., Stamatakis, A., Friedrich, F., Hünefeld,

F., Niehuis, O., Beutel, R.G., Misof, B., 2014. The evolutionary history of holometabolous insects inferred from transcriptome-based phylogeny and comprehensive morphological data. BMC Evol. Biol. 14, 52. doi:10.1186/1471-2148-14-52

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., Huelsenbeck, J.P., 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61, 539–542. doi:10.1093/sysbio/sys029

Sanner, M.F., 1999. Python: a programming language for software integration and development. J. Mol. Graph. Model. 17, 57–61.

Shelby, K.S., Perera, O.P., Snodgrass, G.L., 2015. Expression profiles of astakine-like transcripts in the tarnished plant bug, Lygus lineolaris, exposed to fungal spores of Beauveria bassiana. Insect Mol. Biol. 24, 480–490. doi:10.1111/imb.12175

Söderhäll, I., Kim, Y.-A., Jiravanichpaisal, P., Lee, S.-Y., Söderhäll, K., 2005. An ancient role for a prokineticin domain in invertebrate hematopoiesis. J. Immunol. Baltim. Md 1950 174, 6153–6160.

Suen, G., Teiling, C., Li, L., Holt, C., Abouheif, E., Bornberg-Bauer, E., Bouffard, P., Caldera, E.J., Cash, E., Cavanaugh, A., Denas, O., Elhaik, E., Favé, M.-J., Gadau, J., Gibson, J.D., Graur, D., Grubbs, K.J., Hagen, D.E., Harkins, T.T., Helmkampf, M., Hu, H., Johnson, B.R., Kim, J., Marsh, S.E., Moeller, J.A., Muñoz-Torres, M.C., Murphy, M.C., Naughton, M.C., Nigam, S., Overson, R., Rajakumar, R., Reese, J.T., Scott, J.J., Smith, C.R., Tao, S., Tsutsui, N.D., Viljakainen, L., Wissler, L., Yandell, M.D., Zimmer, F., Taylor, J., Slater, S.C., Clifton, S.W., Warren, W.C., Elsik, C.G., Smith, C.D., Weinstock, G.M., Gerardo, N.M., Currie, C.R., 2011. The genome sequence of the leaf-cutter ant Atta cephalotes reveals insights into its obligate symbiotic lifestyle. PLoS Genet. 7, e1002007. doi:10.1371/journal.pgen.1002007

Thomas, A., Sudheer, N.S., Kiron, V., Bright Singh, I.S., Narayanan, R.B., 2016. Expression

profile of key immune-related genes in Penaeus monodon juveniles after oral administration of recombinant envelope protein VP28 of white spot syndrome virus. Microb. Pathog. 96, 72–79. doi:10.1016/j.micpath.2016.05.002

Wechselberger, C., Puglisi, R., Engel, E., Lepperdinger, G., Boitani, C., Kreil, G., 1999. The mammalian homologues of frog Bv8 are mainly expressed in spermatocytes. FEBS Lett. 462, 177–181.

Whelan, S., Goldman, N., 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18, 691–699.

Wilson, W., Lowman, D., Antony, S.P., Puthumana, J., Bright Singh, I.S., Philip, R., 2015. Immune gene expression profile of Penaeus monodon in response to marine yeast glucan application and white spot syndrome virus challenge. Fish Shellfish Immunol. 43, 346–356. doi:10.1016/j.fsi.2014.12.032

Wurm, Y., Wang, J., Riba-Grognuz, O., Corona, M., Nygaard, S., Hunt, B.G., Ingram, K.K., Falquet, L., Nipitwattanaphon, M., Gotzek, D., Dijkstra, M.B., Oettler, J., Comtesse, F., Shih, C.-J., Wu, W.-J., Yang, C.-C., Thomas, J., Beaudoing, E., Pradervand, S., Flegel, V., Cook, E.D., Fabbretti, R., Stockinger, H., Long, L., Farmerie, W.G., Oakey, J., Boomsma, J.J., Pamilo, P., Yi, S.V., Heinze, J., Goodisman, M.A.D., Farinelli, L., Harshman, K., Hulo, N., Cerutti, L., Xenarios, I., Shoemaker, D., Keller, L., 2011. The genome of the fire ant Solenopsis invicta. Proc. Natl. Acad. Sci. U. S. A. 108, 5679–5684. doi:10.1073/pnas.1009690108

646

647

648 Figure legends

649 Figure 1. Sequence comparison of the prokineticin domain of arthropod astakines.

650 Highly conserved residues are colored green. Other residues are colored according to their

651 chemical properties. Ten cysteine residues, a proline residue between C_{79} and C_{81} and two

652 motifs ($R_{22}Y_{23}S_{24}$ and $Y_{57}P_{58}$) are conserved among most of the astakine sequences.

653

654 Figure 2. Deduced exon-intron structure of the astakine gene of a representative each from the

655 insects, chelicerates and crustaceans. All astakine genes investigated in the current study

656 contain the exon colored yellow in the figure. This exon ends with CCXX or CCXXX.

657

658 Figure 3a. A phylogenetic tree of 79 astakine protein sequences from arthropods analyzed

659 using the maximum likelihood method with the IQ-TREE software. The astakine-like

660 sequence from the hexapod *F. candida* was used as root sequence.

661 Bootstrap values are given at the nodes, and light grey or dark grey shading indicates clades

662 of closely related taxa. The scale bar indicates substitutions per site.

663 Figure 3b. A phylogenetic tree of astakine protein sequences from arthropods analyzed using

664 the MrBayes software. The astakine-like sequence from the hexapod *F. candida* was used as

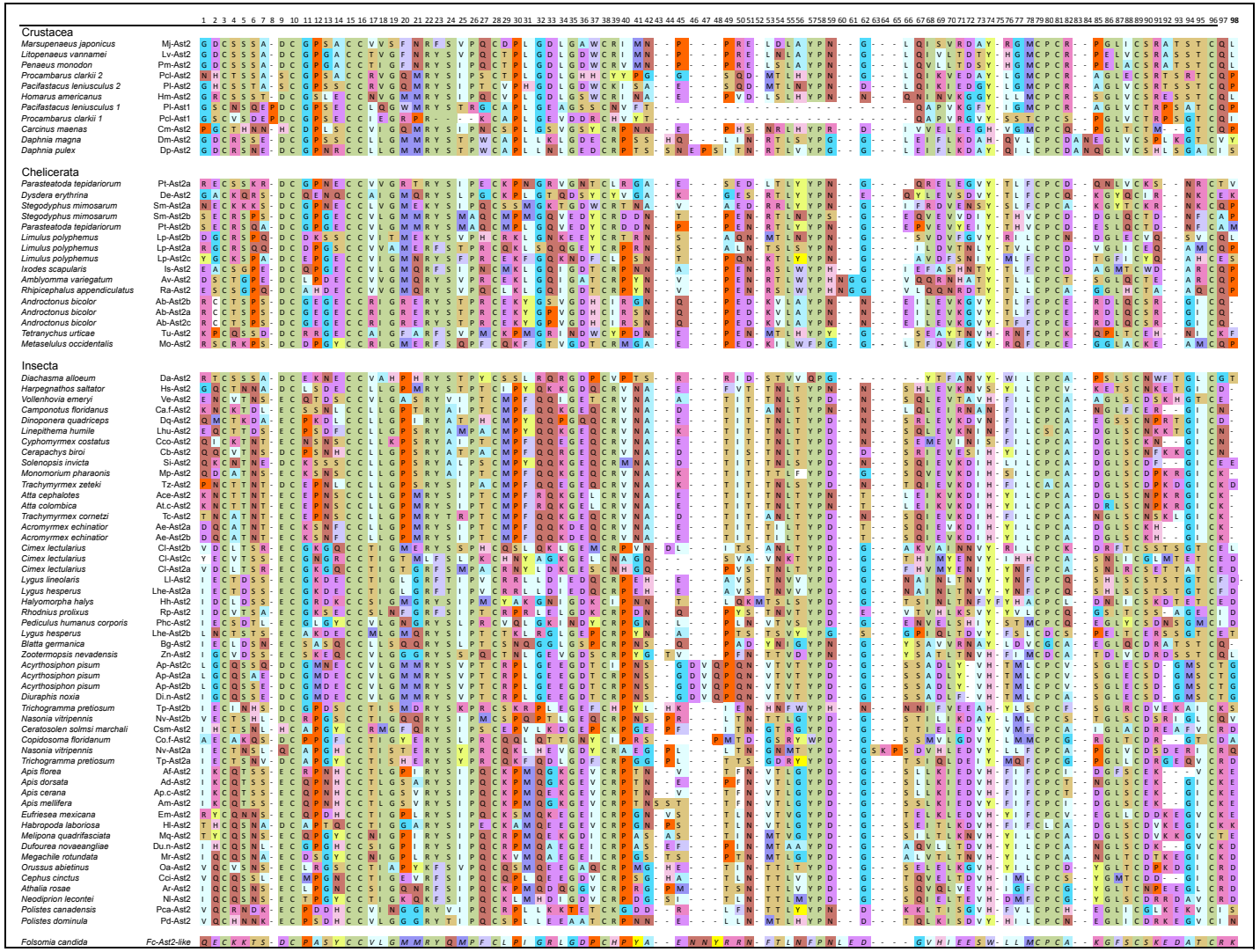
665 root sequence.

666 Node support values are given at the nodes, and light grey or dark grey shading indicates

667 clades of closely related taxa. The scale bar indicates substitutions per site.

668

669 Figure 4a. A phylogenetic tree of 33 astakine protein sequences with highest similarity to *P.*
670 *leniusculus* astakine 2 analyzed using the maximum likelihood method with the IQ-TREE
671 software. The astakine-like sequence from the hexapod *F. candida* was used as root sequence.
672 Bootstrap values are given at the nodes, and the scale bar indicates substitutions per site.
673
674 Figure 4b. A phylogenetic tree of 33 astakine protein sequences with highest similarity to *P.*
675 *leniusculus* astakine 2 analyzed with MrBayes software. The astakine-like sequence from the
676 hexapod *F. candida* was used as root sequence.
677 Node support values are given at the nodes, and the scale bar indicates substitutions per site.
678
679 Figure 5a. Overall fold and the core of six crustacean astakines, determined by using Mamba
680 intestinal toxin 1 as the template. The highly conserved cysteine residues and the RYS and
681 YP(N/D) motifs are colored, and the indel regions are shown by yellow crosses.
682
683 Figure 5b. RYS and YP(N/D) structures, showing the charge potential mapped to the surface,
684 of six crustacean astakines, determined by using Mamba intestinal toxin 1 as the template.
685


Crustacea
Marsupenaeus japonicus Mj-Ast2
Litopenaeus vannamei Lv-Ast2
Penaeus monodon Pm-Ast2
Procambarus clarkii 2 Pcl-Ast2
Pacifastacus leniusculus 2 Pl-Ast2
Homarus americanus Hm-Ast2
Pacifastacus leniusculus 1 Pl-Ast1
Procambarus clarkii 1 Pcl-Ast1
Carcinus maenas Cm-Ast2
Daphnia magna Dm-Ast2
Daphnia pulex Dp-Ast2
Chelicerata
Parasteatoda tepidariorum Pt-Ast2a
Dysdera erythrina De-Ast2a
Stegodyphus mimosarum Sm-Ast2a
Stegodyphus mimosarum Sm-Ast2b
Parasteatoda tepidariorum Pt-Ast2b
Limulus polyphemus Lp-Ast2b
Limulus polyphemus Lp-Ast2a
Limulus polyphemus Lp-Ast2c
Ixodes scapularis Is-Ast2
Amblyomma variegatum Av-Ast2
Rhipicephalus appendiculatus Ra-Ast2
Androctonus bicolor Ab-Ast2b
Androctonus bicolor Ab-Ast2a
Androctonus bicolor Ab-Ast2c
Tetranychus urticae Tu-Ast2
Metaseiulus occidentalis Mo-Ast2
Insecta
Diachasma alloeum Da-Ast2
Harpegnathos saltator Hs-Ast2
Vollenhovia emeryi Ve-Ast2
Camponotus floridanus Ca.f-Ast2
Dinoponera quadriceps Dq-Ast2
Linepithema humile Lhu-Ast2
Cyphomyrmex costatus Cco-Ast2
Cerapachys biroi Cb-Ast2
Solenopsis invicta Si-Ast2
Monomorium pharaonis Mp-Ast2
Trachymyrmex zeteki Tz-Ast2
Atta cephalotes Ace-Ast2
Atta colombica At.c-Ast2
Trachymyrmex cornetzi Tc-Ast2
Acromyrmex echinatior Ae-Ast2a
Acromyrmex echinatior Ae-Ast2b
Cimex lectularius Cl-Ast2b
Cimex lectularius Cl-Ast2c
Cimex lectularius Cl-Ast2a
Lygus lineolaris Ll-Ast2
Lygus hesperus Lhe-Ast2a
Halyomorpha halys Hh-Ast2
Rhodnius prolixus Rp-Ast2
Pediculus humanus corporis Ph-Ast2
Lygus hesperus Lhe-Ast2b
Blatta germanica Bg-Ast2
Zootermopsis nevadensis Zn-Ast2
Acyrthosiphon pisum Ap-Ast2c
Acyrthosiphon pisum Ap-Ast2a
Acyrthosiphon pisum Ap-Ast2b
Diuraphis noxia Dn-Ast2
Trichogramma pretiosum Tp-Ast2b
Nasonia vitripennis Nv-Ast2b
Ceratosolen solmsi marchali Csm-Ast2
Copidosoma floridanum Co.f-Ast2
Nasonia vitripennis Nv-Ast2a
Trichogramma pretiosum Tp-Ast2a
Apis florea Af-Ast2
Apis dorsata Ad-Ast2
Apis cerana Ap.c-Ast2
Apis mellifera Am-Ast2
Eufriesea mexicana Em-Ast2
Habropoda laboriosa Hl-Ast2
Melipona quadrifasciata Mq-Ast2
Dufourea novaeangliae Du.n-Ast2
Megachile rotundata Mr-Ast2
Orussus abietinus Oa-Ast2
Cephus cinctus Cci-Ast2
Athalia rosae Ar-Ast2
Neodiprion lecontei Nl-Ast2
Pcа-Ast2
Polistes canadensis
Polistes dominula Pd-Ast2
Folsomia candida Fc-Ast2-like

INSECTA

Megachile rotundata (Leafcutter Bee)

MTPIFVTLFLLFVLSCSSRAQTNRPDYIQCQSNAECDSGYCCNI
GPLRYSIPQCKVMQAEGEICRPGSTSPTNMTLGYPDGALVTLTN
VHYILCPCANGLTCDTKEGICKDTGEGHDTNRLFEEHKRHD

CHELICERATA

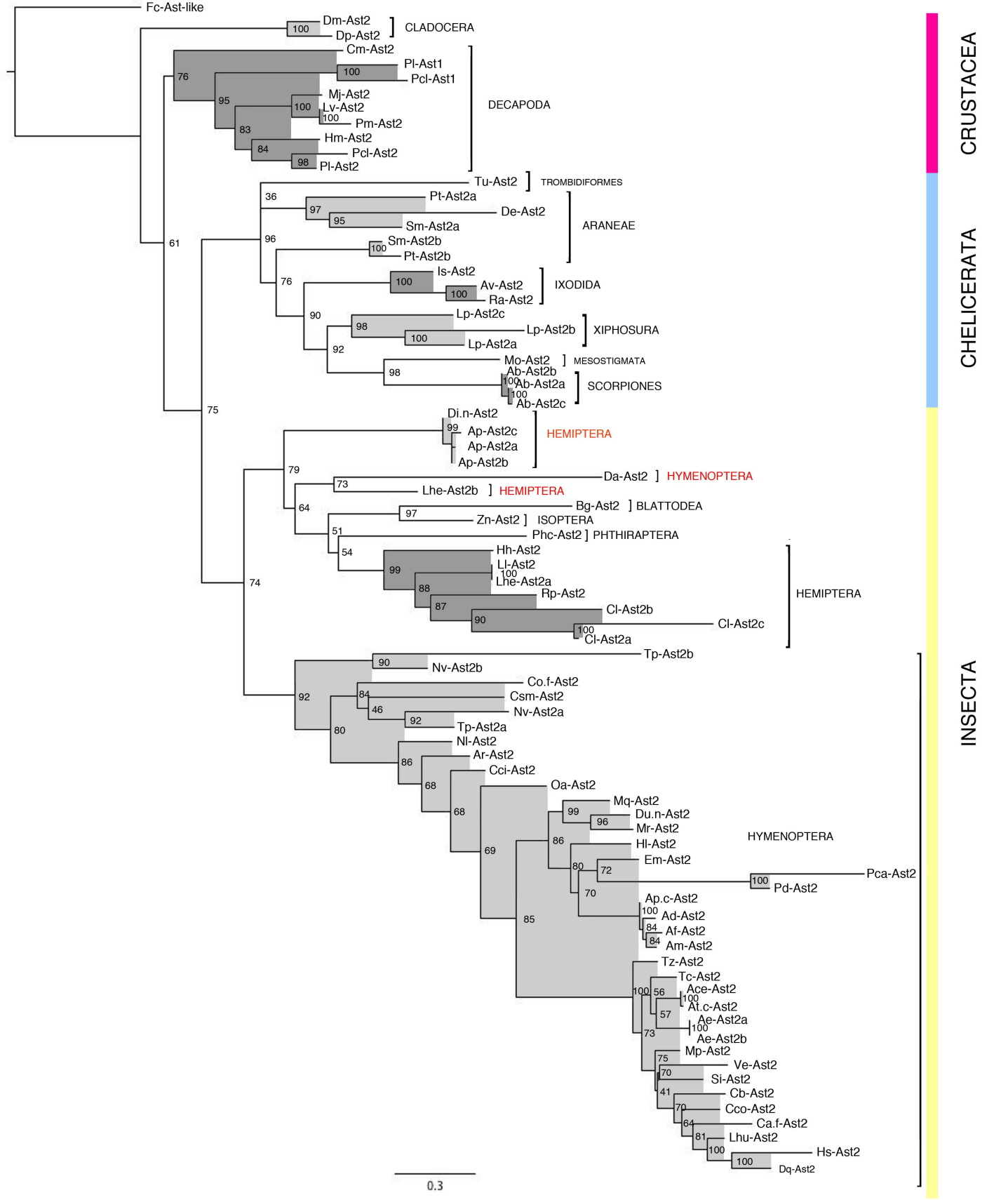
Parasteatoda tepidariorum (American House Spider)

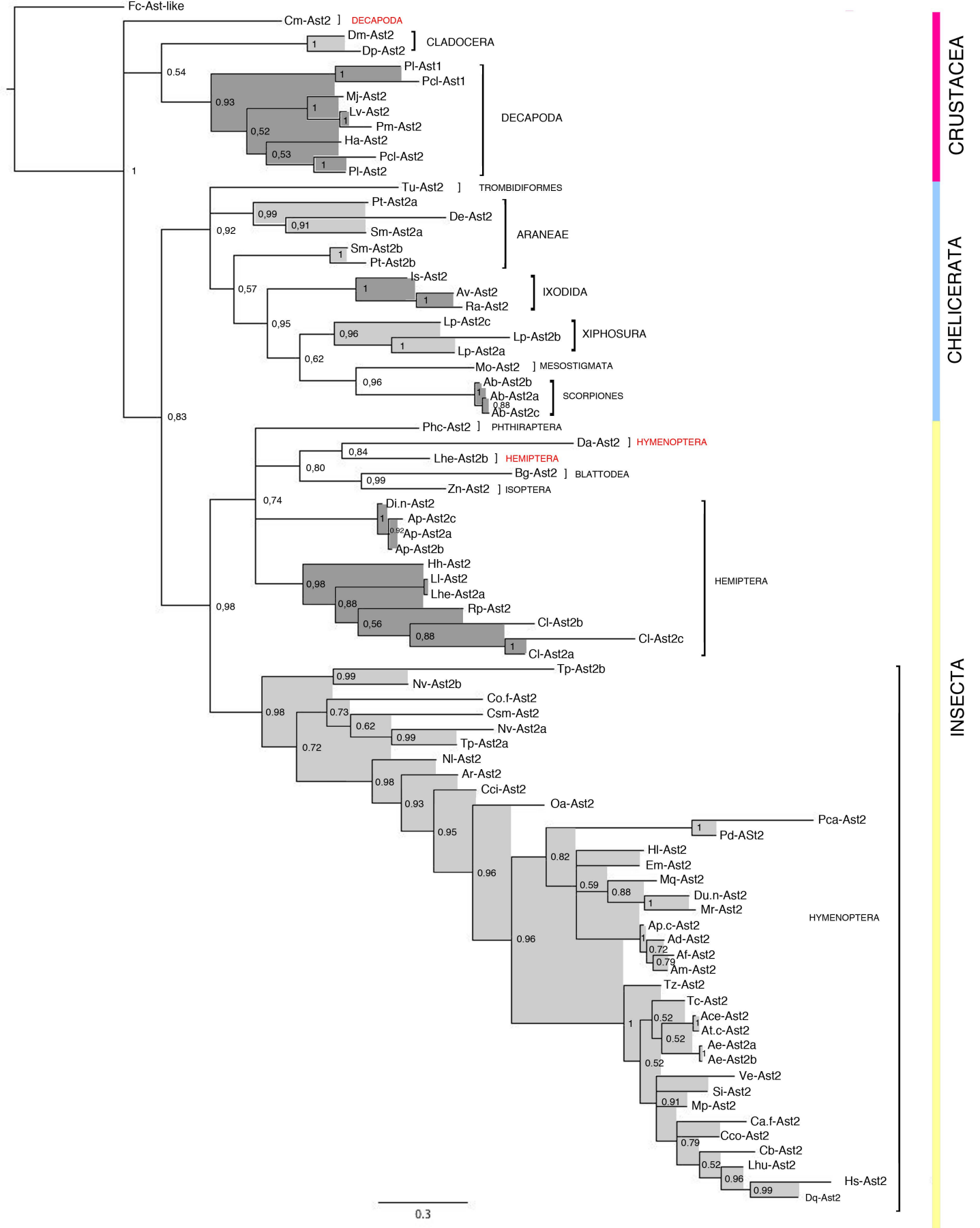
FTLSIVVSLLFQ
VCICNTPRECSSKRDCGPNECCVV
GRTRYSIPECKPNGRVGNTCLRGAESEDLTLYYPNGQRELEG
VYTLFCPCDQNLVCKSNRCTV

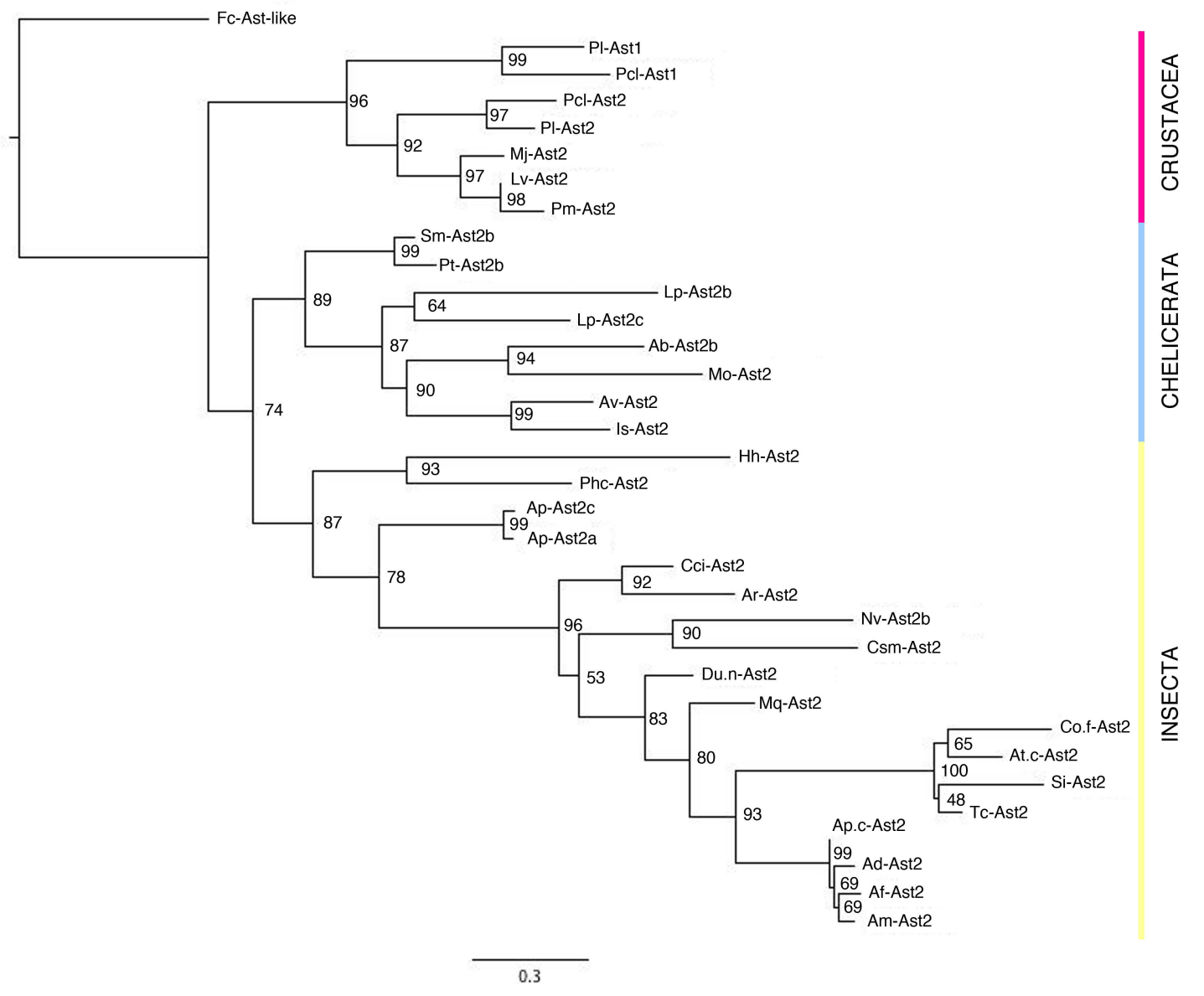
CRUSTACEA

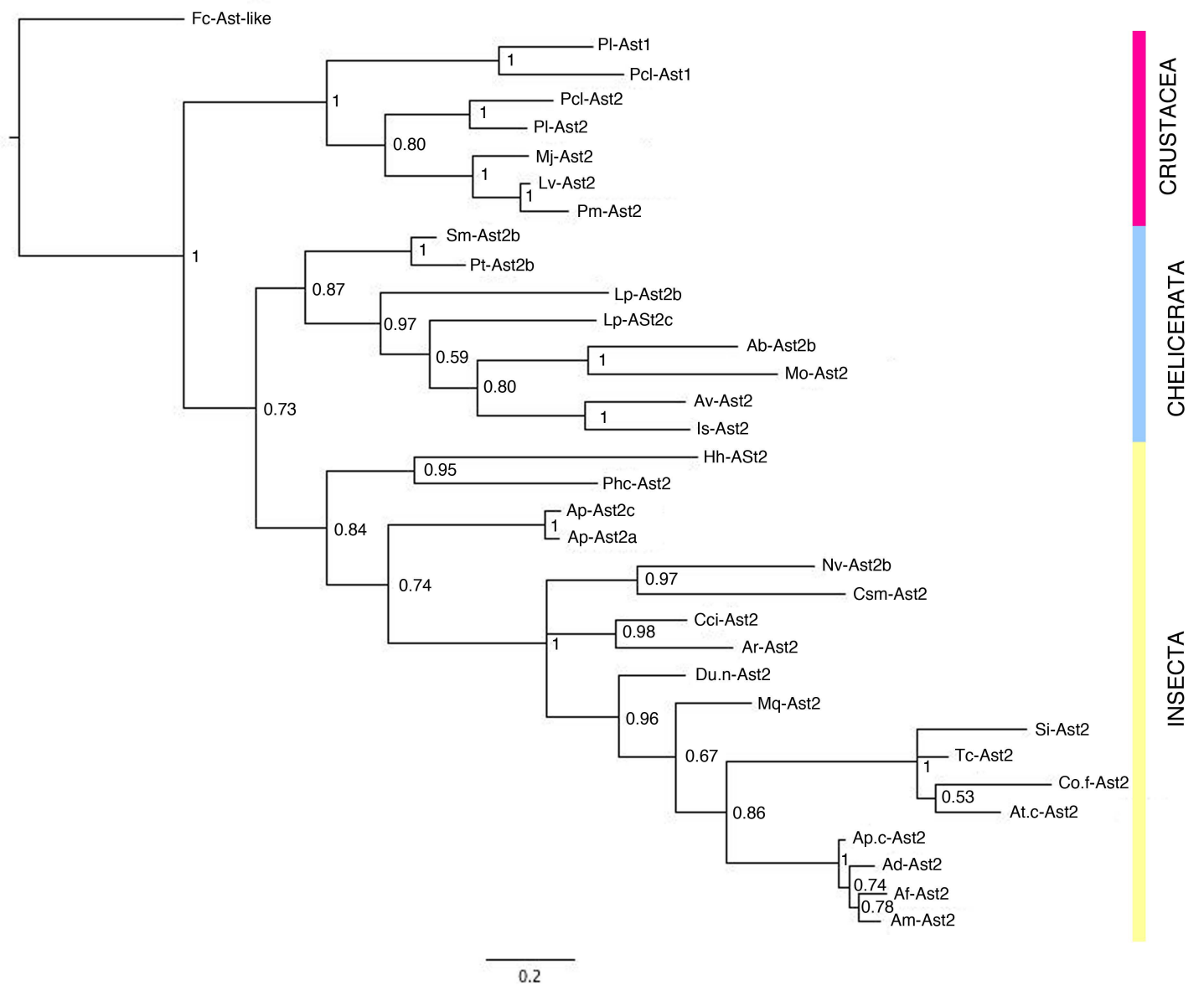
Daphnia magna (Water Flea)

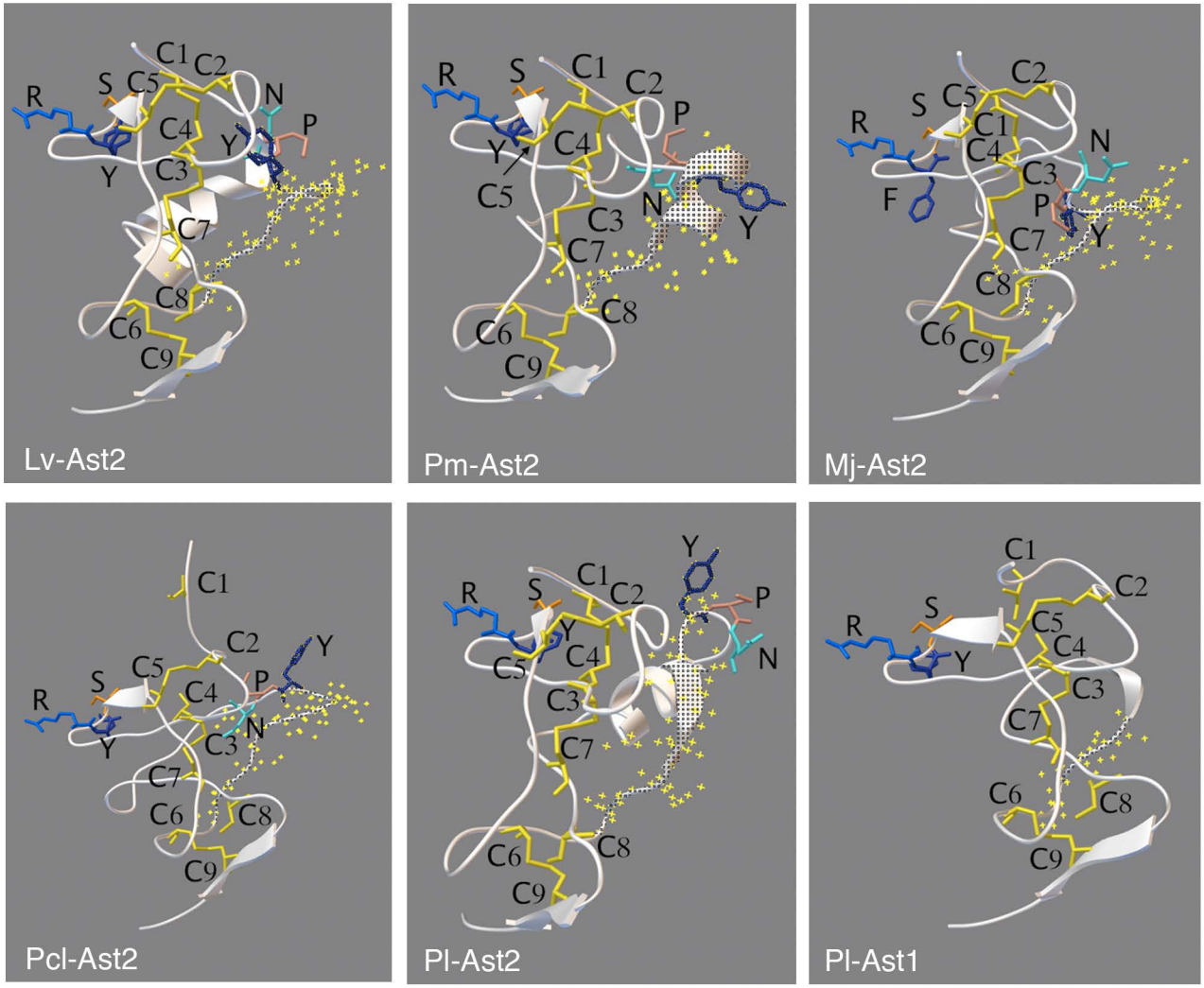
MLKECSLLFVCWTTLALTATLQPLPSYGVTGDCRSSEDCGPSSCCLL
GMMRYSTPWCAPLLKLGDECRPSSHQLINRTLSYPGGLEIFLKDAHQV
LCPCDANEGLVCSPLKGTCVYDVANDITPL

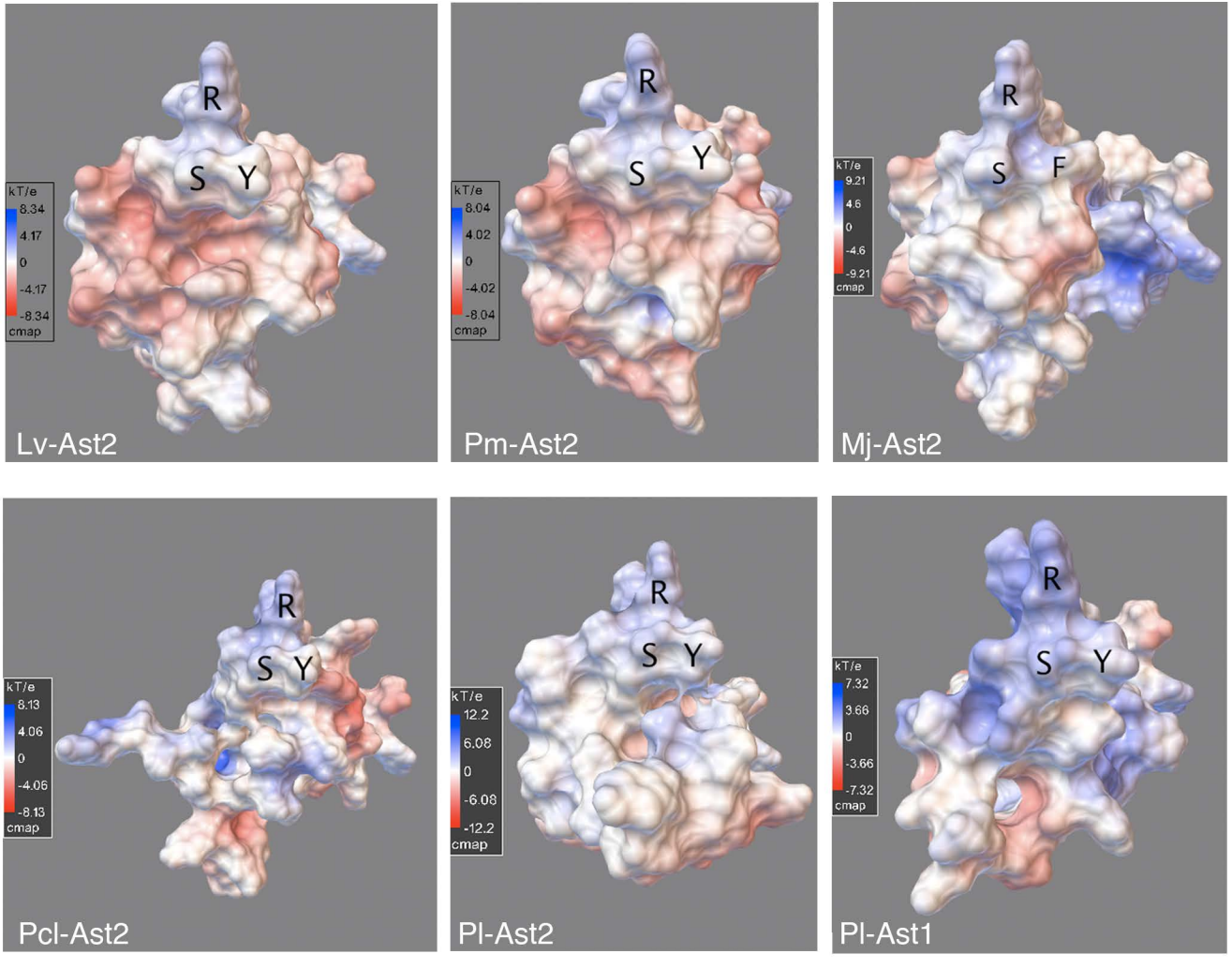

Fc-Ast-like
Dm-Ast2
Dp-Ast2
CLADOCERA
Cm-Ast2
Pl-Ast1
Pcl-Ast1
Mj-Ast2
Lv-Ast2
Pm-Ast2
Hm-Ast2
Pcl-Ast2
Pl-Ast2
DECAPODA
Tu-Ast2
TROMBIDIFORMES
Pt-Ast2a
De-Ast2
Sm-Ast2a
ARANEAE
Sm-Ast2b
Pt-Ast2b
Is-Ast2
Av-Ast2
Ra-Ast2
IXODIDA
Lp-Ast2c
Lp-Ast2b
Lp-Ast2a
XIPHOSURA
Mo-Ast2
MESOSTIGMATA
Ab-Ast2b
Ab-Ast2a
Ab-Ast2c
SCORPIONES
Di.n-Ast2
Ap-Ast2c
Ap-Ast2a
Ap-Ast2b
HEMIPTERA
Da-Ast2
HYMENOPTERA
Lhe-Ast2b
HEMIPTERA
Bg-Ast2
BLATTODEA
Zn-Ast2
ISOPTERA
Phc-Ast2
PHTHIRAPTERA
Hh-Ast2
Ll-Ast2
Lhe-Ast2a
Rp-Ast2
Cl-Ast2b
Cl-Ast2c
Cl-Ast2a
HEMIPTERA
Tp-Ast2b
Nv-Ast2b
Co.f-Ast2
Csm-Ast2
Nv-Ast2a
Tp-Ast2a
Nl-Ast2
Ar-Ast2
Cci-Ast2
Oa-Ast2
Mq-Ast2
Du.n-Ast2
Mr-Ast2
Hl-Ast2
Em-Ast2
Pca-Ast2
Pd-Ast2
HYMENOPTERA
Ap.c-Ast2
Ad-Ast2
Af-Ast2
Am-Ast2
Tz-Ast2
Tc-Ast2
Ace-Ast2
At.c-Ast2
Ae-Ast2a
Ae-Ast2b
Mp-Ast2
Ve-Ast2
Si-Ast2
Cb-Ast2
Cco-Ast2
Ca.f-Ast2
Lhu-Ast2
Hs-Ast2
Dq-Ast2
0.3
CRUSTACEA
CHELICERATA
INSECTA

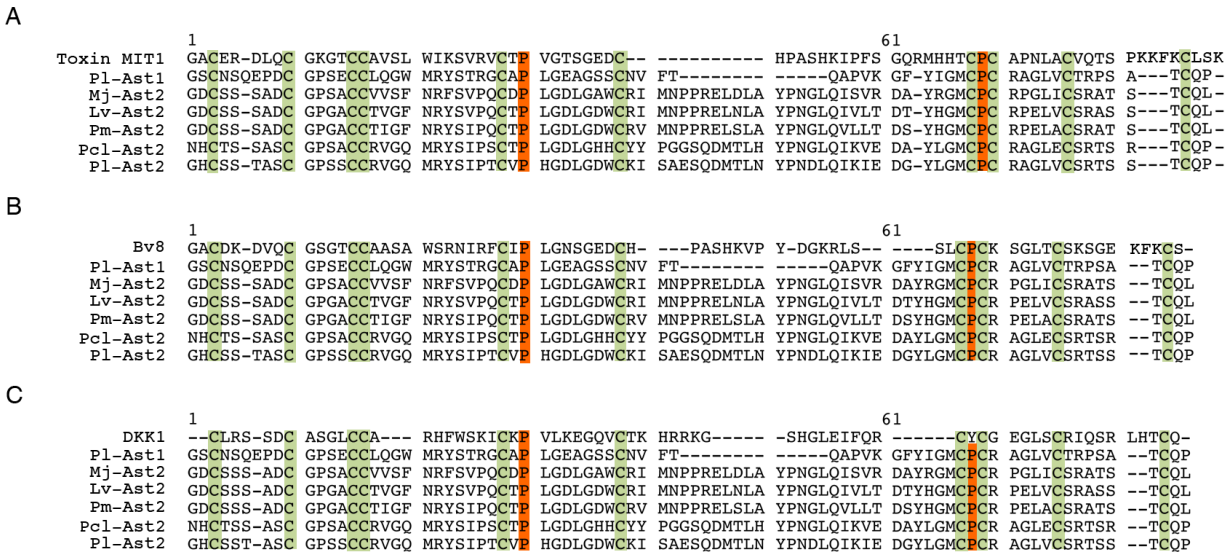

Fc-Ast-like
Cm-Ast2
DECAPODA
Dm-Ast2
Dp-Ast2
CLADOCERA
Pl-Ast1
Pcl-Ast1
Mj-Ast2
Lv-Ast2
Pm-Ast2
Ha-Ast2
Pcl-Ast2
Pl-Ast2
DECAPODA
Tu-Ast2
TROMBIDIFORMES
Pt-Ast2a
De-Ast2
Sm-Ast2a
ARANEAE
Sm-Ast2b
Pt-Ast2b
Is-Ast2
Av-Ast2
Ra-Ast2
IXODIDA
Lp-Ast2c
Lp-Ast2b
Lp-Ast2a
XIPHOSURA
Mo-Ast2
MESOSTIGMATA
Ab-Ast2b
Ab-Ast2a
Ab-Ast2c
SCORPIONES
Phc-Ast2
PHTHIRAPTERA
Da-Ast2
HYMENOPTERA
Lhe-Ast2b
HEMIPTERA
Bg-Ast2
BLATTODEA
Zn-Ast2
ISOPTERA
Di.n-Ast2
Ap-Ast2c
Ap-Ast2a
Ap-Ast2b
Hh-Ast2
Ll-Ast2
Lhe-Ast2a
Rp-Ast2
Cl-Ast2b
Cl-Ast2c
Cl-Ast2a
HEMIPTERA
Tp-Ast2b
Nv-Ast2b
Co.f-Ast2
Csm-Ast2
Nv-Ast2a
Tp-Ast2a
Nl-Ast2
Ar-Ast2
Cci-Ast2
Oa-Ast2
Pca-Ast2
Pd-ASt2
Hl-Ast2
Em-Ast2
Mq-Ast2
Du.n-Ast2
Mr-Ast2
Ap.c-Ast2
Ad-Ast2
Af-Ast2
Am-Ast2
HYMENOPTERA
Tz-Ast2
Tc-Ast2
Ace-Ast2
At.c-Ast2
Ae-Ast2a
Ae-Ast2b
Ve-Ast2
Si-Ast2
Mp-Ast2
Ca.f-Ast2
Cco-Ast2
Cb-Ast2
Lhu-Ast2
Hs-Ast2
Dq-Ast2
0.3
CRUSTACEA
CHELICERATA
INSECTA


Fc-Ast-like
Pl-Ast1
Pcl-Ast1
Pcl-Ast2
Pl-Ast2
Mj-Ast2
Lv-Ast2
Pm-Ast2
Sm-Ast2b
Pt-Ast2b
Lp-Ast2b
Lp-Ast2c
Ab-Ast2b
Mo-Ast2
Av-Ast2
Is-Ast2
Hh-Ast2
Phc-Ast2
Ap-Ast2c
Ap-Ast2a
Cci-Ast2
Ar-Ast2
Nv-Ast2b
Csm-Ast2
Du.n-Ast2
Mq-Ast2
Co.f-Ast2
At.c-Ast2
Si-Ast2
Tc-Ast2
Ap.c-Ast2
Ad-Ast2
Af-Ast2
Am-Ast2
99
96
97
92
97
98
99
89
64
87
94
90
99
74
93
87
99
78
92
96
90
53
83
80
65
100
48
93
99
69
69
0.3
CRUSTACEA
CHELICERATA
INSECTA

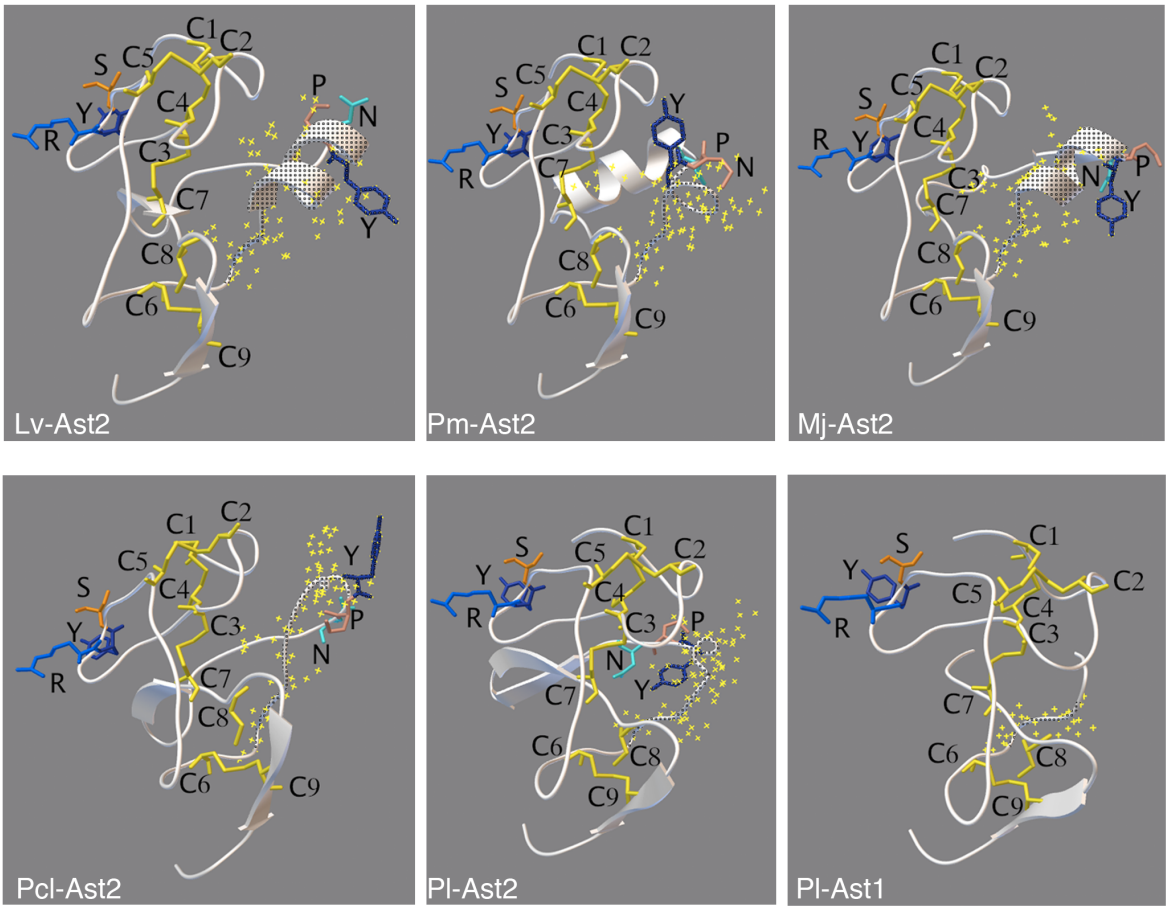

Fc-Ast-like
Pl-Ast1
Pcl-Ast1
Pcl-Ast2
Pl-Ast2
Mj-Ast2
Lv-Ast2
Pm-Ast2
Sm-Ast2b
Pt-Ast2b
Lp-Ast2b
Lp-ASt2c
Ab-Ast2b
Mo-Ast2
Av-Ast2
Is-Ast2
Hh-ASt2
Phc-Ast2
Ap-Ast2c
Ap-Ast2a
Nv-Ast2b
Csm-Ast2
Cci-Ast2
Ar-Ast2
Du.n-Ast2
Mq-Ast2
Si-Ast2
Tc-Ast2
Co.f-Ast2
At.c-Ast2
Ap.c-Ast2
Ad-Ast2
Af-Ast2
Am-Ast2
1
0.80
0.87
0.97
0.59
0.73
0.95
0.84
0.74
0.98
0.96
0.67
0.53
0.86
0.78
CRUSTACEA
CHELICERATA
INSECTA
0.2


Lv-Ast2
Pm-Ast2
Mj-Ast2
Pcl-Ast2
Pl-Ast2
Pl-Ast1

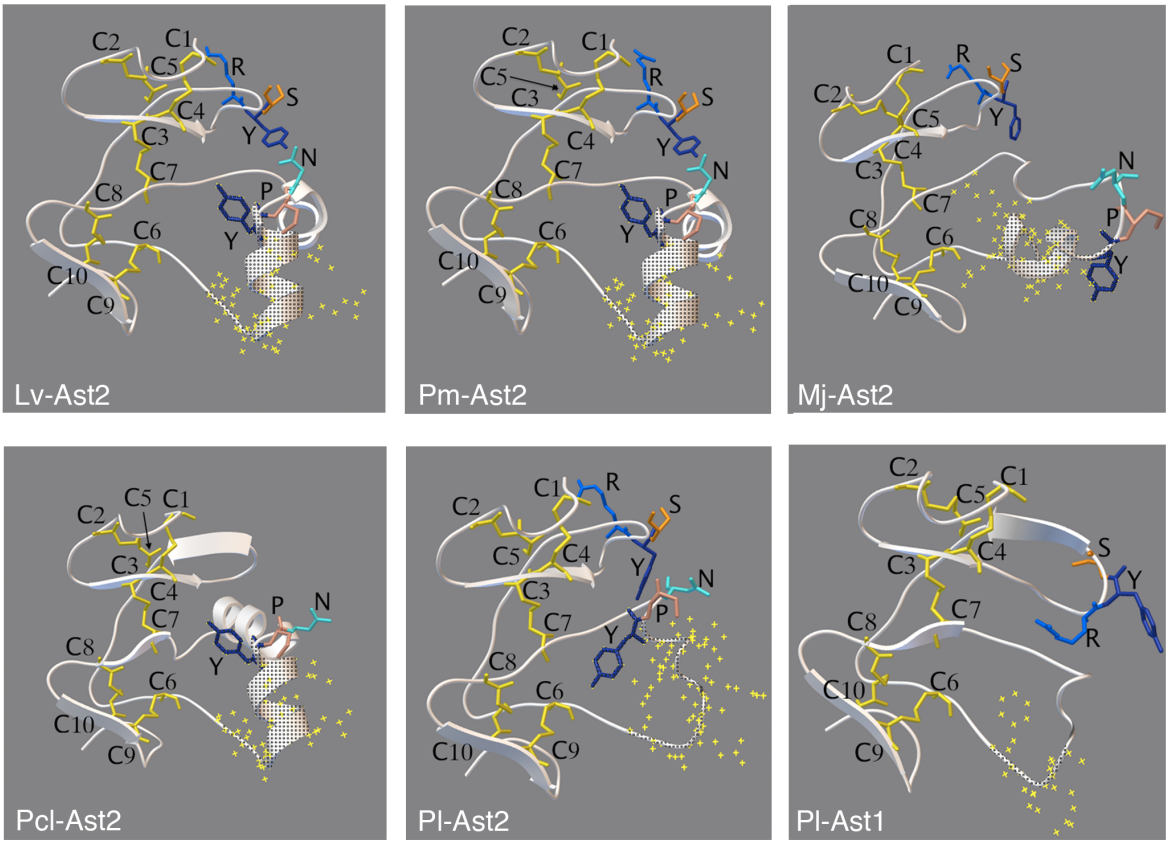

R
S Y
kT/e
8.34
4.17
0
-4.17
-8.34
cmap
Lv-Ast2
R
S Y
kT/e
8.04
4.02
0
-4.02
-8.04
cmap
Pm-Ast2
R
S F
kT/e
9.21
4.6
0
-4.6
-9.21
cmap
Mj-Ast2
R
S Y
kT/e
8.13
4.06
0
-4.06
-8.13
cmap
Pcl-Ast2
R
S Y
kT/e
12.2
6.08
0
-6.08
-12.2
cmap
Pl-Ast2
R
S Y
kT/e
7.32
3.66
0
-3.66
-7.32
cmap
Pl-Ast1



Supplementary Figure 1. Sequence alignment using Muscle software of the three top-scoring proteins used as templates in Phyre2 to build three-dimensional models of six crustacean astakines A) MIT1, mamba intestinal toxin 1 from *Dendroaspis polyepis* (PDB climtA), B) Bv8 from *Bombina variegata* (PDB 2KRA), C) Dickkopf-related protein 1 (DKK1) from *Homo sapiens* (PDB 3S8V). Highly conserved cysteine residues and one proline residue conserved in almost all sequences are colored.



Supplementary Figure 2. The overall fold and the core of six crustacean astakines using the prokineticin Bv8 from *Bombina variegata* as model.
The highly conserved cystein residues and the RYS and YP(N/D) motifs are colored and the indels regions are shown by yellow crosses.



Supplementary Figure 3. The overall fold and the core of six crustacean astakines using Dickkopf-related protein 1 from *Homo sapiens* as model.
The highly conserved cystein residues and the RYS and YP(N/D) motifs are colored and the indels regions are shown by yellow crosses.

Supplementary Table 1. List of astakines included in the study.

Arthropods	Protein name	Species	pI	Mw	Accession number
Crustacea					
	Mj-Ast2	*Marsupenaeus japonicus*	4.88	11.28	**BAJ34645.1**
	Lv-Ast2	*Litopenaeus vannamei*	4.68	11.32	**ADM53424.1**
	Pm-Ast2	*Penaeus monodon*	5.13	11.30	**AAX14636.1**
	Pcl-Ast2	*Procambarus clarkii*	7.69	10.60	**AEC50077.1**
	Pl-Ast2	*Pacifastacus leniusculus*	7.04	11.19	**ABQ23255.1**
	Ha-Ast2	*Homarus americanus*	4.84	10.28	*FE535609*
	Pl-Ast1	*Pacifastacus leniusculus*	5.00	8.73	**AAX14635.1**
	Pcl-Ast1	*Procambarus clarkii*	5.08	8.15	*
	Cm-Ast2	*Carcinus maenas*	5.07	10.25	*DW585080*
	Dm-Ast2	*Daphnia magna*	4.85	11.48	**KZS05559.1,** *LRGB01002901.1*
	Dp-Ast2	*Daphnia pulex*	4.54	11.95	*FE329237*
Chelicerata					
	Pt-Ast2a	*Parasteatoda tepidariorum*	7.62	9.27	**XP_015917287.1**, *XM_016061801.1*
	De-Ast	*Dysdera erythrina*	5.31	10.18	*CV178181*
	Sm-Ast2a	*Stegodyphus mimosarum*	8.54	9.85	**KFM62184**
	Sm-Ast2b	*Stegodyphus mimosarum*	3.93	10.66	**KFM69031.1**
	Pt-Ast2b	*Parasteatoda tepidariorum*	3.92	11.02	**XP_015920703**
	Lp-Ast2b	*Limulus polyphemus*	6.07	10.20	**XP_013776682.1** *XM_013921228.1*
	Lp-Ast2a	*Limulus polyphemus*	6.70	10.40	**XP_013775587.1** *XM_013920133.1*
	Lp-Ast2c	*Limulus polyphemus*	5.09	9.61	**XP_013785672.1**
	Is-Ast2	*Ixodes scapularis*	4.60	11.29	*EW845057*
	Av-Ast2	*Amblyomma variegatum*	4.92	11.73	**DAA34752.1**
	Ra-Ast2	*Rhipicephalus appendiculatus*	5.06	11.62	*CD794853*
	Ab-Ast2b	*Androctonus bicolor*	5.31	8.74	**AIX87718.1**
	Ab-Ast2a	*Androctonus bicolor*	5.31	8.78	**AIX87717.1**
	Ab-Ast2c	*Androctonus bicolor*	5.31	8.81	**AIX87719.1**
	Tu-Ast2	*Tetranychus urticae*	7.66	11.46	**XP_015792820.1,** *XM_015937334.1*
	Mo-Ast2	*Metaselulus occidentalis*	4.59	13.76	**XP_003743670,** *XM_003743622.1*
Insecta	Da-Ast2	*Diachasma alloeum*	6.94	10.98	**XP_015127605.1**
	Hs-Ast2	*Harpegnathos saltator*	5.15	9.29	**EFN86043.1,** *GL447755.1*
	Ve-Ast2	*Vollenhovia emeryi*	5.81	14.95	**XP_011861926.1**,

					XM_012006536.1
	Ca.f-Ast2	*Camponotus floridanus*	7.63	9.63	**XP_011266500.1**, *XM_011268198.1*
	Dq-Ast2	*Dinoponera quadriceps*	5.11	9.19	**XP_014486827.1**
	Lhu-Ast2	*Linepithema humile*	9.07	16.07	**XP_012228085.1**, *XM_012372662.1*
	Cco-Ast2	*Cyphomyrmex costatus*	4.60	9.38	**KYN05408.1**
	Cb-Ast2	*Cerapachys biroi*	6.05	9.69	**EZA54888.1,** *KK107235.1*
	Si-Ast2	*Solenopsis invicta*	4.53	14.80	**EFZ14050.1,** *GL767121.1*
	Mp-Ast2	*Monomorium pharaonis*	7.43	9.84	**XP_012539106.1,** *XM_012683652.1*
	Tz-Ast2	*Trachymyrmex zeteki*	4.49	12.59	**KYQ52101.1**
	Ace-Ast2	*Atta cephalotes*	9.27	17.89	**XP_012063524.1,** *XM_012208134.1*
	At.c-Ast2	*Atta colombica*	8.57	9.92	**KYM75707.1**
	Tc-Ast2	*Trachymyrmex cornetzi*	6.48	9.71	**KYN16810.1**
	Ae-Ast2a	*Acromyrmex echinatior*	5.49	9.87	**EGI67870.1,** *GL888084.1*
	Ae-Ast2b	*Acromyrmex echinatior*	8.67	17.00	**XP_011050611.1**
	Cl-Ast2b	*Cimex lectularius*	8.58	10.75	**XP_014256495.1**
	Cl-Ast2c	*Cimex lectularius*	7.50	10.08	**XP_014256561**
	Cl-Ast2a	*Cimex lectularius*	4.94	11.47	**XP_014256494.1**
	Ll-Ast2a	*Lygus lineolaris*	4.19	12.39	**AJR27902.1**
	Lhe-Ast2a	*Lygus hesperus*	4.19	12.39	**JAG12052.1**
	Hh-Ast2	*Halyomorpha halys*	5.53	12.79	**XP_014294052.1,** *XM_014438566.1*
	Rp-Ast2	*Rhodnius prolixus*	4.33	11.80	**JAA75272.1**
	Phc-Ast2	*Pediculus humanus corporis*	4.27	10.73	**XP_002431167.1,** *XM_002431122.1*
	Lhe-Ast2b	*Lygus hesperus*	5.25	11.95	**JAG7549.1**
	Bg-Ast2	*Blatta germanica*	4.46	9.41	**FG125716**
	Zn-Ast2	*Zootermopsis nevadensis*	5.13	11.26	**KDR15950.1**
	Ap-Ast2c	*Acyrthosiphon pisum*	4.24	10.15	*EX635976.1*
	Ap-Ast2a	*Acyrthosiphon pisum*	4.16	10.18	*FF336283.1*
	Ap-Ast2b	*Acyrthosiphon pisum*	4.02	10.02	*EX610914.1*
	Di.n-Ast2	*Diuraphis noxia*	4.29	12.41	**XP_015368848.1**
	Tp-Ast2b	*Trichogramma pretiosum*	6.63	10.87	**XP_014230684.1,** *XM_014375198.1*
	Nv-Ast2b	*Nasonia vitripennis*	5.16	11.02	**XP_001605660.1**, *XM_001605610.3*
	Csm-Ast2	*Ceratosolen solmsi marchali*	5.02	11.94	**XP_011499199.1**, *XM_011500897.1*
	Co.f-Ast2	*Copidosoma floridanum*	7.63	9.63	**XP_011266500.1**, *XM_011268198.1*
	Nv-Ast2a	*Nasonia vitripennis*	5.07	13.41	**XP_008213060.1,** *XM_008214838*
	Tp-Ast2a	*Trichogramma pretiosum*	4.29	12.59	**XP_014235233.1,** *XM_014379747.1*
	Af-Ast2	*Apis florea*	7.64	9.74	**XP_012345115.1,** *XM_012489661.1*
	Ad-Ast2	*Apis dorsata*	6.86	9.65	**XP_006617146.1**, *XM_006617083.1*
	Ap.c-Ast2	*Apis cerana*	5.56	9.71	**XP_016909725.1**,

					XM_017054236.1
	Am-Ast2	*Apis mellifera*	6.71	9.84	**XP_003250271.1,** *XM_003250223*
	Em-Ast2	*Eufriesea mexicana*	4.72	10.50	**OAD53282.1**
	Hl-Ast2	*Habropoda laboriosa*	5.10	11.85	**KOC63709.1**
	Mq-Ast2	*Melipona quadrifasciata*	7.99	14.09	**KOX80272.1**
	Du.n-Ast2	*Dufourea novaeangliae*	4.87	11.97	**KZC05910.1**
	Mr-Ast2	*Megachile rotundata*	5.19	12.02	**XP_003708605.1**, *XM_003708557.1*
	Oa-Ast2	*Orussus abietinus*	6.50	17.02	**XP_012283423.1,** *XM_012428000.1*
	Cci-Ast2	*Cephus cinctus*	4.84	14.50	**XP_015602971.1**
	Ar-Ast2	*Athalia rosae*	5.21	12.19	**XP_012261298.1,** *XM_012405875.1*
	Nl-Ast2	*Neodiprion lecontei*	4.86	12.51	**XP_015520701.1**
	Pca-Ast2	*Polistes canadensis*	6.98	10.71	**XP_014616409.1,** *XM_014760923.1*
	Pd-Ast2	*Polistes dominula*	5.18	10.56	**XP_015184135.1,** *XM_015328649.1*
Root sequence	Fc-Ast-like	*Folsomia candida*	5.08	12.81	**XP_021968286.1**

Supplementary Table 2. Arthropod astakine gene structures.
(https://www.ncbi.nlm.nih.gov/genome/annotation_euk/all/)

1) Insecta

Harpegnathos saltator (taxonomy ID: 610380) Hs-Ast2

MSSILSILLIITVGLVFSSNGQCTNNADCLSDECCLL
Intron – 71 bp –
GPMRYSTPTCIPYQKKGDQCRVNAEFVTTNLTYPNNSHLEVKNVSYILCPCVKETSCNKETGICD

Camponotus floridanus (taxonomy ID: 104421) Ca.f-Ast2

MSLMSNVLLLITLAGIVPAFPFNSFKNCKTDLECSSNLCCLL
Intron – 225 bp –
*LL*GPTRYAIPTCMPFQQKGEQCRVNADTITANLTYPNNLQLEIRNANFILCPCANGLFCERGICN

Cerapachys biro (Taxonomy ID: 443821) Cb-Ast2

MSSILGLLLLISIAVAVPTSRTQQCVTNSDCPSNHCCLL
Intron – 183 bp –
GPSRYATPACMPFQQRGEQCRVNADTISTNLTYPDDSRIEVESIHYILCPCADGLSCNFKKGICN

Monomorium pharaonis (taxonomy ID: 307658) Mp-Ast2

MSPISGILIFVISIVATSNIGSVTSSSQDCATNSECKSNSCCLL
Intron – 236 bp –
GPSRYAIPTCMPFQQKGEQCRVNAKTITTTLFYPDGSQVEVKDIHSILCPCADGLSCDPKRGICK

Nasonia vitripennis (taxonomy ID: 7425) Nv-Ast2a

MKMIMRLGLLLLCAMVINTKALARFPRNWNSHIECTNSLQCAPGHCCTI
Intron – 128 bp –
STERYSYPRCQKLHEVGDYCRAEGPLLTNGNMTYPDGSKPSDVHLEDVYLLFCPCAPGLVCDSDERI
CRQPSDMKDFNYLKEQETGSNKSDD

Nasonia vitripennis (taxonomy ID: 7425) Nv-Ast2b

MTSAVLLLSLMIGSLYAAAIEEQSIPPSWVECTSHLDCRPGSCCTI
Intron – 238 bp –
GQQRYSIPMCSPQPTLGEQCRPNSPRLTNTTLGYPDGSTILIKDAYLMLCPCSTGLSCDSRIGLCQV
KQQAEA

Trichogramma pretiosum (taxonomy ID: 7493) Tp-Ast2a

MLAARIKFSLSLLLLVGAVNAATLHEQTGLLADEIECTSNVDCAPGYCCTI
Intron – 86 bp –
*S*HERYSYPRCQKFQDLGDFCRPGGPLTTSGDRYYPDGTSIQLDEIYMQFCPCGPGLLCDRGEQVCRD
ASDFNSVQLNQSGKSDD

Trichogramma pretiosum (taxonomy ID: 7493) Tp-Ast2b

MCRLFSFLFLSFAIIHLIDANDYRSGVIECINHSDCGPDSCCTIS
Intron – 523 bp –
MDRYSKPRCSKRPLEGEFCHPYLHKIENHNFWYPHNNNIFVEEAHYLSCPCFSGLRCDVEKAICKSK
IA

Ceratosolen solmsi marchali (taxonomy ID: 326594) Csm-Ast2

MSRLLLSLLLLFHITAILIRDVRAGHPSWIHCTSNLHCAPGYCCRM
Intron – 1686 bp –
GFQRYSIPSCEPVLKDGEPCKPGEPFITNGTRGYPDGTTIELEDVYVMFCPCAIGLACDREAFVCRD
ASEMKDFNHLSGKSDKTDD

Apis mellifera (taxonomy ID: 7460) Am-Ast2

MMTSIFEISFLLFVLAYPCHAQNDYIKCQTSSECQPNHCCT*L*
Intron – 96 bp –
*L*GSVRYSIPQCKPMQGKGEVCRPTNSSTTFNVTLGYPDGSSLKIEDVYFIFCPCIDGLSCEKGICKE
KN

Apis dorsata (taxonomy ID: 7462) Ad-Ast2

MMTSIFVISFLLFVLAYPCHAQNDYIKCQTSSECQPNHCCTL
Intron – 110 bp –
*L*GSARYSIPQCKPMQGKGEVCRPTNEPFNVTLGYPDGSLLKIEDVHFIFCPCTNGLSCEKGICKEKN

Apis florea (taxonomy ID: 7463) Af-Ast2

MMTSIFVISFLLFVLAYPCHTQNDYIKCQTSSECRPNHCCTL
Intron – 95 bp –
GPIRYSIPQCKPMQGKGEVCRPTNVTFNVTLGYPDGSLLKIEDVHFIFCPCIDGFSCEKGVCKEKN

Megachile rotundata (taxonomy ID: 143995) Mr-Ast2

MTPIFVTLFLLFVLSCSSRAQTNRPDYIQCQSNAECDSGYCCNI
Intron – 95 bp –
GPLRYSIPQCKVMQAEGEICRPGSTSPTNMTLGYPDGALVTLTNVHYILCPCANGLTCDTKEGICKD
TGEGHDTNRLFEEHKRHD

Acyrthosiphon pisum (taxonomy ID: 7029) Ap-Ast2c

MNTNIMGKLSLVAVTILVATVAAYPSKPSFLGCQSSQDCGMNECCVL*G*
Intron – ? bp –
*G*MMRYSVPTCRPLGEEGDTCIPNSGDVQPQNVTVTYPDGSSADLYVHTMLCPCVSGLECSDGMSCTG
LNGAGKLMPAPRLHGRG

Orussus abietinus (taxonomy ID: 222816) Oa-Ast2

MTQKLILAMILAIAGMSGPIIAGVASNRPSHVQCVSNSECLRGSCCTI
Intron – 946 bp –
APYKFSVPQCQSMQEEGAQCRPMGHETINTTLTYPDGSELELKGVHYILCPCDYGLTCDPKDGICRD
VSQRRDFNHLQNEAIAHED

Athalia rosae (taxonomy ID: 37344) Ar-Ast2

MTQSLTFAAIVVIVGICQVGYTSARVTMRPPYIQCQSNSECLPGNCCSI
Intron – 69 bp –
GQNRFSIPQCKPMQDQGGVCRPRGPMTSNTTLVYPDGSQVQLVEVHIGFCPCGYGLTCNPEEGLCRD
PSQRRGFNSLLDEASVQDD

Polistes canadensis (taxonomy ID. 91411) Pca-Ast2

MSSTISTSFTILLLLGLVSFLFAASIKQEEPPDVQCRNDKECPDDHCCV*I*
Intron – 81 bp –
*I*NGGRYVIPQCRPLLKKTETCKGDDRLFNTTLYYPNDKKLTISGVHFVLCPCHEGLICGLKEKVCIS
NN

Polistes dominula (taxonomy ID: 743375) Pd-Ast2

MSSTISTSFTLLLLLGLVSFLFAASIKQEEPANVQCHNNKECPSDHCCVL
Intron – 68 bp –
GGGRYTIPQCSPLLEEAATCRPNNELLNMTLHYPNDTQLKISDVYHILCPCNEGLICDRKEGVCINN
N

Halyomorpha halys (taxonomy ID: 2867706) Hh-Ast2

MSMTLFQLGAIASIFLTVYAMPNDRPGYIDCLDSSECGRDKCCSI
Intron – 2710 bp –
*CSI*GMGRYSIPMCYAKGNIGDKCIPNNTLQKMTSLSYPDGTSINLTNFYFYHACPCLDNLICSKDTE
TCEDPLF
Intron – 1105 bp –
NYDFRGRYYQHTMGRF

Atta Cephalotes (taxonomy ID: 12957) Ace-Ast2

MCQNAERTQRYKTMRSSVRAIIRRAXCTASSGPRFKTT
Intron – 1108 bp –
EIIKSKQAIMLPMLNVVILITSIVVFPNIDPVTSSSPVFQKNCTTNTECEPNSCCL*LG*
Intron – 1242 bp –
*LG*PMRYSIPTCMPFRQKGELCRVNAETITTNLTYPNTLEIKVKDIHYILCPCADGLSCNPKRGICK

Linepithema humile (taxonomy ID: 83485) L.hu-Ast2

MKWRYKMMRLSLRTVIRRAVCDKCGSRIAL*QL*
Intron – 518 bp –
*QL*VKSKQTAIMSPILVALLFISLAIA**A**PPLIPSEQCTTDSECPSDFCCLL
Intron – 139 bp –
GPSRYAMPACMPYQQKGEQCRVNAKTITTNLTYPDNSQLEVKNINFILCSCADGLSCNKKTGICN

Solenopsis invicta (taxonomy ID: 13686) Si-Ast2

MRSSVGAIIRRAXCAANSGPRIRND
Intron – 1049 bp –
RDYQTQTAIMTPISGILILVISIMATSSISSVVPSYQKCNTNEDCKSSSCCLL
Intron – 189 bp –
GPSRYALPSCMPYQQKGEQCRMNADTITTNLSYPDNSQIEVKDIHLILCPCADGLSCDFGICEEDA

Vollenhovia emeryi (taxonomy ID: 411798) Ve-Ast2

MRSSVRAIIHASCVAAGSGPRARND
Intron – 1225 bp –
SGYQAQTAIMSPMPGVLLFISIVTLPNISSIPLSSENCVTNSECQTDSCCVL
Intron – 254 bp –
GASRYVIPTCMPFQQIGETCRVNAATITTNLSYPDNSQLEVTAVHFILCPCAAGLSCDSKHGTCE

2) Chelicerate

Parasteatoda tepidariorum (taxonomy ID: 114398) Pt-Ast2a

FTLSIVVSLLFQ
Intron – 1913 bp –
VCICNTPRECSSKRDCGPNECCVV*G*
Intron – 3377 bp –
*G*RTRYSIPECKPNGRVGNTCLRGAESEDLTLYYPNGQRELEGVYTLFCPCDQNLVCKSNRCTV

Parasteatoda tepidariorum (taxonomy ID: 114398) Pt-Ast2b

MGTPVHMYAFLAAMLVCCFS*Q*
Intron – 6785 bp –
*Q*VSSYTLATSECRSQADCGPGECCVL*G*
Intron – 7829 bp –
MMRYSMAQCMPLGQVEDYCRDDNPPENRTLYYPNGEPVEVYEIYTHVCPCDESLQCTDNFCAMDESY
ENNYLY

Tetranychus urticae (taxonomy ID: 32264) Tu-Ast2

MLCYSTKFIIIFAL*M*
Intron – 549 bp –
*M*VTVSGRTWNFFALNSPKPCQSSDDCRRGECCAI*G*
Intron – 240 bp –
*G*FARFSVPMCKPMGRINDWCYPDNEPENMTLHYPYGSEAYTNVHRNFCPCKQPLTCEHNICKFERFN
YY

Limulus polyphemus (taxonomy ID: 6850) Lp-Ast2a

MRTLVATIIILVAQ
Intron – 4625 bp –
MAQSFPGFRGCRSQQDCDPGSCCVV A?
Intron – 429 bp –
MERFSTPRCQKLSQQGEYCRPRNSALNTSLSYPNGILDVTNLYTVLCPCDVGLICEQAMCQPNTFLQ
SNHLA

Limulus polyphemus (taxonomy ID: 6850) Lp-Ast2b

MKTIVCVFLILLELQ
Intron – ? bp –
VILSFPGFDGCRSPQDCDKSSCCVI T
Intron – 429 bp –

MEKYSVPHCRKLGNKEEYCRTRNSAQNMTLNYPNGSVDVFGVYRILCPCNDGLECVQSVCQLLHSDT
IL

Limulus polyphemus (taxonomy ID: 6850) Lp-Ast2c

MIMRPEITLFILF*TIIM*
Intron – 2450 bp –
*TIIM*LAIGVPYFYGCKSPADCEPGECCVI*G*
Intron – 921 bp –
MNRYSFPRCEKFGQKNDFCLPSNTPQNKTLYYPNGAVDFSNIYMLFCPCDTGFICYQAHCESA

Metaseiulus occidentalis (taxonomy ID: ?) Mo-Ast2

MRSSMERLSLLLISLSVFLEFAACEDVEVRSCRKPSDCDPGYCCRI
Intron – 1268 bp –
GMERFSQPFCQKFGTVGDTCRMGAEPEDKILWFPGGLTFDVFGVYRQFCPCE*G*
Intron – 153 bp –
*G*GLACKEAMCQPESAKIAAPTKKYNIDDFDYESLDNRAKSDNFAEFDI

3) Crustacea

Daphnia pulex (taxonom ID: 6669) Dp-Ast2

MKECGLLFVCWATVVLAGILQPLPSHSAMGDCRSNEDCGPNRCCLL*G*
Intron – 74 bp –
*G*MMRYSTPWCAPLLNLGEDCRPTSSNEPSITNRTLVYPGGLEIFLKDAYQ
Intron – 75 bp –
ILCPCDANQGLVCSHLSGACISDESLNDISPL

Daphnia magna (taxonomy ID: 35525) Dm-Ast2

MLKECSLLFVCWTTLALTATLQPLPSYGVTGDCRSSEDCGPSSCCLL
Intron – 74 bp –
GMMRYSTPWCAPLLKLGDECRPSSHQLINRTLSYPGGLEIFLKDAHQV
Intron – 74 bp –
LCPCDANEGLVCSPLKGTCVYDVANDITPL

Supplementary Table 3. Prediction of proportion of disorder in Phyre2 for the five sequences studied.

Astakine sequence	Proportion of disorder (%)
Pl-Ast1	49
Pl-Ast2	61
Pcl-Ast2	50
Mj-Ast2	62
Pm-Ast2	64
Lv-Ast2	66